PHENOPATIENT: CLINICALLY PLAUSIBLE AND DIVERSE DE NOVO VIRTUAL PATIENTS ALIGNED WITH REAL-WORLD PHENOTYPE DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive virtual patients support medical-AI training and evaluation, but many existing simulators depend on real clinical records. Limited shareability, privacy and data-use restrictions, and scarce rare-disease records constrain their coverage. De novo simulation removes one-to-one record dependence but creates a central tension between individual coherence and cohort diversity: maximizing typicality can produce stereotyped cases, whereas expanding coverage can introduce within-patient contradictions. We introduce PHENOPATIENT, which reconciles these objectives while targeting real-world disease-phenotype distributions. A medical literature retrieval engine supports a severity-conditioned phenotype atlas; probabilistic sampling, cross-modal correction, and an immutable, rule-checked ledger produce diverse, coherent patients. Evaluation spans individual-level plausibility, population-level plausibility, and within-disease diversity. Across 1,000 completed encounters per implementation and 50 diseases, PHENOPATIENT achieves individual-level plausibility comparable to record-grounded simulators, with higher observed scores on most population-level measures and all five within-disease coverage and diversity measures. Under fixed interaction protocols, it yields the largest observed GPT-5.6 Sol–Qwen3.8-27B gaps in broad-category and specific-disease diagnosis, consistent with heterogeneous cases exposing doctor-model differences. Replacing GPT-5.6 Sol with Qwen3.8-27B for patient dialogue produces the smallest mean quality decline among evaluated simulators, consistent with ledger-supported smaller-model operation. Blinded physicians assign the highest reported mean plausibility, logical-consistency, and patient-likeness ratings. These findings support scalable, disease-diverse cases for model evaluation, training-data construction, and controlled clinical-reasoning training.

1 INTRODUCTION

Medical diagnosis sequentially reveals evidence through questions, examinations, and tests, a process absent from static medical question answering. Recent systems couple doctor models with patient, examination, and test agents (Li et al., 2024; Schmidgall et al., 2024; Fan et al., 2024; Almansoori et al., 2025). Many instantiate patients from records or curated profiles, inheriting limited shareability, sparse rare and atypical presentations, and privacy and data-use restrictions (Price & Cohen, 2019; Johnson et al., 2023). Dialogue variation alone cannot create controlled new latent clinical states.

De novo simulation must satisfy three coupled objectives: *individual-level plausibility* requires coherent symptoms, signs, tests, numerical values, and temporal course; *population-level plausibility* requires agreement between aggregated phenotypes and external literature prevalence and phenotype-frequency statistics; and *within-disease diversity* requires varied manifestations of the same diagnosis. Per-case typicality alone can yield uniformly textbook cohorts, sacrificing population structure and diversity. Conversely, unconstrained coverage can combine clinically plausible findings into contradictory patients. Cohort variation therefore requires constraints on each realization.

For diseases with sufficient retrievable evidence, PHENOPATIENT separates disease-level population specification, patient-level realization, and fact-grounded interaction (Figure 1). Its immutable ledger constrains the patient, examination, and diagnostic-test agents throughout dialogue.

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.

Our contributions are:

- We introduce PHENOPATIENT, a de novo framework that compiles disease- and severity-conditioned atomic phenotype probabilities across five clinical modalities into interactive patients. Sampling, cross-modal correction, and rule-checked ledger freezing produce immutable patient states; action-gated role views expose the same state throughout interaction. This integrates population variation with individual consistency without one real record per patient.
- We introduce a three-dimensional evaluation framework for individual-level plausibility, population-level plausibility, and within-disease diversity. Population agreement compares interaction-visible phenotypes with frozen, externally retrieved literature targets, rather than targets estimated from the atlas or generated cohort, enabling a common assessment of simulators with and without phenotype ledgers.
- Across 1,000 completed encounters per implementation covering 50 diseases, PHENOPATIENT achieves individual-level plausibility comparable to record-grounded simulators and leads three of four population-level measures and all five within-disease coverage and diversity measures. Separate fixed-protocol experiments yield the largest observed GPT-5.6 Sol–Qwen3.8-27B gaps in broad and specific diagnosis and the smallest mean quality decline when Qwen replaces Sol for patient dialogue among evaluated simulators.

2 RELATED WORK

Interactive clinical evaluation. MEDIQ evaluates question asking under incomplete information (Li et al., 2024); AGENTCLINIC evaluates sequential diagnosis, multimodal acquisition, and tool use (Schmidgall et al., 2024); and AI HOSPITAL and MEDAGENTSIM organize doctor, patient, examination, and evaluator roles in multi-agent simulations (Fan et al., 2024; Almansoori et al., 2025). Work on the inquiry–diagnosis relationship shows that doctor questioning and patient simulation jointly determine downstream accuracy (Liu et al., 2025). Thus, diagnosis is a property of the full interaction protocol, not of the doctor model alone.

Grounded and behaviorally realistic patients. EVOPATIENT uses agent coevolution to improve standardized-patient behavior (Du et al., 2025). PATIENTSIM uses clinically grounded personas to vary communication and recall (Kyung et al., 2025). PATIENTS WITH PERSONALITY controls behavioral diversity and selective disclosure (Schlager et al., 2026), whereas VERISIM adds controlled communication noise over a fixed medical state (Mansouri et al., 2026). State-aware simulators make evolving case state explicit (Liao et al., 2024), and PATIENTHUB pursues a unified abstraction for patient simulation (Sabour et al., 2026). PSYCHE represents structured psychiatric profiles and behaviors (Lee et al., 2025), and PATIENT-ψ uses LLM patients for mental-health training (Wang et al., 2024). Together, these systems motivate separate evaluation of factual fidelity, clinical state, disclosure, and behavioral variation.

Record-free patient generation. PATIENT-ZERO (Lai et al., 2026) combines stratified attribute sampling, hierarchical record synthesis, and NLI-verified dialogue. Its explicit probabilities concern demographic, socioeconomic, and lifestyle attributes; LLMs generate symptoms and examination findings under clinical constraints, without explicit per-phenotype occurrence probabilities. PHENOPATIENT instead samples atomic phenotypes using disease- and severity-conditioned probabilities across five clinical modalities, enabling direct control and evaluation of within-disease phenotype distributions. Samples become rule-checked, immutable ledgers with action-gated role views. Although Patient-Zero provides prompts, its public repository contained only a README and license on 24 September 2026;[1] we therefore omit comparison using an official implementation.

[1] Patient-Zero repository snapshot, accessed 24 September 2026.

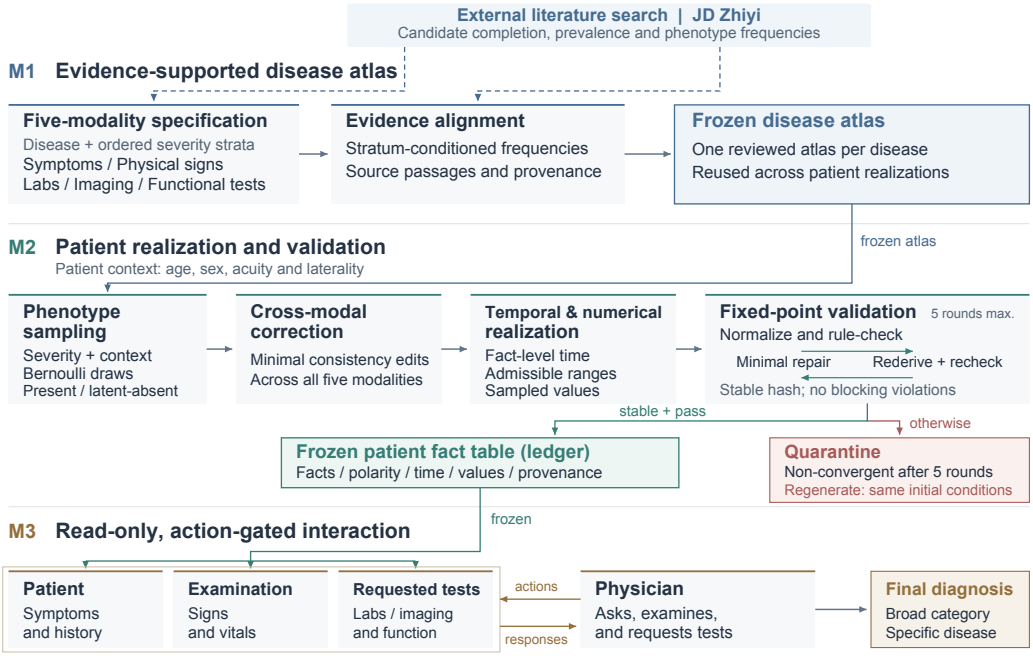

Figure 1: PHENOPATIENT workflow. M1 freezes five-modality phenotype knowledge with literature-derived frequencies. JD Zhiyi supports candidate completion and population targets through external literature search (dashed arrows), only during M1. M2 samples, realizes, corrects, and validates each patient before freezing its fact table. Non-convergent candidates are quarantined and regenerated under the same initial conditions. M3 exposes action-gated, read-only patient, examination, and requested-test views; the physician returns broad-category and specific-disease diagnoses.

3 METHOD

3.1 OVERVIEW

PHENOPATIENT comprises disease-level knowledge construction (M1), patient generation, validation, and freezing (M2), and fact-grounded interaction (M3). M1 freezes an evidence-supported probabilistic phenotype atlas. M2 samples a patient, corrects cross-modal contradictions, realizes temporal and quantitative details, and iterates rule checks and minimal repair until stable. M3 exposes role-specific views of the same immutable patient fact table.

