PHYSIOLOGY AS TOPOLOGY: INTERVENABLE ORGAN GRAPHS FOR LONGITUDINAL PATIENT FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Do language models forecast patient trajectories by reasoning through physiology, or by fitting surface statistics? We introduce PhysioCircuit, a multi-agent agent system that represents each organ as an agent and each physiological relation as an editable mechanism card with direction, lag, and evidence. Its expert-defined graph is frozen at inference, enabling controlled edge interventions. We also introduce PhysioCircuit-Bench, a mechanism-controlled benchmark spanning three organ systems, treatment events, and real ICU admissions. We evaluate intervention-direction agreement (IDAR), which requires the affected organ to change in the expected direction within the expected lag window. The expert graph substantially outperforms all topology controls on edge-IDAR, including fully connected, random, and reversed graphs. Independent clinician adjudication supports the counterfactual results. PhysioCircuit transfers to new organ systems by configuration alone, without retraining, while preserving inspectable and editable relations.

1 INTRODUCTION

Large language models (LLMs), built on transformer architectures and instruction tuning (Vaswani et al., 2017; Brown et al., 2020; Bommasani et al., 2021; Ouyang et al., 2022; Touvron et al., 2023), now forecast patient trajectories from electronic health records (Achiam et al., 2023; Makarov et al., 2025; Li et al., 2025; Singhal et al., 2023; Nori et al., 2023; Moor et al., 2023; Khan et al., 2025). Yet predictive accuracy alone does not reveal whether a model followed the right physiology. The distinction is decisive when organs interact: a change in one organ reaches another with a specific direction and delay, and a forecaster that lacks this structure cannot be inspected, intervened on, or trusted.

The human body does not admit a precise mathematical model. Coupled organs, overlapping time scales, and pathological cross-talk defeat any equation-level simulator. We therefore build the simulation on LLM capabilities, which already integrate heterogeneous clinical knowledge and describe local state changes. The design goal is structural: an explicit physiology that constrains this flexibility and approximates real organ interactions as closely as the evidence supports. This builds on deep learning's representational foundations (Hinton et al., 2006).

Figure 1 shows why. After cholecystectomy, a patient develops bile reflux. Gastric injury worsens while intestinal fat absorption improves. A global predictor averages these conflicting signals into one mixed trend. PhysioCircuit keeps them separate, and the conflict runs through one editable edge, `bile_reflux_injury`. Deleting that single edge flips the gastric trajectory from worsening to improvement and moves the patient from decompensation to recovery, with every other input fixed. Details are in Appendix A.1.

Existing systems leave this requirement unsolved. Data-driven organ simulators store physiology in learned parameters (Chang et al., 2025; Silveira et al., 2025; Chen et al., 2024); protocol systems reason at the global level (Qu & Färber, 2026); graph models infer couplings that carry no explicit clinical meaning (Veličković et al., 2017); rule systems compile biology into fixed equations (Johnson et al., 2025; Wang et al., 2025). Hallucination and unverifiable mechanism claims remain open risks in clinical LLM systems (Ji et al., 2023). They share one limit: predictive accuracy is their only score, and it cannot reveal whether a model used the correct mechanism.

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

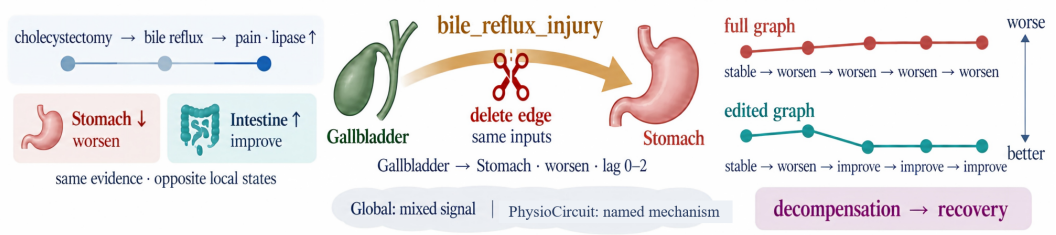


Figure 1: Same patient, one editable mechanism, different trajectory. A post-cholecystectomy patient develops bile reflux, worsening gastric injury, and improving intestinal fat absorption. Global aggregation collapses the conflicting local states into a mixed signal without exposing a named cause. PhysioCircuit routes the conflict through the editable mechanism edge `bile_reflux_injury`. Deleting only this edge changes the gastric trajectory from persistent worsening to improvement and shifts the macro-course from decompensation to recovery, while all other inputs remain fixed.

We introduce **PhysioCircuit**, a multi-organ LLM framework that makes physiology the topology. Each organ is an agent with its own inspectable knowledge base. Expert-defined edges carry named mechanisms, and typed messages pass deterministic gates before receiving organs accept, partially accept, or reject them and apply bounded state updates. The graph remains frozen at inference, so deleting, reversing, or delaying one edge directly tests that mechanism.

We make three contributions.

(1) **Physiology is the topology.** PhysioCircuit follows expert-specified physiological structure when defining agent connections. Each mechanism edge is explicit, editable, intervenable, and verifiable.
(2) **Every organ is an agent with its own knowledge.** Each organ is an LLM agent backed by a dedicated, inspectable knowledge base. The framework uses off-the-shelf LLMs without gradient updates.
(3) **Runtime-adjustable dynamics.** Organ functions and the physiological topology can be reconfigured on demand to simulate new physiological and pathological processes, without retraining the LLM.

Together, these results establish physiology as an inspectable and intervenable topology for clinical trajectory forecasting.