Positioning. PHENOPATIENT contributes a structured orchestration contract that compiles population specifications into sampled, rule-checked, role-partitioned ledgers, rather than new retrieval, sampling, or role-play algorithms. It complements Synthea (Walonoski et al., 2018), HPO (Kohler et al., 2021), and provenance-oriented documentation (Gebru et al., 2021).

3.2 EVIDENCE-AUGMENTED PHENOTYPE ATLAS

M1 constructs an ordered two-to-four-level severity axis and five modality-specific libraries: symptoms, physical signs, laboratory abnormalities, imaging findings, and functional-test abnormalities. Unless a level is anchored to disease-specific clinical criteria, it is an internal generative severity stratum rather than a validated clinical stage. Each atomic phenotype has a stratum-conditioned occurrence probability; symptoms additionally have duration, trigger, and qualitative attributes. Composite diagnoses are decomposed before candidate sets are merged and deduplicated.

The formal experiment used **JD Zhiyi** for literature evidence retrieval during M1 atlas construction. A language model first proposed atomic candidates. JD Zhiyi then received disease–modality queries to supplement missing candidates and disease–phenotype queries to retrieve epidemiological prevalence, phenotype-frequency, and occurrence evidence from external medical literature. The

interface made 203.1 $\pm$ 52.1 queries per disease (mean $\pm$ SD across diseases). For conflicting probability descriptions, the model read the relevant passages and froze the arithmetic mean of their reported values. The retrieved numerical statistics, source passages, and metadata defined frozen population targets before atlas construction; the atlas indexes rather than estimates them. We reused one evidence-augmented atlas per disease across the full method and targeted variants, holding retrieval evidence fixed. JD Zhiyi was not queried during patient sampling or interaction.

3.3 Patient realization and consistency correction

Generation is stratified by disease: fix n_d patient slots with $\sum_d n_d = N$ and $w_d = n_d/N$. Within disease d, slot i draws a severity stratum and atomic phenotypes:

$$s_{di} \sim \text{Categorical}(\rho_d), \qquad z_{dik} \mid s_{di} = s \sim \text{Bernoulli}(\theta_{dsk}), \quad \theta_{dsk} = P(z_k = 1 \mid d, s). \quad (1)$$

Here ρ_d is the prescribed severity mixture and θ_{dsk} the frozen occurrence probability of phenotype k. Age, sex, acuity, and laterality are fixed per slot. Phenotypes are drawn independently before correction; eligible comorbidities can add objective findings. Latent negatives are hidden simulation assignments, not clinical observations: not elicited, not examined, and not tested facts remain unavailable to the doctor. Cross-modal correction minimally resolves incompatible findings across all five modalities. The state is accepted only when no blocking rule violation remains and a canonical state hash stabilizes, within at most five repair rounds.

The temporal layer separates disease duration, current-episode duration, and fact-level time. Signs and tests are anchored to observation time even when their underlying abnormality predates the encounter. For quantifiable findings, the system first generates a rule-admissible range (mean, standard deviation, minimum, and maximum) and then samples a patient-specific value. Direction, unit, threshold, and selected physiological relations are rule-checked jointly.

3.4 Rule-checked patient fact table and interaction

Within M2, the validation-and-freezing step compiles every present or absent fact into a table containing canonical identity, modality, polarity, temporal context, value, provenance, and feasibility metadata. The fixed-point cycle normalizes, rule-checks, minimally patches, rederives dependent quantities, and rechecks. Checks cover polarity, severity, temporal order, laterality, sex, units, thresholds, shared measurements, and selected physiological equations. Hash verification establishes state integrity and convergence, not medical correctness; guarantees are limited to these constraints. A non-convergent candidate is quarantined, then resampled for the same slot with its original disease, severity, and requested context until accepted. This preserves the preset allocation of initial conditions, although correction and rejection can alter phenotype frequencies.

M3 derives patient, examination, and requested-test views; the physician starts without the diagnosis. Matched positive or latent-negative facts return from the frozen ledger. For an unenumerated question, examination, or test, the model may provide an interaction-local, context-compatible completion from that ledger, comorbidities, and dialogue, defaulting to normal unless context supports an abnormal result. It must not contradict or mutate frozen facts, persist as new patient state, or reveal the diagnosis, and is stored only in the encounter trace.

3.5 Targeted implementation variants

NoCorr removes the semantic cross-modal consistency-correction loop while retaining deterministic downstream checks and the ledger. **LLM Selection** replaces Bernoulli sampling with one GPT-5.5 selection over the complete, filtered candidate universe; unselected candidates become latent negatives and cannot be renamed. **Direct Values** replaces range modeling and local sampling with direct model-generated values while retaining unit, threshold, physiological, and ledger checks. Each variant changes one configured upstream component while retaining the remaining interfaces and downstream checks; it does not isolate that component's causal effect.

4 EXPERIMENTAL DESIGN

4.1 SYSTEMS, PANELS, AND ROLES

We divide the compared implementations by how they construct the latent patient state. The *MIMIC-IV record-grounded* family comprises AGENTCLINIC, MEDAGENTSIM, PATIENTSIM, AI HOSPITAL, and MIMIC-OSCE-DIRECT. In this experiment, all five start from the same deidentified MIMIC-IV inpatient cases (Johnson et al., 2023), transformed into each simulator's native profile. A canonical record combines admission demographics with available history, examination, vital-sign, laboratory, microbiology, and radiology fields. Diagnosis labels, codes, discharge diagnoses, and subsequent hospital-course information are evaluator-only. AgentClinic and MedAgentSim receive doctor-safe OSCE profiles; PatientSim receives a flat interview persona plus a deterministic examination/test sidecar; AI Hospital receives a visible-only localized record; and MIMIC-OSCE-DIRECT exposes the converted OSCE state directly. This taxonomy describes the adapters evaluated here, not every possible use of the original frameworks.

The *de novo clinical-state* family comprises DEMODX-ONLY and PHENOPATIENT. For matched evaluation, both retain a minimal specification containing age, sex, target disease, and task-level acuity metadata, but receive no record-specific history, physical examination, laboratory result, imaging report, discharge diagnosis, or subsequent course. DEMODX-ONLY generates its state directly from this minimal seed; PHENOPATIENT instead constructs and samples the evidence-augmented atlas. Thus, *de novo* refers to the clinical state rather than independence from panel-matching variables. This source-aware grouping separates the closest de novo construction comparison from contextual record-grounded references, while retaining all seven systems in the same metric matrix.

All seven implementations are aligned by runtime case identifier and disease allocation. In the main comparison and variants, GPT-5.5 plays patient and physician in Chinese, physician-initiated encounters with up to 15 history turns, one examination stage, eight test requests, and one final diagnosis. Every method and variant completes 1,000 encounters across five systems, ten diseases per system, and 20 patients per disease. Thus $n_d = 20$ and $w_d = 1/50$ specify a disease-balanced benchmark, not disease prevalence in the general population. For the main PHENOPATIENT cohort, we instantiated 1,532 candidate patients; 532 remained non-convergent after permitted repair rounds and were excluded before dialogue or scoring. Resampling retained each failed slot's initial conditions until it was filled, leaving 1,000 finalized patients with the prescribed allocation unchanged. Native logic and profiles are preserved; model roles, language, budgets, case matching, and trace schema are harmonized.

The doctor-model experiment evaluates every simulator–doctor condition on the same completed 500-case panel, with ten patients for each of the same 50 diseases. The selected patient slots—case identifiers, disease assignments, and matched patient-selection variables—were frozen before inference and reused across all seven methods and all four doctor models; each simulator instantiated its native representation for the same matched slots, and no doctor model received a separately redrawn patient sample. Patient dialogue is fixed to GPT-5.5 while GPT-5.6 Sol, GPT-5.5, Claude Opus 5, and Qwen3.8-27B play the doctor. We report the method-level aggregate broad-category and specific-disease diagnostic accuracy under a common diagnosis-normalization rule; cross-method summaries are unweighted means of the seven method-level results. The patient-backbone experiment likewise evaluates every simulator–patient-model condition on a completed 500-case panel, fixes GPT-5.5 as doctor, and changes the patient dialogue endpoint among the same four models. For PHENOPATIENT, atlas construction and the sampled latent profile remain fixed, isolating the robustness of ledger-grounded dialogue realization.

Eleven physicians in internal medicine or surgery performed the paired, within-patient blinded study. The identical set of 50 matched patient cases was sampled by disease—five disease systems, ten diseases per system, one patient per disease—and evaluated under all seven methods; no method received a separately redrawn human-evaluation sample. Patient order was randomized at the patient level, and the same physician evaluated method-specific realizations within a matched-case comparison session. Three ratings per realization gave each method 50 patients and 150 physician–trace rating events, or 350 method–patient realizations, 1,050 rating events, and 5,250 scalar Likert scores overall. Physicians were blinded to method, backend, and diagnosis; the export has anonymous rating slots but no stable physician identifiers.

4.2 METRICS AND REPORTING

Three metric families assess individual-level plausibility and coherence, population-level agreement with external literature-derived prevalence and phenotype-frequency targets, and within-disease coverage and diversity. Targets are retrieved before atlas construction and frozen for scoring; they are not estimated from the atlas structure or generated cohort. We evaluate phenotypes revealed in completed interactions: DEMODX-ONLY has no frozen phenotype ledger, and real-world clinical frequencies are estimated from elicited, examined, tested, or documented findings rather than inaccessible latent states. This evaluates the cohort actually presented to the clinician under the shared interaction protocol. Table 1 and Appendix B define the components. Quality measures use 0–100 scales except severity entropy (0–1); *information leakage* is a separate, lower-is-better percentage. Doctor evaluation reports broad and specific diagnostic accuracy; physician ratings use five 1–5 Likert dimensions.

Severity metrics use strata recognized from completed interactions, rather than hidden generation labels. We report all patient-quality components, leakage, both diagnostic accuracies, and all five physician dimensions. Physician inference uses three-rating method–case means, 50 paired case blocks, 100,000 disease-system-stratified case bootstraps, Friedman tests, and paired Wilcoxon tests jointly Holm-adjusted across 30 PHENOPATIENT comparisons. Krippendorff's ordinal α uses 20,000 case bootstraps; rater effects and two-way ICC are not identifiable without stable IDs. Appendix E gives full methods and reporting rules.

5 RESULTS

Table 1: Main scores (left) and variants relative to Full PHENOPATIENT (right; Δ); 1,000 encounters each. Bold: best absolute value; **red bold**: largest adverse variant change. Severity entropy is on 0–1; other quality metrics are on 0–100. Information leakage is separate; its diagnostic-leakage rate is a lower-is-better percentage with point deltas. S/S: symptoms/signs; E/T: examinations/tests. LLM/Direct: LLM Selection/Direct Values. AC/MAS/PS/AIH/M-dir/Demo: AgentClinic/MedAgentSim/PatientSim/AI Hospital/MIMIC-OSCE-direct/DemoDx-only.

	MIMIC-IV record-grounded					De novo		Ablation Δ vs. Full		
Metric	AC	MAS	PS	AIH	M-dir	Demo	Pheno	NoCorr	LLM	Direct
Individual-level plausibility and coherence										
S/S plausibility	81.72	82.45	79.70	78.78	82.27	**83.63**	83.30	−4.01	−0.51	−0.06
E/T plausibility	81.43	76.32	77.45	77.48	**84.18**	83.29	80.54	−4.82	−0.39	−2.37
Consistency	91.53	90.37	**95.12**	86.30	94.76	88.95	91.07	−4.96	+0.34	−0.96
Course plausibility	76.91	79.67	71.60	74.54	74.56	80.87	**83.18**	−2.96	−1.07	−1.18
Test return	72.94	27.72	47.04	81.52	**99.51**	89.45	73.02	−1.53	−0.55	−1.18
Population-level plausibility										
Severity agreement	29.71	40.07	24.08	27.87	24.72	**53.14**	40.30	−1.26	−1.86	−0.71
S/S rate agreement	33.65	35.85	32.27	29.34	28.58	48.76	**48.93**	−1.86	−5.99	−0.02
E/T rate agreement	14.24	11.57	11.54	23.38	17.31	34.36	**48.26**	−3.17	−8.18	−1.57
Course-length agr.	29.34	31.69	15.84	25.26	18.62	42.85	**75.70**	−1.65	−1.15	+0.24
Within-disease coverage and diversity										
S/S coverage	39.31	41.45	37.83	34.44	33.93	35.65	**42.18**	+0.18	−6.20	−0.92
E/T coverage	16.63	11.67	13.89	24.14	20.02	24.52	**41.66**	−0.11	−6.72	−1.90
IQR matching	48.47	54.33	47.23	54.32	48.94	36.53	**82.01**	−4.24	−6.25	**−11.61**
Severity entropy	.610	.629	.548	.626	.513	.235	**.642**	−.012	−.005	**−.016**
Course diversity	22.52	18.12	32.05	13.69	14.59	67.17	**88.78**	−0.93	**−7.81**	−1.78
Information leakage (lower is better)										
Diagnostic leakage	2.3	2.1	2.7	2.9	2.6	4.1	**1.7**	+0.7	+0.4	**+0.9**

5.1 OVERALL SIMULATION QUALITY

Table 1 jointly reports the complete main comparison and the three targeted implementation variants. The left block separates record-grounded from de novo clinical-state generation and reports absolute scores; the right block reports each variant's change from the complete PHENOPATIENT system. PHENOPATIENT maintains individual-level plausibility comparable to the record-grounded methods while showing its clearest advantages in population-level plausibility and within-disease diversity. It leads three of four population-level measures and all five within-disease coverage and diversity measures. Its margins over the strongest alternative are +32.85 for course-length agreement, +27.68 for quantitative-IQR matching, +21.61 for course diversity, +17.14 for examination coverage, and +13.90 for examination-positive-rate agreement. It also leads symptoms/signs coverage by +0.73 and severity entropy by +.013. Its diagnostic leakage is lowest at 1.7%, 0.4–2.4 points below alternatives.

5.2 TARGETED IMPLEMENTATION VARIANTS

The right-hand block of Table 1 makes the component sensitivities directly comparable with the complete system. Removing semantic cross-modal correction (NoCorr) produces the largest decrease for all five individual-level plausibility metrics and reduces the first four by 2.96–4.96 points. Replacing probabilistic sampling with LLM Selection produces the largest decrease on six population-structure or diversity measures, including positive-rate agreement, coverage, and course diversity by 5.99–8.18 points. Direct Values produces the largest decrease on quantitative-IQR matching (-11.61) and severity entropy ($-.016$). All variants raise leakage by 0.4–0.9 points. The small favorable changes on isolated quality cells are expected descriptive variation. These contrasts are consistent with, but do not isolate, the intended contribution of each component.

5.3 PROTOCOL-SPECIFIC DOCTOR ENDPOINT DIFFERENCES

Table 2 reports diagnostic accuracy and Sol–Qwen separation. Across seven methods, unweighted mean broad-category accuracies are 83.6%, 83.1%, 81.5%, and 62.3% for Sol, Opus, GPT-5.5, and Qwen, respectively; specific-disease accuracies are 69.9%, 67.7%, 64.3%, and 40.1%. Mean Sol–Qwen gaps are 21.3 and 29.7 percentage points. On PHENOPATIENT, Sol attains 99.2% broad and 91.6% specific correctness versus Qwen's 74.0% and 56.6%: the largest gaps, 25.2 and 35.0 points, compared with next-largest gaps of 22.0 and 31.4. Identical 500-case slots exclude separately drawn samples as an explanation. With patient selection, disclosure, interaction budget, evaluator, and label mapping fixed, these results are consistent with heterogeneous realizations exposing doctor-model differences. They measure protocol-specific discrimination, not general clinical competence or safety. The main protocol separately records PHENOPATIENT's lowest observed diagnostic leakage, 1.7%.

Table 2: Diagnostic accuracy (%) for four doctor models using the same 500 patient slots per method; GPT-5.5 drives patient dialogue. $\Delta_{\mathrm{S-Q}}$ is GPT-5.6 Sol minus Qwen3.8-27B; bold marks the largest gap in each diagnostic endpoint. Sol: GPT-5.6 Sol; Opus: Claude Opus 5.

	Broad category					Specific disease				
Method	Sol	GPT-5.5	Opus	Qwen	$\Delta_{\mathrm{S-Q}}$	Sol	GPT-5.5	Opus	Qwen	$\Delta_{\mathrm{S-Q}}$
MIMIC-IV record-grounded										
AgentClinic	79.4	77.2	78.4	60.6	18.8	68.2	62.2	68.4	38.2	30.0
MedAgentSim	73.8	70.6	74.4	51.8	22.0	54.4	47.4	51.2	23.6	30.8
PatientSim	74.2	71.4	72.6	52.2	22.0	51.8	44.0	48.2	20.4	31.4
AI Hospital	80.8	78.4	79.8	59.8	21.0	61.6	55.2	58.6	31.2	30.4
MIMIC-OSCE-direct	77.8	75.2	78.0	56.8	21.0	62.6	56.0	59.4	32.4	30.2
De novo clinical state										
DemoDx-only	99.8	99.4	99.6	80.6	19.2	98.8	98.4	98.6	78.4	20.4
PHENOPATIENT	99.2	98.4	98.8	74.0	**25.2**	91.6	86.6	89.4	56.6	**35.0**

5.4 SENSITIVITY TO THE PATIENT DIALOGUE BACKBONE

With GPT-5.5 fixed as physician, Table 3 changes the patient backbone from GPT-5.6 Sol to Qwen3.8-27B. PHENOPATIENT has the smallest mean decline and the smallest decline on 11 of 14 higher-is-better measures: 1.0–2.4 points on 0–100 measures and .008 in severity entropy. This is consistent with its detailed frozen fact table reducing reconstruction demands on a weaker dialogue model. Appendix Table 7 gives all four backbones; leakage remains 1.4–1.8%.

Table 3: Patient-backbone change, Qwen3.8-27B minus GPT-5.6 Sol (500 encounters per condition; GPT-5.5 physician). For 14 higher-is-better measures, negative values indicate degradation and bold marks the smallest decline before rounding. Severity entropy is on 0–1; other quality measures are on 0–100; leakage is separate in percentage points.