2 Related Work

Clinical simulation falls into five lines. Trajectory models forecast from EHR sequences (Makarov et al., 2025; Li et al., 2025; Zhou et al., 2025; Naveed et al., 2025), and their internals never expose an organ mechanism one can check or change. Organ-Agents trains organ-level simulators across nine systems (Chang et al., 2025), yet the physiology lives in weights after fine-tuning and reinforcement learning, so a new system means a new training run and no edge can be edited at inference. Vital Trace constrains longitudinal reasoning with a global protocol (Qu & Färber, 2026); propagation moves through patient-level rules, and no organ edge carries direction or lag. Graph models infer couplings from data (Kipf & Welling, 2016; Veličković et al., 2017), and an inferred edge names no mechanism, direction, or delay. Rule-based simulators encode biology in equations that freeze at compile time.

PhysioCircuit makes physiology the editable inference structure: organs are LLM agents, links are named mechanism cards, and every edge can be deleted, reversed, or delayed at evaluation time. Table 1 positions it against these lines, and Appendix A.2 details the comparison.

The decisive test is to intervene on these explicit edges and ask whether macro outcomes move as medicine predicts, using a two-source protocol that separates truth generation from mechanism cards (Section 4).

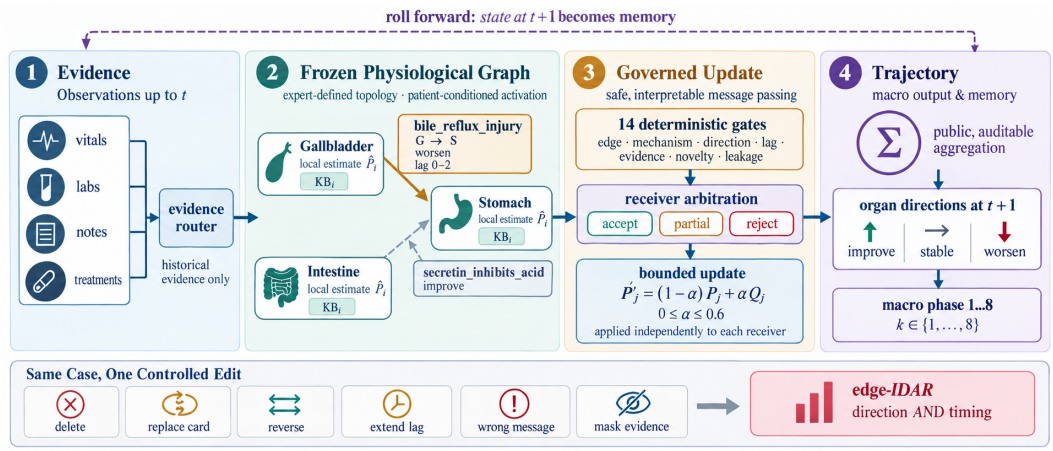


Figure 2: One inference step in PhysioCircuit. Each organ estimates locally from its own evidence; typed mechanism messages cross a frozen, expert-written organ graph under fourteen deterministic gates; surviving messages are arbitrated (accept / partial / reject) and applied through a bounded residual update; a public deterministic aggregation maps revised organ states to per-organ directions and one of eight macro phases. The graph supports intervention: every edge refers to a named mechanism card, and six operations act on one named physiological mechanism: delete edge, replace card, reverse direction, extend lag, inject wrong message, or mask source evidence.

Framework	Knowledge location	Macro course	Mechanism-level intervention	Transfer
Organ-Agents	model weights	data fitting + coordination	✗	↑ retrain
Vital Trace	protocol rules	rule-constrained reasoning	✓ partial	↑ rewrite
Graph learning	learned weights	learned graph readout	✗	↑ re-learn
Cell grammar	compiled rules	equation simulation	✗	↑ recompile
PhysioCircuit	**mechanism cards**	**typed propagation + arbitration**	✓	↓ config only

Table 1: Positioning of PhysioCircuit against the closest lines of work. The table compares where knowledge is stored, how the macro course is produced, relative support for semantic intervention, and transfer cost to a new physiological system (↓ = lower).

3 METHOD

3.1 PROBLEM SETUP

We study longitudinal forecasting over multiple organs. A case is an ordered sequence of observations $D = \{(o_1, t_1), \ldots, (o_N, t_N)\}$, where each o_k carries symptoms, laboratory results, interventions, and elapsed time. From the observed prefix, the model outputs each organ direction (improve, stable, or worsen) and one of eight patient-level phases.

The forecast runs on an expert-defined mechanism graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{M})$: nodes are organs, directed edges connect organs, and $\mathcal{M}$ attaches a mechanism card to every edge. Clinicians specify the graph and freeze it before evaluation, so each connection is a named physiological relation. A new system is obtained by changing the graph and case inputs while the inference code stays fixed. Figure 2 summarizes one inference step.

3.2 MECHANISM CARDS AND TYPED MESSAGES

A mechanism card fixes a source, a target, the variables involved, the allowed direction, the lag range, supporting evidence and counterevidence, and treatment bindings. It constrains what a mechanism can do during inference; it never supplies the answer to a case. The full schema is given in Table A2 (Appendix A.3).

Each organ first forecasts its own state from local evidence. When an edge activates, the source emits a typed message with sender, receiver, mechanism identifier, direction, strength, lag, and evidence IDs. Deterministic gates then check edge existence, direction, lag, evidence, duplicate detection, and verdict before delivery, and they bar future observations from the current step.

3.3 Arbitration and patient dynamics

The receiving organ accepts, partially accepts, or rejects each surviving message and records a reason. Let $\hat{P}_j$ be its independent estimate and $\hat{Q}_j$ be the estimate conditioned on the message. The update is bounded:

$$\hat{P}'_j = (1-\alpha)\hat{P}_j + \alpha\hat{Q}_j, \qquad 0 \le \alpha \le \alpha_{\max} < 1. \tag{1}$$

Rejection sets $\alpha = 0$; stronger acceptance permits a larger revision, while the cap prevents oscillation and over-correction across steps.

A public, deterministic aggregation maps the revised organ states to one of eight macro phases. It sits outside the gold labels, so every trajectory is reproducible; the same layer emits the uncertainty used for calibration.