	MIMIC-IV record-grounded					De novo	
Metric	AC	MAS	PS	AIH	M-dir	Demo	Pheno
Individual-level plausibility and coherence							
Symptoms/signs plausibility	-3.3	-3.4	-3.3	-3.4	-2.9	-10.0	**-1.8**
Exams/tests plausibility	-1.9	-2.6	-2.5	-2.7	-1.9	-5.4	**-1.6**
Phenotype consistency	-2.0	-3.0	-2.1	-2.5	-1.9	-8.0	**-1.0**
Course plausibility	-4.4	-3.7	-4.3	-3.2	-2.9	-12.2	**-1.5**
Requested-test return	-1.5	-2.5	**-1.2**	-2.9	-1.6	-6.0	-1.8
Population-level plausibility							
Severity-distribution agreement	-3.2	-1.9	-2.6	-2.3	-2.5	-7.0	**-1.0**
Symptoms/signs rate agreement	-1.9	-2.9	-3.1	-3.2	-2.0	-8.2	**-1.9**
Exams/tests rate agreement	-2.9	-3.1	-2.4	-2.5	-2.4	-8.7	**-2.4**
Course-length agreement	-3.9	-3.6	-3.4	-3.9	-1.6	-8.7	**-1.2**
Within-disease coverage and diversity							
Symptoms/signs coverage	-2.2	-2.0	-1.6	-2.3	**-1.5**	-8.9	-1.6
Exams/tests coverage	-3.5	-3.8	-3.9	-3.7	-3.3	-11.3	**-1.9**
Quantitative-IQR matching	-4.9	-4.3	-3.7	-3.8	-2.9	-12.2	**-1.5**
Severity entropy	-.017	-.009	-.010	-.017	**-.006**	-.026	-.008
Course diversity	-4.1	-5.3	-4.6	-5.7	-3.4	-13.0	**-2.0**
Information leakage (lower is better)							
Diagnostic leakage rate	+0.6	-0.6	+0.6	-0.8	-0.2	+1.0	-0.4

5.5 BLINDED PHYSICIAN EVALUATION

Table 4 gives PHENOPATIENT the highest observed plausibility (4.3), logic (4.3), and patient-likeness (4.3). All omnibus differences survive Holm adjustment ($p < 0.001$; $W = .076$–$.276$). Versus each dimension's highest alternative, plausibility improves by $+0.247$ (95% CI $[0.120, 0.373]$; Holm $p = 0.007$); the other four are indistinguishable. Ordinal α is .030–.103, with 94.4% within-one-point agreement; Appendix E gives details.

6 DISCUSSION

Results support three complementary axes: typicality alone can collapse cohorts, whereas unconstrained coverage can create within-patient conflicts. PHENOPATIENT uses literature-conditioned sampling, correction, and ledger freezing to retain record-comparable individual-level plausibility while improving cohort plausibility and diversity without added leakage. Fixed-panel results are consistent with heterogeneous cases exposing doctor-model gaps and the detailed ledger supporting weaker patient backbones. Intended uses are offline evaluation and training, not clinical care; prospective population and multi-seed validation remain future work.

Table 4: Blinded paired physician ratings (mean ± SD; 1–5): 50 matched cases and 150 raw scores per method/dimension from 11 physicians. SD describes dispersion across the 150 raw scores; inference uses case means. Bold is the highest observed mean, not significance. Ranks and tests use unrounded values.

Method	Plausibility	Logic	Course	Completeness	Patient-likeness
MIMIC-IV record-grounded					
AgentClinic	3.9 ± 0.7	3.9 ± 0.6	4.1 ± 0.6	**4.1** ± 0.6	3.9 ± 0.5
MedAgentSim	3.8 ± 0.7	3.9 ± 0.5	3.9 ± 0.5	4.0 ± 0.5	4.0 ± 0.5
PatientSim	3.8 ± 0.7	3.8 ± 0.5	4.0 ± 0.6	3.9 ± 0.5	4.2 ± 0.5
AI Hospital	3.7 ± 0.7	3.8 ± 0.6	3.8 ± 0.5	3.9 ± 0.6	4.0 ± 0.6
MIMIC-OSCE-direct	4.0 ± 0.6	4.0 ± 0.5	**4.3** ± 0.5	4.0 ± 0.6	4.0 ± 0.6
De novo clinical state					
DemoDx-only	4.1 ± 0.6	4.2 ± 0.5	4.2 ± 0.5	3.8 ± 0.7	3.8 ± 0.6
PHENOPATIENT	**4.3** ± 0.6	**4.3** ± 0.6	4.2 ± 0.6	4.1 ± 0.5	**4.3** ± 0.6

7 CONCLUSION

PHENOPATIENT provides literature-aligned de novo patients without one real record each, balancing individual- and population-level plausibility with diversity. Its discrimination and robustness results support scalable evaluation, training-data construction, and controlled clinical-reasoning training.

ETHICS STATEMENT

This work studies simulated encounters, not diagnosis, treatment, triage, or autonomous patient-facing deployment. Simulators may produce incorrect, incomplete, stereotyped, or misleading information; benchmark performance is not clinical safety. The physician-rating study was exempt from ethics approval; ratings are reported in aggregate without identifying information. Source material retains its licenses and access controls. Risks include overstated competence, representation bias, and unreviewed case reuse.

REPRODUCIBILITY STATEMENT

We report case counts, shared panels, model roles, variants, metric definitions, and aggregate matrices. Reproduction requires manifests, prompts, interaction settings, evaluation code, frozen retrieval outputs, atlas hashes, and per-example outputs. Anonymous review artifact: anonymous.4open.science/r/iclr-review-artifact-7F3C/.

AI USE STATEMENT

Generative AI served virtual-patient, doctor, and evaluation roles and assisted research critique, literature organization, implementation, data processing, interpretation, drafting, and editing. The authors checked numbers against stored artifacts, citations against sources, and prose for unsupported claims, and take responsibility for all content.

REFERENCES

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Medagentsim: Self-evolving multi-agent simulations for realistic clinical interactions. *arXiv preprint arXiv:2503.22678*, 2025.

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haochao Ying. Llms can simulate standardized patients via agent coevolution. *arXiv preprint arXiv:2412.11716*, 2025.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xie, Fei Huang, and Jingren Zhou. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*, 2024.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021. doi: 10.1145/3458723.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Sebastian Kohler, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M. Brower, et al. The human phenotype ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217, 2021. doi: 10.1093/nar/gkaa1043.

Daeun Kyung, Hyunseung Chung, Seongsu Bae, Jiho Kim, Jae Ho Sohn, Taerim Kim, Soo Kyung Kim, and Edward Choi. Patientsim: A persona-driven simulator for realistic doctor-patient interactions. *arXiv preprint arXiv:2505.17818*, 2025.

Yunghwei Lai, Ziyue Wang, Weizhi Ma, and Yang Liu. Patient-Zero: Scaling synthetic patient agents to real-world distributions without real patient data. *arXiv preprint arXiv:2509.11078v2*, 2026. URL https://arxiv.org/abs/2509.11078v2.

Jingoo Lee, Kyungho Lim, Young-Chul Jung, and Byung-Hoon Kim. Psyche: A multi-faceted patient simulation framework for evaluation of psychiatric assessment conversational agents. *arXiv preprint arXiv:2501.01594*, 2025.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *arXiv preprint arXiv:2406.00922*, 2024.

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yu Wang, and Yanfeng Wang. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*, 2024.

Zhaocheng Liu, Quan Tu, Wen Ye, Yu Xiao, Zhishou Zhang, and Hengfu Cui. Exploring the inquiry-diagnosis relationship with advanced patient simulators. *arXiv preprint arXiv:2501.09484*, 2025.

Sina Mansouri, Mohit Marvania, Vibhavari Ashok Shihorkar, Han Ngoc Tran, Kazhal Shafiei, Mehrdad Fazli, Yikuan Li, and Ziwei Zhu. Verisim: A configurable framework for evaluating medical ai under realistic patient noise. *arXiv preprint arXiv:2604.10441*, 2026.

W. Nicholson II Price and I. Glenn Cohen. Privacy in the age of medical big data. *Nature Medicine*, 25(1):37–43, 2019. doi: 10.1038/s41591-018-0272-7.

Sahand Sabour, TszYam Ng, and Minlie Huang. Patienthub: A unified framework for patient simulation. *arXiv preprint arXiv:2602.11684*, 2026.

Moritz Schlager, Friederike Jungmann, Samuel Schmidgall, Philipp Raffler, Franziska Hartl, Eva Wende, Paula Rossmuller, Conrad Ketzer, Avinatan Hassidim, Dale R. Webster, Yossi Matias, and Yun Liu. Patients with personality: Realistic patient simulation through controlled diversity and selective disclosure. *arXiv preprint arXiv:2606.17441*, 2026.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, 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.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 2018. doi: 10.1093/jamia/ocx079.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate V. Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12772–12797, 2024. doi: 10.18653/v1/2024.emnlp-main.711.