3.4 Mechanism interventions and IDAR

Because every edge refers to a mechanism card, PhysioCircuit supports controlled intervention. We use six operations: delete a mechanism, replace its card, reverse its direction, extend its lag, inject a wrong message, or mask the source evidence. Each operation is applied to the same case with all other inputs fixed. The expected response is derived from the truth team's frozen executable generator, using the same rules that produce the gold trajectories; the mechanism team that writes the cards never sees or signs these labels. The expected response specifies both the affected organ and an admissible time window.

We report three related quantities. Direction Agreement (DA) is the fraction of interventions for which the affected organ moves in the expected direction. Timing Agreement (TA) is the fraction for which the first affected-organ change falls inside the specified window. Their joint version is intervention-direction agreement (IDAR):

$$\mathrm{DA} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[d_i = \hat{d}_i], \qquad \mathrm{TA} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[t_i \in W_i], \tag{2}$$

$$\mathrm{IDAR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[d_i = \hat{d}_i \ \wedge \ t_i \in W_i]. \tag{3}$$

Here d_i and t_i are the observed direction and first-change time, and $\hat{d}_i$ and W_i are the expected direction and window. A non-response fails both checks. Interventions reach the model through two channels scored by the same formulas. The graph channel edits the named card or edge; we call it edge-IDAR. The edge-less channel states the same mechanism descriptor as an input-level counterfactual while the case, model, and decoding stay fixed; we call it input-IDAR.

3.5 Inference procedure

Each rollout follows a fixed order: local organ forecasting → typed message emission → deterministic gating → arbitration with bounded updates → public macro-phase aggregation, after which the state rolls forward. Every message, evidence item, verdict, and update is logged, making each trajectory inspectable and reproducible apart from LLM sampling. Mechanism-card fields are listed in Appendix A.3; gate checks are detailed in Section 3.2 and the six intervention operations in Section 3.4.

4 PHYSIOCIRCUIT-BENCH: THE MECHANISM-CONTROLLED BENCHMARK

4.1 DESIGN AND INDEPENDENCE

Standard longitudinal records describe what happened, but they do not identify which mechanism caused a transition or what would happen if that mechanism were changed. We therefore build **PhysioCircuit-Bench**, a mechanism-controlled benchmark with three complementary components. Synthetic cases provide trajectories generated from explicit physiological rules. Treatment events test whether the same rules predict responses to physiological inputs. Real ICU trajectories test whether the framework remains useful on observed clinical data.

The synthetic component uses two independently built sources. One generates patient trajectories and labels; the other writes mechanism cards and their evidence. If the model and the labels encode the same rules, the benchmark cannot distinguish mechanism reasoning from rule memorization. The two sources stay independent, the answer key for every intervention remains with the truth team, and a perturbation probe confirms that PhysioCircuit answers from the cards it was given.

4.2 PHYSIOLOGICAL SYSTEMS

The benchmark covers three systems with separate organ sets, mechanism graphs, and 120 independent synthetic cases per system. Digestive is the source system with three organs. Pneumonia and cardiovascular-sepsis are six-organ extensions under the same protocol. The systems share no cases. The two six-organ systems share only the heart and kidney, and the digestive system shares no organ with them, so the partial overlap isolates reusable organ structure from system-specific graph content. Table 2 summarizes the systems.

Cases are authored by a truth team independent of the mechanism-card writers, independently reviewed by three physicians (Fleiss $\kappa \approx 0.72$), and split into 20 development and 100 blind-test cases per system. A third-party audit, a below-30% knowledge-overlap bound, and an 18-rule independence probe provide evidence that the benchmark does not reduce to rule memorization. Full details of case construction, the treatment layer, counterfactual pairs, and the independence analysis are provided in Appendix A.5.

System	Organs	Cards	Backbone	Cases	Treat. events	CF pairs
Digestive	3	12	Qwen3-235B-A22B-Instruct-2507	120	480	30
Pneumonia	6	14	Qwen3-235B-A22B-Instruct-2507	120	480	24
Cardiovascular-sepsis	6	12	Qwen3-235B-A22B-Instruct-2507	120	480	24

Table 2: The three systems of the benchmark. Cards denote the number of mechanism cards, i.e. the graph size to specify when the system changes. Cases, treatment events, and mechanism counterfactual pairs are reported per system. The three systems share code but no cases.

4.3 EVALUATION PROTOCOL

We compare PhysioCircuit with twelve component-matched baselines under matched cases, horizons, model version, and decoding settings (temperature 0.2, ≥ 3 runs; Table A3; Appendix A.5). The evaluation asks two questions. First, do methods predict local organ directions and macro phases accurately? Second, among methods exposing a semantic intervention interface, does changing a named edge move the affected organ in the expected direction and time? Graph-bearing systems are scored by edge-IDAR on named mechanism edits; edge-less systems by input-IDAR on controlled input-level counterfactuals (Appendix A.5.2).

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Main experiments use Qwen3-235B-A22B-Instruct-2507; cross-backbone checks use Mistral-Large-3-675B-Instruct-2512 and DeepSeek-V4-Pro-0813. All runs use temperature 0.2, top-p 1.0, and a 1,024-token limit. No model is fine-tuned. Ablations use the digestive system.

5.2 PREDICTION: PHYSIOCIRCUIT WINS WHERE ORGANS CONFLICT

Table 3 shows PhysioCircuit ahead in every system: 91.3% on digestive, 87.8% on pneumonia, 89.0% on cardiovascular-sepsis. It beats the global single agent by 9.2, 9.4, and 9.4 points, optimal voting by 7.8, 7.7, and 7.7, and the strongest learned-graph baseline (L) by 3.4, 4.7, and 3.6. The ordering repeats across systems, so the gain lives in the organ structure.

The gap over the strongest baseline (L) is reliable: paired case-level bootstrap 95% CIs are [2.1, 4.7], [3.3, 6.1], and [2.3, 4.9] (20,000 resamples, case as unit), and McNemar exact $p < 0.01$ on discordant pairs in all three systems. All results are mean $\pm$ std over three independent runs.

Method	Digest.	Pneum.	Cardio.	Tokens	Backbone
Global single agent	82.1±1.2	78.4±1.1	79.6±1.1	12.3/12.5/12.5K	Qwen
Single agent + RAG	83.4±1.1	79.9±1.1	81.1±1.1	14.1/14.4/14.4K	Qwen
Organ + optimal voting (H0)	83.5±1.1	80.1±1.0	81.3±1.0	15.6/15.9/15.9K	Qwen
Compute-matched, no graph	84.2±1.0	80.8±1.0	82.0±1.0	34.8/35.4/35.4K	Qwen
Implicit graph learning (L)	87.9±0.8	83.1±0.8	85.4±0.8	18.9/19.2/19.2K	Qwen
Vital Trace-style	85.6±0.9	82.0±0.9	83.2±0.9	22.4/22.8/22.8K	Qwen
PhysioCircuit (PC)	**91.3±0.7**	**87.8±0.7**	**89.0±0.7**	41.7/42.4/42.4K	Qwen

Table 3: Overall direction accuracy (%, mean $\pm$ std over three runs). Columns are digestive, pneumonia, and cardiovascular-sepsis. All main results use Qwen3-235B-A22B-Instruct-2507; cross-backbone checks use Mistral-Large-3-675B-Instruct-2512 and DeepSeek-V4-Pro-0813. Token entries follow the same system order.

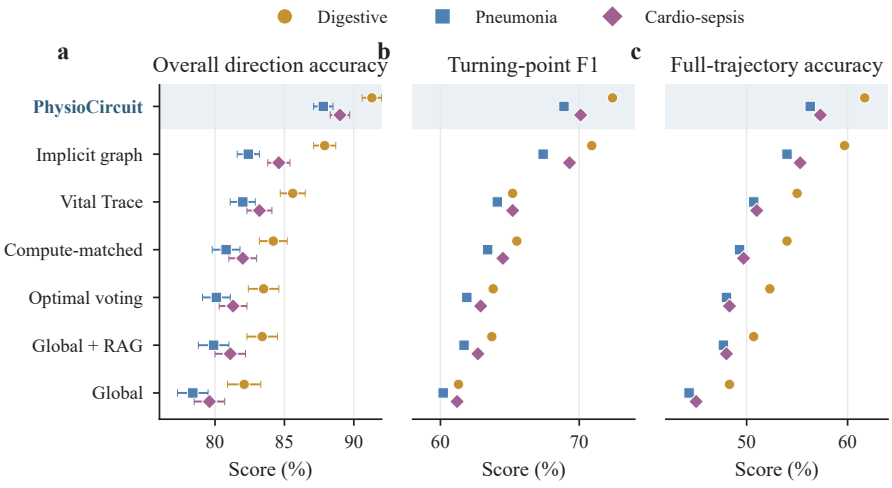


Figure 3: Predictive performance across three physiological systems. **a**, Overall direction accuracy; error bars show the reported standard deviation over three runs. **b**, Turning-point F1. **c**, Full-trajectory accuracy. Marker color and shape denote the organ system, and the shaded row highlights PhysioCircuit.

The gain concentrates where it should. Stratified by case difficulty (Figure 6; Appendix B.2), PhysioCircuit adds only 1.2, 0.8, and 0.8 points over the global agent on simple cases, where one organ dominates, and 13.6, 12.3, and 14.0 points of turning-point F1 on turning-point cases, where a macro transition needs arbitration. On conflict cases, PhysioCircuit leads implicit graph learning by 3.1 points on digestive (92.3 versus 89.2) and 3.6 on cardiovascular (87.6 versus 84.0). This

is exactly the regime in which a global average folds two competing organ trends into one wrong call, whereas PhysioCircuit keeps each organ tied to its own edge. Full-trajectory accuracy, which requires every organ at every step to be correct, shows the same ordering, with PhysioCircuit leading by up to 13 points (Figure 3c).

The gain is structure, not scale. The compute-matched no-graph baseline spends nearly three times the tokens (34.8K versus 12.3K for the global agent) and still trails PhysioCircuit by 7.1 points. RAG and optimal voting each recover a few points from the global agent, and neither reaches the graph.

5.3 THE TOPOLOGY CARRIES THE MECHANISM

Prediction accuracy cannot show that a model used the right organ edge. We intervene, following the causal-inference principle that a mechanism is verified by changing it and observing the effect (Schölkopf et al., 2021). Figure 4(a) puts the graph on trial. In the digestive ablation, edge-IDAR is 92.7% for the correct expert graph, 71.3% for a fully connected graph, 31.2% for a random graph, and 27.9% for a graph with reversed edges. The correct graph leads the fully connected one by 21 points, and the fully connected graph remains substantially above the random and reversed graphs. Randomizing or reversing the edges collapses the score to near-chance level. The result repeats in every system: 92.7/92.2/93.1 for the correct graph versus 29.8–32.4 for random graphs and 26.5–28.7 for reversed graphs (Table A13).

These expected responses carry an external clinical anchor. Three clinicians independent of the truth team adjudicated the expected direction and lag window for all 78 mechanism counterfactual pairs under blinded conditions; they reached majority consensus on 78.2% of pairs overall and only 63.3% of conflict pairs, and PhysioCircuit matches that independently adjudicated consensus on 91.8% of the answerable subset (Table A5, Appendix A.5.6). The headline score is therefore not measured only against the authors' own labels.

The DA/TA decomposition pins down the failure mode (Table A11). The correct graph scores 94.1 for direction and 95.2 for timing. The fully connected graph drops direction to 74.2: the extra edges fire on spurious paths and misdirect the receiver. The reversed graph is the diagnostic. Its timing agreement remains at 78.4, well above its direction score (33.5) but below the correct graph (95.2) and the fully connected graph (82.1). The timing machinery partially survives; the direction is inverted. This is the cleanest evidence that the model reads the topology, not a prior about the patient.