A Complementary Patient-Quality Profiles

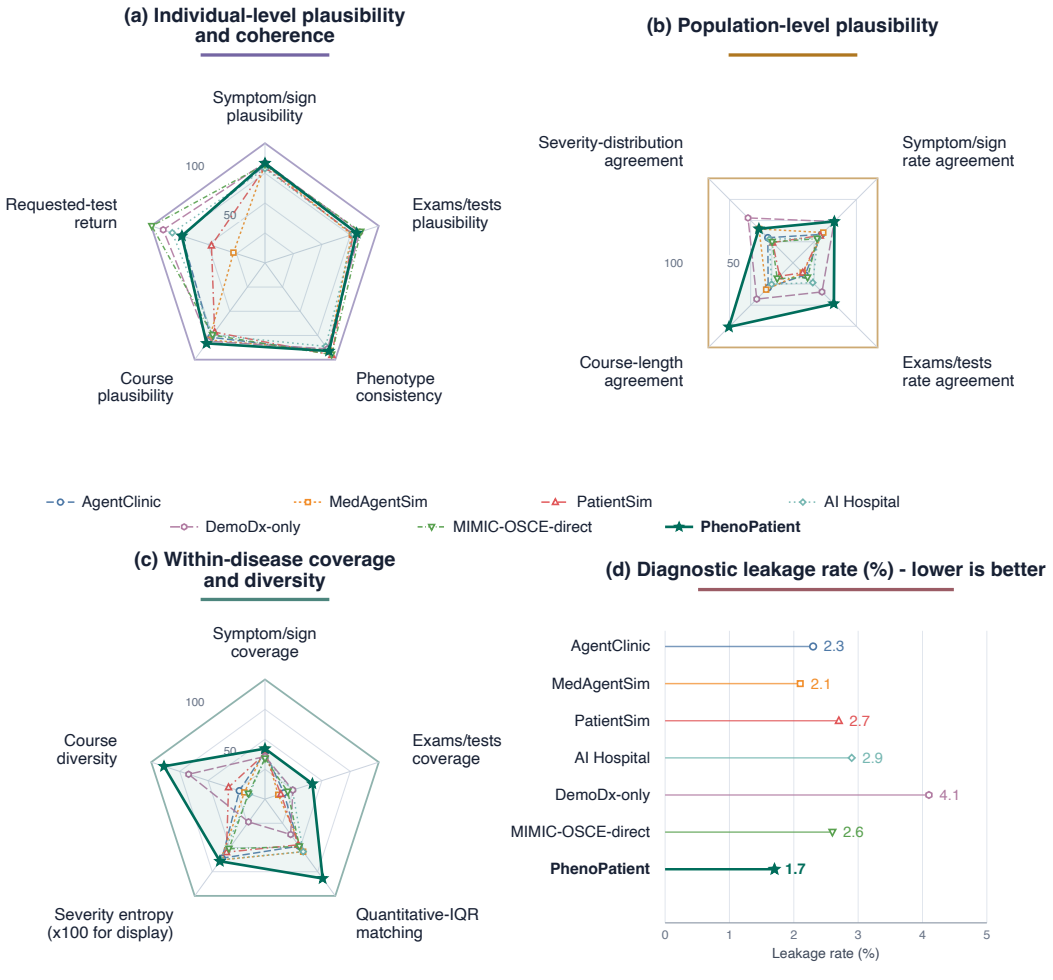

Figure 2: Complete main-method profiles across all 15 reported metrics. Panels (a)–(c) show all seven methods on the five individual-level, four population-level, and five within-disease quality metrics, respectively (0–100; higher is better). Severity entropy, whose original range is 0–1, is multiplied by 100 for visualization only. Panel (d) reports diagnostic leakage separately (lower is better). Radar area is not an aggregate score; Table 1 gives the exact values.

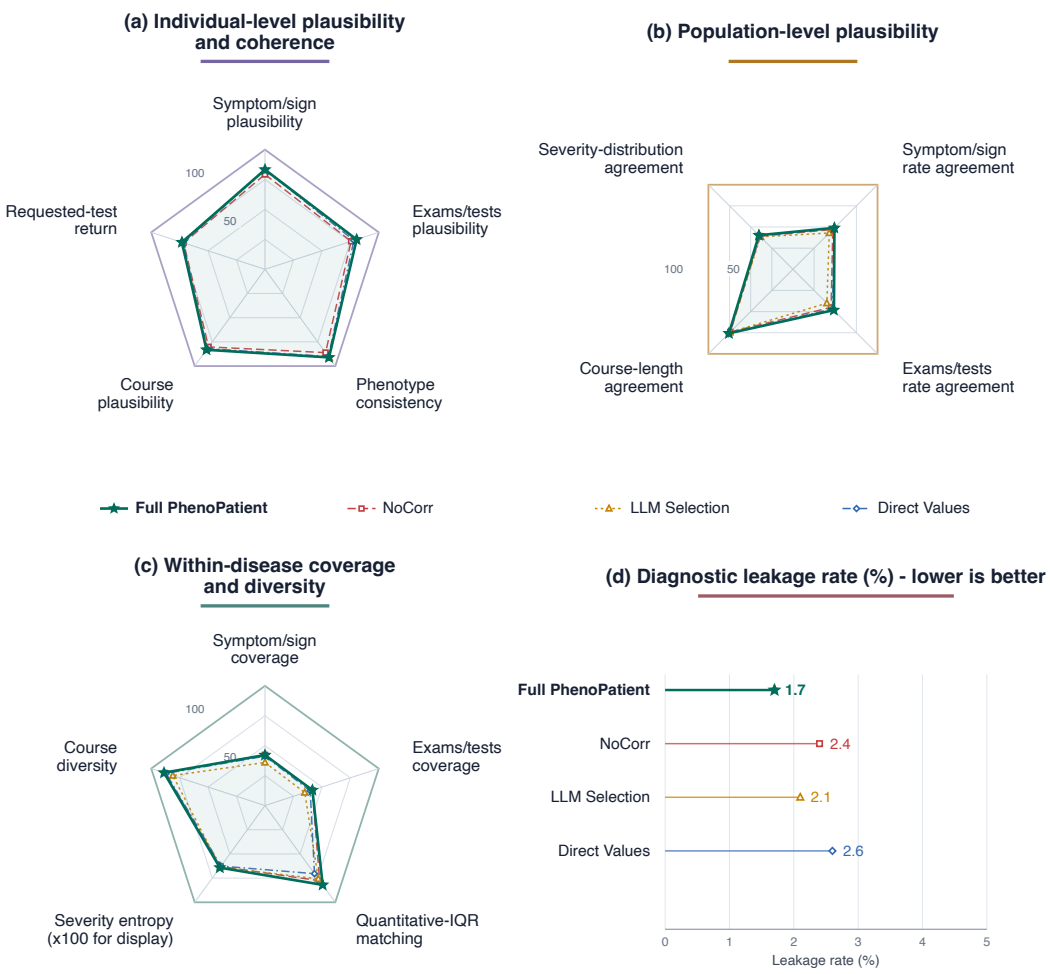

Figure 3: Complete targeted-variant profiles across the same 15 metrics for Full PHENOPATIENT, NoCorr, LLM Selection, and Direct Values. Panels (a)–(c) include every quality metric; severity entropy is multiplied by 100 for visualization only. Panel (d) keeps diagnostic leakage separate and lower-is-better. Exact absolute values and changes from Full appear in Table 1.

B OPERATIONAL METRIC DEFINITIONS

All phenotype-distribution metrics use the completed encounter trace, not hidden generation tables. This supplies a common observable representation even for DEMODX-ONLY, which lacks a frozen phenotype ledger, and reflects the ascertainment required to count phenotypes in empirical clinical data. The resulting agreement is conditional on the shared questioning, examination, and testing protocol; it is not an estimate of unobserved phenotype prevalence independent of that protocol. An unrevealed fact remains unavailable rather than becoming a clinical negative. The missingness and recognition rules below apply separately to each metric.

Severity is recognized from the history, examination findings, and test results visible in each completed interaction, using the disease's fixed severity categories. Let $q_{stratum}$ be the distribution of recognized categories and $c_{stratum}$ the fraction of encounters with a recognizable category. Both severity metrics use this interaction-derived distribution; recognition may change with the dialogue backbone even when the latent patient is fixed. Coverage uses all n_d completed encounters for disease d: $n_d = 20$ in the main comparison and variants, and $n_d = 10$ in the patient-backbone experiment. The sets S_i and E_i contain symptom/sign and examination/test atoms observed in encounter i, respectively.

Table 5: Operational definitions for individual-level plausibility and the first two population-level plausibility measures. Encounter- and disease-level aggregation follows the unit stated for each metric.

Metric	Scoring function, evaluation unit, reference, and aggregation
Symptoms/signs plausibility	Medical plausibility of symptoms and signs visible in one encounter, scored 0–100. *Reference / evaluator:* Fixed GPT-5.5 rubric aggregate. *Aggregation and missingness:* Mean over judge-evaluable encounters.
Exams/tests plausibility	Medical plausibility of visible examination/test atoms, scored 0–100. *Reference / evaluator:* Fixed GPT-5.5 rubric aggregate. *Aggregation and missingness:* Mean over judge-evaluable encounters.
Phenotype consistency	Degree of contradiction freedom among visible atoms, scored 0–100; $100 - \text{score}$ is not an error-event rate. *Reference / evaluator:* Fixed GPT-5.5 rubric aggregate. *Aggregation and missingness:* Mean over judge-evaluable encounters.
Course plausibility	Clinical plausibility of the visible disease course, scored 0–100. *Reference / evaluator:* Fixed GPT-5.5 rubric aggregate. *Aggregation and missingness:* Mean over encounters with sufficient visible course information.
Requested-test return	$100\,(N_{returned} + 0.5N_{partial})/N_{requested}$ for top-level requested tests; unavailable or omitted items receive 0. *Reference / evaluator:* Requested/returned test log. *Aggregation and missingness:* Per encounter, then aggregate; this is request completion, not guideline completeness.
Severity-distribution agreement	$100[1 - \text{JSD}_2(q_{stratum}, p_{lit})]c_{stratum}$, where $q_{stratum}$ is the interaction-recognized stratum distribution, p_{lit} the frozen literature-derived target, and $c_{stratum}$ the recognizable-encounter fraction. *Reference / evaluator:* Interaction-recognized strata and frozen external literature targets. *Aggregation and missingness:* Disease-level distributions with coverage penalty; not clinical-stage validity.
Symptoms/signs rate agreement	$100[1 - \text{BC}(q_{visible}, p_{lit})]$ for symptom/sign positive-rate vectors, where p_{lit} is retrieved from external medical literature. *Reference / evaluator:* Frozen external literature targets. *Aggregation and missingness:* Disease-level vector comparison; not the proportion of symptomatic patients.

Table 6: Operational definitions for the remaining population-level plausibility and within-disease diversity measures, plus the separate information-leakage category. BC is Bray–Curtis dissimilarity; JSD is base-2 Jensen–Shannon divergence.