The six intervention operations all show the expected effect (Table A12). Delete, replace, reverse direction, extend lag, inject wrong message, and mask source evidence average 92.7/92.0/93.2 across systems. The two adversarial operations, inject wrong message and mask source evidence, score highest, 94.0/93.3/94.4 and 94.5/93.9/95.0, because they force the receiving organ to arbitrate instead of propagating blindly.

5.4 CARDS ALONE ARE NOT ENOUGH

The two IDAR channels answer different questions. Edge-IDAR edits named edges on the graph and measures how sensitive the system is to topological change. On the digestive system, the correct expert graph reaches 92.7%, a fully connected graph reaches 71.3%, a random graph reaches 31.2%, and a graph with reversed edges reaches 27.9%. The closer the graph matches real physiology, the more faithful the response; this ordering holds across all three systems. This gradient shows that response quality is determined by topological quality, not by patient-level priors in the model.

Input-IDAR states the same mechanism descriptor as input text to edge-less systems and measures how much of the mechanism response can be recovered from linguistic and card knowledge alone. On input-IDAR, the global agent scores 35.4%, 33.7%, and 36.0% across the three systems; optimal voting scores 38.2/36.5/39.1; no communication scores 37.0/35.8/38.3; Vital Trace-style scores 45.2/43.8/46.1; and implicit graph learning scores 58.4/56.7/59.2. The PhysioCircuit prompt-only control is an edge-less upper bound: it keeps the same LLM and card text but disables typed messages and graph propagation, reaching 62.1/59.8/63.4. Because the intervention reaches the model through different paths, the absolute values of the two channels are not directly comparable; however, the ordering within each channel is consistent across systems, and the qualitative conclusion

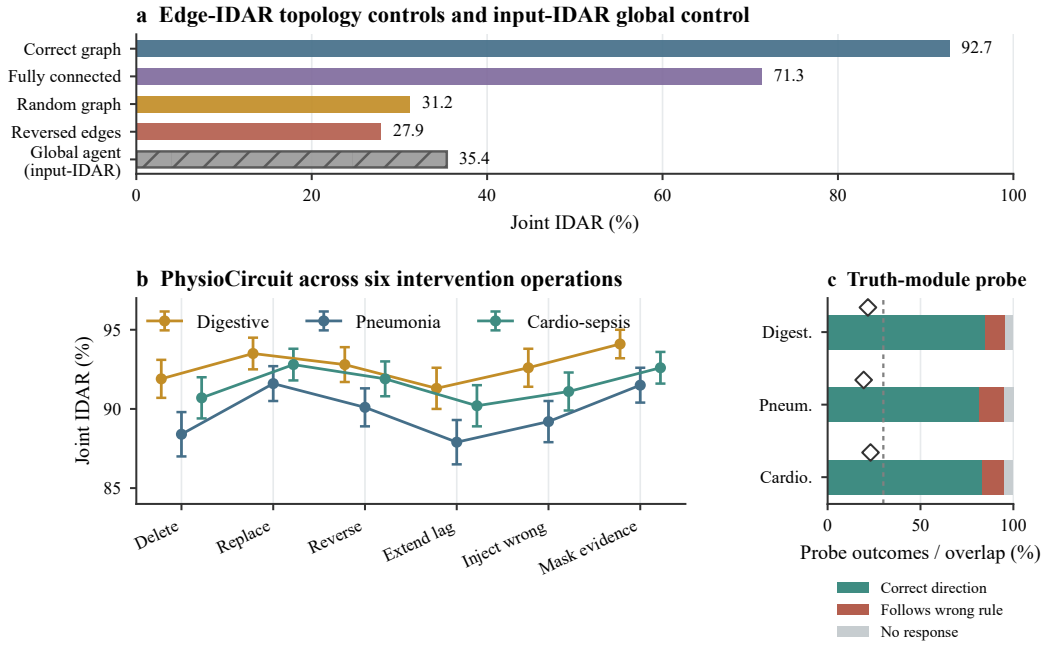


Figure 4: Semantic intervention faithfulness. **a**, Edge-IDAR for graph-bearing digestive topology controls (solid bars), with edge-less input-IDAR controls as hatched bars. A vertical divider separates the two channels. **b**, PhysioCircuit edge-IDAR across the three systems for six intervention operations. **c**, Outcome composition of the truth-module perturbation probe.

is robust: explicit graph structure reaches a level of mechanistic faithfulness that linguistic and card knowledge alone cannot.

The independence probe confirms that the model answers from the cards it was given. When we inject 18 wrong truth rules, PhysioCircuit retains the medically correct direction on 84.6/81.5/83.2% of probes, follows the injected rule on only 11.2/13.6/12.0%, and reaches a joint probe IDAR of 74.5/70.4/72.8%. A truth-retrieval control follows the injected rules 92.3% of the time, indicating that the gap is not explained by retrieval.

5.5 Each component earns its place

Ablations remove one part at a time on the digestive system (Figure 7; Appendix B.2). Removing the mechanism cards costs 4.6 points (91.3 to 86.7); replacing typed messages with free text costs 3.4 (to 87.9); optimal voting over organ states costs 7.8. Replacing the correct graph with a random one costs 10.8 points (91.3 versus 80.5, paired CI [8.9, 12.7], $p < 0.001$); reversed edges drop to 79.8; dropping organ communication drops to 81.6. A fully connected graph recovers part of the loss at 85.1 but stays 6.2 points below PhysioCircuit.

The gates do real work. The fourteen gates span six atomic checks: edge existence, direction, lag, evidence, duplicate detection, and verdict. The ablation removes four functional groups: (i) edge existence + duplicate detection, (ii) direction + lag, (iii) evidence, (iv) verdict. This cuts invalid-message interception from 87.6% to 74.2%, 68.9%, 66.1%, and 70.5%, respectively; removing all gates raises error spread from 6.8 to 26.3.

The bounded update similarly earns its place. With $\alpha_{\max} = 0.6$, five-step error growth is halved (0.42 versus 0.89) and direction flips drop from 7.1 to 3.1 per case. One-step accuracy is nearly flat across the bound, ranging from 88.8 at 0.3 to 88.4 at 1.0, so accuracy does not determine the choice; a looser bound of 1.0 accumulates error across steps.

Figure 5 shows that the written graph transfers to a new system by configuration alone, while the implicit graph learner must relearn its couplings from data.

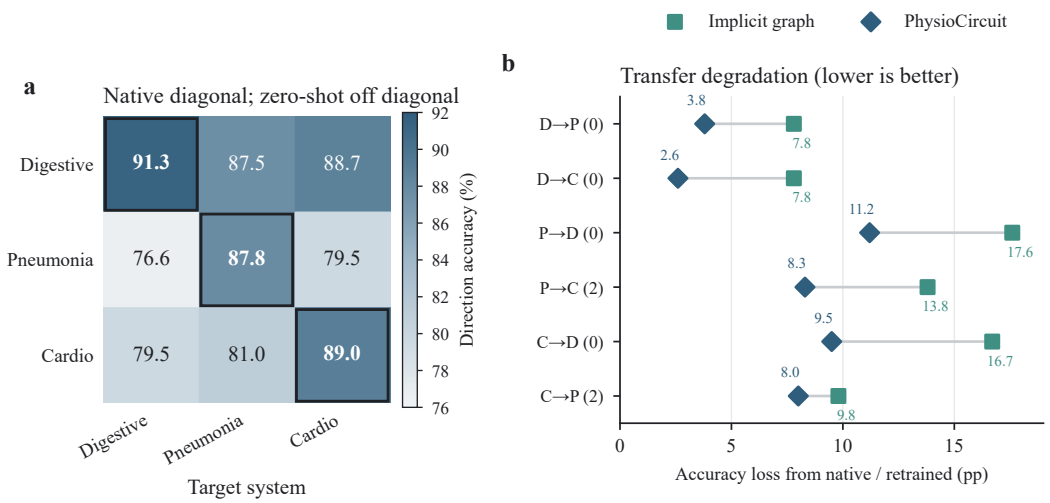


Figure 5: Cross-system transfer of explicit and learned graphs. **a**, Direction accuracy when each source graph is evaluated on each target system; outlined diagonal cells are native configuration-only runs and off-diagonal cells are zero-shot transfers. **b**, Accuracy loss relative to the native or retrained system for PhysioCircuit and implicit graph learning. Parentheses in the row labels give the number of shared organs.

5.6 Transfer, backbone, and real data

The written graph transfers. Rewritten pneumonia and cardiovascular graphs require 127 and 154 configuration lines and reach 87.5% and 88.7%, only 3.8 and 2.6 points below the digestive source (Figure 5). Zero-shot reuse loses 3.9 to 11.2 points and tracks organ overlap. The implicit graph learner pays 7.8 to 17.6 points to relearn its couplings, and PhysioCircuit leads it in all six transfer directions by 1.8 to 7.2 points.

The result holds across backbones. Across Qwen3, Mistral-Large-3, and DeepSeek-V4-Pro, PhysioCircuit edge-IDAR runs from 88.7 to 93.1, with per-system spread of 1.3, 3.5, and 2.8 points across the three systems, and the ordering inside every channel is unchanged (Table A9, Figure 8).

Treatment events bound to named mechanisms show the same direction and lag behavior (Figure 9, Appendix B.2).

Real records anchor the synthetic result. On 600 MIMIC-IV admissions (Johnson et al., 2023), PhysioCircuit reaches event-prediction AUROC 0.842/0.831/0.838, higher than implicit graph learning (0.834/0.821/0.827), Vital Trace-style (0.822/0.812/0.819), and the global agent (0.798/0.789/0.793) (Figure 10(a,b) in Appendix). Mechanism-direction consistency is 73.8/71.3/74.7, and timing agreement is 70.9/69.0/71.6. These real scores sit below synthetic IDAR because co-interventions go unrecorded and timestamps are coarser, yet the ordering holds, and deterministic aggregation reaches ECE 0.0305.

6 Conclusion

In this work we introduced PhysioCircuit, a multi-organ LLM framework in which physiology defines the topology: each organ is an agent and every relation is an editable mechanism card carrying direction, lag, evidence, and treatment bindings. It simultaneously improves predictive accuracy and mechanism faithfulness on PhysioCircuit-Bench, three organ systems with synthetic cases, treatment counterfactuals, and a 600-admission MIMIC-IV pillar.

The correct expert graph reaches 92.7% edge-IDAR, versus 71.3% fully connected, 31.2% random, and 27.9% reversed. A prompt-only control reaches only 62% input-IDAR, showing that card text alone cannot substitute for explicit topology. The written graph transfers to a new system by configuration alone, with no retraining. Physiology is thus an inspectable and intervenable topology for clinical trajectory forecasting.

AI USE STATEMENT

LLM tools assisted language editing and programming; all numbers, experiments, clinical content, and code were reviewed by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

Human subjects. The 600 ICU admissions are de-identified MIMIC-IV records under its DUA, and no patient-identifiable information was retained at any stage. The three clinicians who rated trajectories under blinded conditions and independently adjudicated the 78 counterfactual expectations did so on synthetic and de-identified material. They were not members of the truth team, gave informed consent before participation, and were reimbursed at market rate, above the local hourly minimum; the adjudication task took approximately two hours per clinician.

Clinical use. PhysioCircuit is a research artifact; trajectories are not cleared for clinical decision-making. The policy counterfactual is evaluated against card-defined expectations and is not treatment guidance.