Metric	Scoring function, evaluation unit, reference, and aggregation
Exams/tests rate agreement	$100[1 - \text{BC}(q_{visible}, p_{lit})]$ for examination/test positive-rate vectors, where p_{lit} is retrieved from external medical literature. *Reference / evaluator:* Frozen external literature targets. *Aggregation and missingness:* Disease-level vector comparison.
Course-length agreement	$100[1 - \text{JSD}_2(q_{course}, p_{lit})]c_{course}$ for five-bin course-length distributions, where p_{lit} is the external literature-derived course target and c_{course} is recognizable-course coverage. *Reference / evaluator:* Frozen external literature targets. *Aggregation and missingness:* Disease-level distributions; missing recognizable courses reduce coverage.

Continued on next page

Metric	Scoring function, evaluation unit, reference, and aggregation
Symptoms/signs coverage	$100 \mid \cup_{i=1}^{n_d} S_i \mid / \mid S_{lit} \mid$ for symptom/sign atoms observed across n_d same-disease encounters, relative to the external evidence-derived candidate set. *Reference / evaluator:* Frozen external evidence set. *Aggregation and missingness:* Per disease, then aggregate; $n_d = 20$ (main/variants), 10 (backbones).
Exams/tests coverage	$100 \mid \cup_{i=1}^{n_d} E_i \mid / \mid E_{lit} \mid$ for examination/test atoms observed across n_d same-disease encounters, relative to the external evidence-derived candidate set. *Reference / evaluator:* Frozen external evidence set. *Aggregation and missingness:* Per disease, then aggregate; $n_d = 20$ (main/variants), 10 (backbones).
Quantitative-IQR matching	For normalized atom–unit pairs with at least three values on each side, $100 \min(IQR_g/IQR_r, IQR_r/IQR_g)$; both zero gives 100 and one zero gives 0. *Reference / evaluator:* Frozen external literature numeric targets. *Aggregation and missingness:* Mean over evaluable atom–unit pairs.
Severity entropy	$H(q_{stratum})/\log K$ over K fixed severity categories, using strata recognized from completed interactions. *Reference / evaluator:* Interaction-recognized stratum labels. *Aggregation and missingness:* Recognizable encounters; 0–1. Uniformity is not target agreement.
Course diversity	$100[1 - \text{mean}_{i<j} J(C_i, C_j)]$, where C_i is the course-atom set of encounter i and J is Jaccard similarity. *Reference / evaluator:* Generated encounter set. *Aggregation and missingness:* Within-disease pairwise diversity; no independent validity gate.
Information leakage (lower is better)	
Diagnostic leakage rate	$100N_{leak}/N_{valid}$, where an encounter is counted once if the patient reveals the evaluator-only current target diagnosis before the physician submits a final diagnosis. *Reference / evaluator:* Dialogue trace and hidden target label. *Aggregation and missingness:* Encounter-level percentage; patient-known prior diagnoses are excluded; lower is better.

Across the complete metric set, non-evaluable items are excluded where the operational definition requires an evaluable judge output, atom–unit pair, recognizable severity category, or valid dialogue trace. Severity-distribution and course-length agreement incorporate recognition coverage directly; severity entropy is conditional on recognizable encounters. No missing value is imputed as a zero score.

C Absolute Patient-Backbone Scores

Table 7 reports the absolute PHENOPATIENT scores across all four patient dialogue backbones. The cross-method Sol-to-Qwen changes and row-wise robustness comparison are reported in the main text (Table 3).

Table 7: PHENOPATIENT under four patient dialogue backbones (500 encounters each; GPT-5.5 physician). Parentheses show change from GPT-5.6 Sol. Severity entropy is on 0–1; other quality measures are on 0–100. Information leakage is separate and reported as a lower-is-better percentage.

Metric	GPT-5.6 Sol	GPT-5.5	Claude Opus 5	Qwen3.8-27B
Individual-level plausibility and coherence				
Symptoms/signs plausibility	83.1	83.5 (+0.4)	83.9 (+0.8)	81.3 (−1.8)
Exams/tests plausibility	81.0	80.8 (−0.2)	80.1 (−0.9)	79.4 (−1.6)
Phenotype consistency	90.7	91.0 (+0.3)	90.9 (+0.2)	89.6 (−1.0)
Course plausibility	82.4	83.0 (+0.6)	82.5 (+0.1)	80.9 (−1.5)
Requested-test return	73.7	73.5 (−0.2)	73.1 (−0.6)	71.9 (−1.8)
Population-level plausibility				
Severity-distribution agreement	39.9	39.8 (−0.1)	40.4 (+0.5)	38.9 (−1.0)
Symptoms/signs rate agreement	49.3	48.4 (−1.0)	49.5 (+0.2)	47.4 (−1.9)
Exams/tests rate agreement	48.9	48.0 (−0.9)	47.9 (−1.0)	46.5 (−2.4)
Course-length agreement	75.1	75.6 (+0.5)	76.4 (+1.3)	73.9 (−1.2)
Within-disease coverage and diversity				
Symptoms/signs coverage	42.3	42.1 (−0.3)	43.0 (+0.7)	40.7 (−1.6)
Exams/tests coverage	41.5	42.1 (+0.6)	42.5 (+1.0)	39.7 (−1.9)
Quantitative-IQR matching	81.2	81.7 (+0.5)	81.6 (+0.4)	79.7 (−1.5)
Severity entropy	.644	.640 (−.004)	.633 (−.011)	.636 (−.008)
Course diversity	88.3	88.5 (+0.3)	88.8 (+0.6)	86.3 (−2.0)
Information leakage (lower is better)				
Diagnostic leakage rate	1.8	1.8 (+0.0)	1.6 (−0.2)	1.4 (−0.4)

D PROTOCOL AND CAPABILITY MATRIX

Table 8: Controls and interpretation boundaries for the five completed evaluation blocks.

Experiment	Controls, deliberate variation, and supported interpretation
Main comparison	*Shared controls:* 1,000 encounters per method; GPT-5.5 patient and doctor; doctor-first interaction *Deliberate variation:* Native profile representation and patient-construction method *Supported interpretation:* Individual-level plausibility, population-level plausibility, and within-disease diversity
Implementation variants	*Shared controls:* 1,000 encounters per variant; same roles and downstream rule checks *Deliberate variation:* One sampling, correction, or numeric-realization component *Supported interpretation:* Descriptive sensitivity associated with each intervention
Doctor endpoints	*Shared controls:* Same 500 selected patient slots across every method–doctor condition; GPT-5.5 patient *Deliberate variation:* Doctor model *Supported interpretation:* Broad and specific diagnostic-accuracy separation
Patient backbones	*Shared controls:* 500 encounters per method–patient condition; GPT-5.5 doctor *Deliberate variation:* Patient dialogue model *Supported interpretation:* Robustness of visible patient quality
Physician study	*Shared controls:* Identical 50 matched cases across methods; no redraw; 150 rating events/method; 1,050 events (5,250 scalar scores); 11 physicians *Deliberate variation:* Simulator method *Supported interpretation:* Paired, blinded physician comparison on five dimensions

All systems use the same disease allocation and doctor-first interaction contract, while each adapter preserves its native profile format. The five record-grounded systems use their MIMIC-IV-derived profiles; DEMODX-ONLY and PHENOPATIENT construct new clinical states from minimal disease specifications. Only PHENOPATIENT uses a literature-supported phenotype atlas.

E PHYSICIAN-EVALUATION STATISTICAL DETAIL

The released score matrix contains 50 matched cases × seven methods × three anonymous rating slots × five dimensions. Initial case selection sampled one patient from each disease. For statistical inference, bootstrap resampling instead draws matched case blocks within each of the five disease systems, each containing ten selected cases. Table 9 reports all descriptive cells and case-bootstrap confidence intervals. The SD is computed across 150 raw scores, whereas the confidence interval resamples 50 matched cases within disease system and retains all ratings and methods attached to each selected case.

Table 9: Complete physician-rating estimates. Each cell is raw-rating mean ± SD, with the disease-system-stratified case-bootstrap 95% CI for the mean on the second line ($B = 100{,}000$; seed 20260923).

Method	Plausibility	Logic	Course
AgentClinic	3.9 ± 0.7 $[3.8, 4.0]$	3.9 ± 0.6 $[3.8, 4.0]$	4.1 ± 0.6 $[4.0, 4.2]$
MedAgentSim	3.8 ± 0.7 $[3.7, 3.9]$	3.9 ± 0.5 $[3.8, 3.9]$	3.9 ± 0.5 $[3.8, 4.0]$
PatientSim	3.8 ± 0.7 $[3.7, 3.9]$	3.8 ± 0.5 $[3.8, 3.9]$	4.0 ± 0.6 $[4.0, 4.1]$
AI Hospital	3.7 ± 0.7 $[3.6, 3.8]$	3.8 ± 0.6 $[3.7, 3.9]$	3.8 ± 0.5 $[3.8, 3.9]$
MIMIC-OSCE-direct	4.0 ± 0.6 $[3.9, 4.1]$	4.0 ± 0.5 $[4.0, 4.1]$	4.3 ± 0.5 $[4.2, 4.3]$
DemoDx-only	4.1 ± 0.6 $[4.0, 4.2]$	4.2 ± 0.5 $[4.2, 4.3]$	4.2 ± 0.5 $[4.1, 4.2]$
PHENOPATIENT	4.3 ± 0.6 $[4.2, 4.4]$	4.3 ± 0.6 $[4.2, 4.4]$	4.2 ± 0.6 $[4.1, 4.3]$

Method	Completeness	Patient-likeness
AgentClinic	4.1 ± 0.6 $[4.0, 4.2]$	3.9 ± 0.5 $[3.8, 4.0]$
MedAgentSim	4.0 ± 0.5 $[3.9, 4.1]$	4.0 ± 0.5 $[3.9, 4.1]$
PatientSim	3.9 ± 0.5 $[3.8, 4.0]$	4.2 ± 0.5 $[4.1, 4.3]$
AI Hospital	3.9 ± 0.6 $[3.8, 4.0]$	4.0 ± 0.6 $[3.9, 4.1]$
MIMIC-OSCE-direct	4.0 ± 0.6 $[3.9, 4.1]$	4.0 ± 0.6 $[3.9, 4.1]$
DemoDx-only	3.8 ± 0.7 $[3.7, 3.9]$	3.8 ± 0.6 $[3.7, 3.9]$
PHENOPATIENT	4.1 ± 0.5 $[4.0, 4.1]$	4.3 ± 0.6 $[4.2, 4.3]$

Table 10 combines the omnibus method tests and agreement analysis. Each Friedman test uses the 50 cases as blocks after averaging the three ratings within a method–case. The five omnibus p values are Holm-adjusted as a family. Ordinal α is calculated across 350 method–case units per dimension; its interval uses 20,000 disease-system-stratified case resamples. Exact and within-one-point columns are raw pairwise agreement rates among the three ratings per unit.