REPRODUCIBILITY STATEMENT

We release the synthetic benchmark, mechanism cards, two-source truth module, and intervention scripts under a permissive license; the MIMIC-derived pillar cannot be redistributed but is documented for reproduction. Code and benchmark are available at `https://anonymous.4open.science/r/physiocircuit-B806/`.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

II Apache. Apache ii: A severity of disease classification system. *Published online*, 1985.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Rihao Chang, He Jiao, Weizhi Nie, Honglin Guo, Keliang Xie, Zhenhua Wu, Lina Zhao, Yunpeng Bai, Yongtao Ma, Lanjun Wang, et al. Organ-agents: Virtual human physiology simulator via llms. *arXiv preprint arXiv:2508.14357*, 2025.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. From llm reasoning to autonomous ai agents: A comprehensive review. *IEEE Access*, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Tiantian He, An Zhao, Elinor Thompson, Anna Schroder, Ahmed Abdulaal, Frederik Barkhof, and Daniel C Alexander. Llm-guided spatio-temporal disease progression modelling.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Yong Huang, Zhongqi Yang, and Amir Rahmani. Mimic-sepsis: A curated benchmark for modeling and learning from sepsis trajectories in the icu. In *2025 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–7. IEEE, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

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

Jeanette AI Johnson, Daniel R Bergman, Heber L Rocha, David L Zhou, Eric Cramer, Ian C Mclean, Yoseph W Dance, Max Booth, Zachary Nicholas, Tamara Lopez-Vidal, et al. Human interpretable grammar encodes multicellular systems biology models to democratize virtual cell laboratories. *Cell*, 188(17):4711–4733, 2025.

Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. Care-ad: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*, 8(1):541, 2025.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Nikita Makarov, Maria Bordukova, Papichaya Quengdaeng, Daniel Garger, Raul Rodriguez-Esteban, Fabian Schmich, and Michael P Menden. Large language models forecast patient health trajectories enabling digital twins. *npj Digital Medicine*, 8(1):588, 2025.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 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.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Chantal Pellegrini, Ege Özsoy, David Bani-Harouni, Nassir Navab, and Matthias Keicher. Ehr2path: Comprehensive pathway-level modeling of longitudinal patient trajectories from multimodal electronic health records.

Zhan Qu and Michael Färber. Vital trace: Protocol-constrained patient-state reasoning for longitudinal clinical trajectories. *arXiv preprint arXiv:2602.12833*, 2026.

Weijieying Ren, Jingxi Zhu, Zehao Liu, Tianxiang Zhao, and Vasant Honavar. A comprehensive survey of electronic health record modeling: From deep learning approaches to large language models. *arXiv preprint arXiv:2507.12774*, 2025.

Moritz Schaefer, Stephan Reichl, Rob ter Horst, Adele M Nicolas, Thomas Krausgruber, Francesco Piras, Peter Stepper, Christoph Bock, and Matthias Samwald. Large language models are universal biomedical simulators. *BioRxiv*, pp. 2023–06, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Andre Lehdermann Silveira, Rodrigo da Rosa Righi, and Cristiano Andre Da Costa. Multi-agent systems for clinical decision support: A systematic review. *Applied soft computing*, pp. 114447, 2025.

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.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Aoqi Wang, Jiajia Liu, Jianguo Wen, Yangyang Luo, Zhiwei Fan, Liren Yang, Xi Hu, Ruihan Luo, Yankai Yu, Sophia Li, et al. A multi-layered framework for modeling human biology: From basic ai agents to a full-body ai agent. *ArXiv*, pp. arXiv–2508, 2025.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*, 2023.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China information sciences*, 68(2):121101, 2025.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Xinyang Zhou, Yongyong Ren, Qianqian Zhao, Daoyi Huang, Xinbo Wang, Tingting Zhao, Zhixing Zhu, Wenyuan He, Shuyuan Li, Yan Xu, et al. An llm-driven multi-agent debate system for mendelian diseases. *arXiv preprint arXiv:2504.07881*, 2025.

A Additional Method and Benchmark Details

A.1 Introductory case study

The case used in Figure 1 is a mechanism-conflict example from the digestive benchmark. After cholecystectomy, the patient develops bile reflux, epigastric pain, and rising lipase. The stomach shows worsening mucosal injury while intestinal fat absorption improves. A global predictor can collapse these opposing local states into a mixed signal. PhysioCircuit keeps them separate: BILE_REFLUX_INJURY (gallbladder→stomach, worsen, lag 0–2) explains the gastric worsening, while intestinal improvement follows a separate pathway. The macro state therefore enters CONFLICT_EMERGES. Table A1 summarizes the observed signals and mechanism-grounded reading.

Observed signals	Local organ states	Mechanism-grounded reading
Day 0: cholecystectomy. Day 3: bile reflux, epigastric pain, rising lipase	Stomach: mucosal injury worsening. Intestine: fat absorption improving	BILE_REFLUX_INJURY (gallbladder→stomach, worsen, lag 0–2) explains the gastric worsening; intestinal improvement is a separate pathway. Macro state enters CONFLICT_EMERGES.

Table A1: Mechanism-conflict case used in the introductory demo. Local organ states point in opposite directions; a faithful macro-course reading composes them through typed mechanism edges with direction and lag.

In the unperturbed run, the macro course is BASELINE, LOAD_ACCUMULATION, CONFLICT_EMERGES, TRANSITION, and DECOMPENSATION. Deleting only the `bile_reflux_injury` edge from gallbladder to stomach changes the gastric sequence from stable, worsen, worsen, worsen, worsen to stable, worsen, improve, improve, improve. The macro course becomes BASELINE, LOAD_ACCUMULATION, COMPENSATION, RECOVERY, RESIDUAL_STATE. All other inputs remain fixed, isolating the effect of one editable mechanism.