Table 10: Omnibus method effects and physician-rating agreement. W: Kendall's concordance; p_{H}: Holm-adjusted p; CI: case-bootstrap 95% confidence interval.

Dimension	$\chi^2(6)$	p_{H}	W
Plausibility	67.496	< 0.001	.225
Logic	82.809	< 0.001	.276
Course	77.690	< 0.001	.259
Completeness	22.677	< 0.001	.076
Patient-likeness	58.706	< 0.001	.196
Pooled agreement	—	—	—

Dimension	Ordinal α [CI]	Exact (%)	Within 1 (%)
Plausibility	.078 [.011, .142]	45.4	91.2
Logic	.103 [.051, .154]	51.3	96.1
Course	.035 [−.014, .083]	51.6	95.8
Completeness	.030 [−.037, .100]	52.4	93.7
Patient-likeness	.084 [.019, .148]	53.0	95.2
Pooled agreement	.071 [.038, .103]	50.7	94.4

Table 11 gives all 30 prespecified paired comparisons. A positive difference favors PHENOPATIENT. Confidence intervals are percentile intervals from 100,000 stratified case resamples; the two-sided Wilcoxon p values are Holm-adjusted jointly across all 30 comparisons. Values below 0.001 are shown as $p < 0.001$.

Table 11: Complete PHENOPATIENT minus comparator physician-rating comparisons ($n = 50$ matched cases per comparison). r_{rb} is the matched-pairs rank-biserial effect.

Comparator	Mean Δ [95% CI]	r_{rb}	p_{Holm}
Plausibility			
AgentClinic	+0.393 [+0.247, +0.547]	+.717	.001
MedAgentSim	+0.493 [+0.360, +0.627]	+.846	< 0.001
PatientSim	+0.527 [+0.373, +0.680]	+.882	< 0.001
AI Hospital	+0.593 [+0.453, +0.733]	+.893	< 0.001
MIMIC-OSCE-direct	+0.320 [+0.200, +0.440]	+.816	< 0.001
DemoDx-only	+0.247 [+0.120, +0.373]	+.689	.007
Logic			
AgentClinic	+0.347 [+0.207, +0.487]	+.705	.002
MedAgentSim	+0.420 [+0.280, +0.553]	+.766	< 0.001
PatientSim	+0.447 [+0.313, +0.580]	+.850	< 0.001
AI Hospital	+0.500 [+0.360, +0.640]	+.831	< 0.001
MIMIC-OSCE-direct	+0.260 [+0.133, +0.380]	+.638	.008
DemoDx-only	+0.047 [−0.067, +0.160]	+.187	1.000
Course			
AgentClinic	+0.153 [0.000, +0.307]	+.260	1.000
MedAgentSim	+0.333 [+0.200, +0.460]	+.696	.002
PatientSim	+0.193 [+0.060, +0.320]	+.470	.082
AI Hospital	+0.413 [+0.293, +0.527]	+.802	< 0.001
MIMIC-OSCE-direct	−0.027 [−0.153, +0.100]	−.086	1.000
DemoDx-only	+0.073 [−0.053, +0.200]	+.185	1.000

Continued on next page

Comparator	Mean Δ [95% CI]	r_{rb}	p_{Holm}
Completeness			
AgentClinic	$-0.020\ [-0.153, +0.107]$	$-.055$	1.000
MedAgentSim	$+0.073\ [-0.027, +0.173]$	$+.328$	1.000
PatientSim	$+0.140\ [+0.027, +0.260]$	$+.504$	.106
AI Hospital	$+0.160\ [+0.033, +0.280]$	$+.465$	.126
MIMIC-OSCE-direct	$+0.033\ [-0.093, +0.160]$	$+.071$	1.000
DemoDx-only	$+0.273\ [+0.140, +0.407]$	$+.563$	.028
Patient-likeness			
AgentClinic	$+0.367\ [+0.253, +0.480]$	$+.850$	< 0.001
MedAgentSim	$+0.287\ [+0.160, +0.413]$	$+.671$	.005
PatientSim	$+0.047\ [-0.060, +0.160]$	$+.197$	1.000
AI Hospital	$+0.220\ [+0.080, +0.360]$	$+.433$	.115
MIMIC-OSCE-direct	$+0.247\ [+0.113, +0.373]$	$+.556$	.035
DemoDx-only	$+0.453\ [+0.327, +0.580]$	$+.826$	< 0.001

The score export contains no stable physician identifier linking its anonymous rater slots across method–case rows. Accordingly, these analyses do not estimate physician random effects, physician-population uncertainty, rater-specific bias, or two-way ICC. The low ordinal α values coexist with high within-one-point agreement because ratings are concentrated in a narrow range: 65.35% of the 5,250 scalar ratings equal 4. This limitation does not alter the paired-case comparisons, whose inferential unit is the matched case, but it limits generalization of agreement beyond the realized rating pool.

F Algorithmic Detail

1. Build and rule-check a two-to-four-level internal severity axis for disease d.
2. For each of five modalities, generate atomic candidates; query the medical literature retrieval engine for missing candidates and phenotype occurrence evidence; merge, deduplicate, and estimate severity-conditioned probabilities.
3. Allocate n_d slots per disease. For each slot, fix its severity and requested context, then draw positive/negative phenotype sets using Equation 1.
4. Correct cross-modal incompatibilities, build hierarchical time, realize numerical ranges and values, and iterate the ledger to a fixed point.
5. Quarantine a non-convergent candidate and redraw its phenotype realization with the same initial disease, severity, and context until the slot is filled. Preserve the initial slot counts and proportions; do not transfer failed slots to easier conditions. Derive the chief complaint and role-specific views only from an accepted ledger; fail closed if it becomes stale or inconsistent.
6. Run doctor-initiated interaction and normalize the final broad and specific diagnoses against the frozen label map.

G AUDITABLE END-TO-END CASE STUDY

We randomly selected one completed case from the archived formal run: source case `gastrointestinal__crohn_disease__06` (runtime case `vp_29805771a212ef25`). It has a three-year disease course with a ten-day acute exacerbation, a complete 13-turn encounter, and eight requested test groups. The appendix presents a case interaction illustration comprising a compact M1 count summary, the M2 positive facts with counts of latent-negative facts, the complete history-taking dialogue, and examination and test panels. Case-level scores and diagnostic outcomes come from the archived evaluation. M1 and M2 remained hidden from the doctor. Manual turn-by-turn review found no patient utterance that explicitly disclosed the target-diagnosis label. This limited check does not exclude strong clinical cues, and the case is illustrative rather than additional quantitative evidence.

G.1 COMPACT M1 ATLAS COUNT SUMMARY

For Crohn disease, M1 organized the candidate space into the following clinical dimensions before patient-level sampling. To keep the case study readable, we report only the number of candidates in each dimension; the full frozen atlas remains part of the archived run artifact.

Table 12: Numbers of candidate entries in the frozen Crohn disease M1 atlas.

M1 dimension	Candidates
Symptoms	78
Physical signs	79
Laboratory findings	78
Imaging findings	9
Functional-test findings	24
Diagnostic subdisease	1
Eligible comorbidities	20
Comorbidity-phenotype candidates	108

In prose, M1 therefore contained 78 symptoms, 79 physical signs, 78 laboratory findings, 9 imaging findings, 24 functional-test findings, 1 diagnostic subdisease, 20 eligible comorbidities, and 108 comorbidity-phenotype candidates.

G.2 M2 PATIENT REALIZATION AND CASE INTERACTION

This completed case has source identifier `gastrointestinal__crohn_disease__06` and runtime identifier `vp_29805771a212ef25`. The patient is a 43-year-old woman with mild activity; current exacerbation 10 days; underlying disease duration 3 years. The frozen comorbidities are iron-deficiency anemia, vitamin D deficiency, anxiety disorder, and tobacco dependence. M2 validation converged; one revalidation round; zero patches; zero blockers. The frozen fact-table SHA-256 is `f9fafe1aceb4856f80ee48ac5e9ba66a2e237486669798276fce2d0ed1178122`.

Case-level outcomes. The archived case evaluation reports individual scores of symptoms/signs 85; examinations/tests 85; contradiction-free 95; course 85. The requested-test return was 6 fully returned + 1 partially returned + 1 unavailable of 8 groups = 81.25%. Stored diagnosis outcomes were broad category correct; specific disease correct.

G.2.1 M2 POSITIVE FACTS

The resulting M2 patient profile contains 51 positive clinical facts: 17 symptoms, 4 physical signs, 24 laboratory findings, 1 imaging finding, and 5 functional-test findings.

Symptoms (17) Recurrent diarrhea; mushy loose stools; abdominal cramping pain; watery diarrhea; mucus in stool; feeling of anal swelling; increased flatus; loss of appetite; oral-ulcer pain; anxiety; dysuria; abdominal distension; periumbilical colic; postprandial

	worsening of abdominal pain; increased bowel sounds; nausea; and predefecatory abdominal pain.
Signs (4)	Right-lower-quadrant tenderness near McBurney's point; periumbilical tenderness; hyperdynamic precordium; and tobacco staining of the teeth or fingers.
Laboratory (24)	Intestinal mucosal biopsy showing chronic active inflammation, focal active inflammation, noncaseating granulomas, and crypt abscesses; elevated neutrophil count (7.6×10^9/L); increased neutrophil ratio (80.18%); elevated erythrocyte sedimentation rate (28.43mm/h); elevated serum amyloid A (29.1mg/L); decreased transferrin saturation (7.64%); elevated total iron-binding capacity (73.61μmol/L); decreased serum 25-hydroxyvitamin D (21.49ng/mL); decreased total protein (60.93g/L); decreased total calcium (1.9559mmol/L); elevated fibrinogen (4.606g/L); elevated fecal calprotectin (336.52μg/g); increased fecal fat (7.75g/24h); positive anti-*Saccharomyces cerevisiae* IgA antibody; reduced hemoglobin (118.08g/L); decreased mean corpuscular volume (74.93fL); decreased serum ferritin (17.73ng/mL); elevated red-cell distribution width (16.83%); elevated parathyroid hormone (88.24pg/mL); positive urine or blood cotinine; and blood calcium reduced or low-normal.
Imaging (1)	Capsule endoscopy showing small-bowel aphthous ulcers.
Functional (5)	Anorectal manometry showing reduced resting anal pressure; increased GAD-7 score; elevated Hamilton Anxiety Rating Scale score; elevated exhaled carbon monoxide (11.64ppm); and increased Nicotine Dependence Scale score.

The frozen M2 state also contained 61 absent symptoms, 78 absent physical signs, 58 absent laboratory findings, 8 absent imaging findings, and 25 absent functional-test findings, together with three prior visits. These latent-negative entries are summarized by count rather than reproduced item by item.

G.2.2 Complete Formal History-Taking Dialogue (13 Turns)

Table 13: Complete formal history-taking dialogue (13 turns).

Turn	GPT-5.5 doctor	GPT-5.5 patient
1	What brings you in today?	For about ten days I have mainly had recurrent watery diarrhea. Recently, after eating, my abdomen also develops cramping pain in waves.
2	About how many bowel movements do you have each day?	About four or five times a day, and up to roughly six times when it is worse. I am more likely to need the toilet after eating.
3	Have you noticed blood or mucus in the stool?	I have not seen blood, and the stool is not black. There has been a small amount of mucus over the past few days.
4	Have you had a fever recently?	No fever. I measured my temperature and it was normal. I have not had chills or shivering either.
5	Where is the abdominal pain located?	Mainly around the umbilicus, with intermittent cramping that is not especially severe. I have not felt persistent pain specifically in the right lower abdomen.
6	Have you taken any medicine for this episode of diarrhea?	I did not take antidiarrheals or antibiotics on my own because I was worried that they might conflict with my previous gastrointestinal problem. I have mainly kept hydrated and eaten lightly.
7	What condition did the doctor diagnose for the chronic gastrointestinal problem you mentioned?	The doctor previously called it a chronic intestinal inflammatory problem. It has been present for about three years, and gastroenterology has followed me and adjusted treatment. I do not know the exact diagnostic name; it was not explained very clearly to me.

Turn	GPT-5.5 doctor	GPT-5.5 patient
8	Have you undergone colonoscopy during these three years?	Yes. I previously had colonoscopy during follow-up, including a biopsy. I do not remember the report details; I only remember being told that it showed chronic, active intestinal inflammation.
9	Which gastrointestinal medicines have you taken during these three years?	I have followed the gastroenterology clinic's long-term regimen, but I cannot recall every drug name. Treatment was adjusted at follow-up visits. I did not add antidiarrheals or antibiotics for this recent episode.
10	Do you have any other chronic conditions?	Yes. I have previously had iron-deficiency anemia and vitamin D deficiency, and I also have some anxiety. I have not been given a definite diagnosis of hypertension or diabetes. I also smoke.
11	Have you had surgery before?	I have not had major surgery or an operation for the intestinal problem. My prior procedures were mainly tests such as colonoscopy.
12	Do you have any medication allergies?	I have no known medication allergy. I have previously taken penicillins and cephalosporins without an obvious reaction. I have no food allergy either.
13	Which gastroenterology clinic has followed you?	I have mainly been followed by the gastroenterology department of a local tertiary hospital, returning at intervals. I did not bring the hospital name or all reports today, but I can provide them later.

G.2.3 Case Interaction: Examination, Tests, and Diagnosis

The doctor requested abdominal inspection, auscultation, palpation, and percussion together with eight test groups. The examination panel provides an M2-based view of all four positive signs, including the precordial and tobacco-staining findings beyond the abdominal request. The test panel includes only items corresponding to the eight listed groups.

Physical-examination findings. Temperature 36.8°C; pulse 75/min; respiratory rate 16/min; blood pressure $124/76\,\mathrm{mmHg}$; oxygen saturation 98%. The positive findings are right-lower-quadrant tenderness near McBurney's point, periumbilical tenderness, hyperdynamic precordium, and tobacco staining of the teeth or fingers, matching the four positive M2 signs.

Requested test groups. Complete blood count; C-reactive protein; stool routine examination and occult blood; stool culture; fecal calprotectin; liver and kidney function and electrolytes; ferritin and transferrin saturation; and 25-hydroxyvitamin D. Table 14 uses neutral test-item names and preserves the direction of each recorded predicate. *Not elevated* and *not decreased* mean that the corresponding high- or low-value predicate was absent; neither establishes a complete normal reference-range result. Numerical positive findings retain their recorded values.

Table 14: Case interaction display for the requested tests; absent predicates retain their directional meaning.

Test item	Observed result
Hematocrit	not decreased
Platelet count	not elevated
Mean corpuscular volume	74.93fL (decreased)
Fecal red blood cells	not elevated
Red-cell distribution width	16.83% (elevated)
Neutrophil ratio	80.18% (elevated)
Hemoglobin	118.08g/L (reduced)
Lymphocyte count	not decreased
Neutrophil count	7.6×10^9/L (elevated)
Mean corpuscular hemoglobin	not decreased
White blood cell count	not elevated
Fecal white blood cells	not elevated
Red blood cell count	not decreased
C-reactive protein	not elevated
Serum sodium	not decreased
Serum phosphate	not decreased
Bicarbonate	not decreased
Blood urea nitrogen	not elevated
Creatinine	not elevated
Blood calcium (qualitative status)	reduced or low-normal (abnormal)
Serum magnesium	not decreased
Alkaline phosphatase (mild-elevation criterion)	not met
Total calcium	1.9559mmol/L (decreased)
Serum chloride	not decreased
Fecal calprotectin	336.52μg/g (elevated)
Alkaline phosphatase (general-elevation criterion)	not met
Serum potassium	not decreased
Transferrin saturation	7.64% (decreased)
Serum ferritin	17.73ng/mL (decreased)
Serum 25-hydroxyvitamin D	21.49ng/mL (decreased)

Requested-test accounting (8 groups).

Table 15: Requested-test accounting (8 groups), with status labels from the archived case evaluation.

ID	Requested group	Status	Evidence
1	Complete blood count	fully returned	MCV, RDW, hemoglobin, and related findings returned
2	C-reactive protein	fully returned	elevation predicate absent (not elevated)
3	Stool routine examination and occult blood	partially returned	stool erythrocytes/leukocytes returned; occult-blood result absent
4	Stool culture	unavailable	no culture result provided
5	Fecal calprotectin	fully returned	$336.52\,\mu\mathrm{g/g}$
6	Liver/kidney function and electrolytes	fully returned	calcium, sodium, potassium, creatinine, and related findings returned
7	Ferritin and transferrin saturation	fully returned	both returned
8	25-hydroxyvitamin D	fully returned	serum 25-hydroxyvitamin D result returned

Final diagnosis. Active Crohn disease, probably involving the terminal ileum/ileocecal region, mild to moderate; accompanied by iron-deficiency anemia, vitamin D insufficiency/deficiency, and hypocalcemia. The case evidence includes a three-year history of chronic active intestinal inflammation, recurrent watery diarrhea, postprandial periumbilical cramping, right-lower-quadrant tenderness, elevated fecal calprotectin, microcytic iron-deficiency anemia, and reduced vitamin D. The diagnosis and the correct broad-category and specific-disease judgments were recorded in the archived evaluation.

Trace-boundary audit. No patient turn explicitly states the target-diagnosis label. Turns 7–8 disclose chronic intestinal inflammation and prior biopsy findings, which are strong clinically informative cues. Turns 6, 9, and 11–13 contain encounter-local contextual completions for medication, surgery, allergy, and care-site questions not represented as dedicated ledger fields. These completions are consistent with the frozen patient context and do not write new persistent facts back to M2. The history documents explicit-label non-disclosure; this case illustration does not establish perfect action gating or utterance-level ledger entailment.