A.2 Extended related work

LLMs for clinical reasoning and trajectory forecasting. Large language models have been used for electronic-health-record risk scoring and trajectory forecasting (Makarov et al., 2025; Li et al., 2025; Ren et al., 2025; Pellegrini et al.; Huang et al., 2025; Yang et al., 2024; Singhal et al., 2025; Chen et al., 2024; Silveira et al., 2025). DT-GPT forecasts clinical trajectories and CARE-AD coordinates specialized agents over longitudinal notes (Makarov et al., 2025; Li et al., 2025). These systems evaluate sequence-level prediction; PhysioCircuit additionally exposes organ mechanisms for inspection.

Multi-agent LLM systems in medicine. Multi-agent collaboration supports medical debate, diagnosis, and simulation (Zhou et al., 2025; Williams et al., 2023; Li et al., 2023; Ferrag et al., 2026; Wang et al., 2023). Retrieval, chain-of-thought, self-consistency, self-refine, reflection, and tool use provide general coordination primitives (Lewis et al., 2020; Wei et al., 2022; Wang et al., 2022; Madaan et al., 2023; Nye et al., 2021; Kojima et al., 2022; Zelikman et al., 2022; Yao et al., 2022; Shinn et al., 2023; Schick et al., 2023; Park et al., 2023; Hong et al., 2024; Wu et al., 2023; Xi et al., 2025). PhysioCircuit uses these primitives with typed messages whose direction, lag, and evidence are defined by physiological mechanisms.

LLM-based physiological simulation. Text-based simulators treat LLMs as biomedical simulators without manual tuning or physics-based equations (Schaefer et al., 2023; Liu et al., 2026). Organ-Agents trains organ-level simulators across nine systems, but the physiological knowledge lives in weights after fine-tuning and reinforcement learning (Chang et al., 2025). Full-Body AI Agent integrates multiple biological levels but does not provide mechanism-level intervention or cross-system transfer evidence (Wang et al., 2025).

Protocol-constrained reasoning. Vital Trace constrains longitudinal reasoning with a curated global protocol and patient-state memory, coordinated by Router, Reasoner, Auditor, and Stew-

Field	Semantics
mechanism_id	Unique identifier of the mechanism
source, target	Endpoints of the directed edge
source_variables, target_variables	Variables involved on each side
allowed_directions	Whether the mechanism worsens, improves, or both
lag_steps	Allowed delays between activation and effect
activation_evidence	Conditions that activate the mechanism
counterevidence	Conditions that suppress the mechanism
treatment_bindings	Treatments that activate or suppress the mechanism
mechanism_description	Natural-language description of the physiology

Table A2: Fields of a mechanism card. A card constrains how a mechanism can act during inference; the case itself is always answered by the rollout.

ard agents (Qu & Färber, 2026). Its state transitions are patient-level protocol rules; the unit of propagation is not a typed organ interaction with explicit direction and lag.

Mechanism knowledge and graph learning. Cell behavior grammars compile natural-language rules into mathematical models, but the compiled rules cannot be arbitrated during inference (Johnson et al., 2025). Graph convolutional and attention networks learn adjacency structures end to end from data (Kipf & Welling, 2016; Veličković et al., 2017); recent exploratory work also probes LLM guidance for discovering such propagation graphs (He et al.). Learned edges carry coupling but no named mechanism semantics, so a user cannot delete, reverse, or delay one physiological process.

A.3 Mechanism card fields

A.4 Implementation details

Macro phases. The deterministic aggregation emits one of eight phases: BASELINE, LOAD_ACCUMULATION, CONFLICT_EMERGES, TRANSITION, DECOMPENSATION, COMPENSATION, RECOVERY, and RESIDUAL_STATE. Phase assignment follows a fixed priority rule over per-organ directions and confidences: worsening organs with confidence above 0.6 advance the phase toward DECOMPENSATION; stable or improving organs advance it toward RECOVERY; conflicting directions across organs produce CONFLICT_EMERGES or TRANSITION. The rule is pure and stateless apart from the previous phase, so the same organ-state vector and previous phase always yield the same phase.

Arbitration and α. Each receiving organ arbitrates incoming messages into accept, partial accept, or reject. Rejection sets $\alpha = 0$ (no state change); partial accept sets $0 < \alpha < \alpha_{\max}$; accept sets $\alpha = \alpha_{\max} = 0.6$. The bounded residual update is $\Delta s = \alpha \cdot \hat{\Delta}$, where $\hat{\Delta}$ is the message-proposed direction and magnitude.

Fourteen gates. The fourteen deterministic gates span six atomic checks: edge existence (2 gates: source–target edge present, edge active at current step), direction (3 gates: direction in card-admissible set, direction not contradicted by evidence, direction monotonic with lag), lag (2 gates: current step within card lag window, lag not exceeded), evidence (3 gates: at least one evidence ID cited, evidence IDs exist in case record, no counterevidence active), duplicate detection (2 gates: same mechanism not already delivered this step, message hash not seen in current rollout), and verdict (2 gates: verdict field well-formed, verdict consistent with direction).

A.5 Benchmark construction details

A.5.1 Two-source independence protocol

Two teams construct each system from disjoint sources. The truth team writes state-transition rules from clinical guidelines and textbooks, freezes them into an executable generator, produces the gold

ID	Method	Rules out
A	Global single agent (no RAG)	organ decomposition has no value
B	Single agent + mechanism RAG	gains come only from external knowledge
C	Independent organs, no communication	gains are only from splitting organs
D	Multi-agent, communication without gates	program gates are unnecessary
E	Organ prediction + optimal voting (H0)	macro is just a combination of local states
F	Compute-matched multi-agent without graph	gains are only from more calls
G	Free-text multi-agent (no mechanism codes)	typed mechanism messages are unnecessary
H	Random graph	correct topology is unnecessary
I	Reversed / fully connected / random graph	direction and sparsity do not matter
J	No gates / unbounded residual	gating and bounded update are unnecessary
K	Vital Trace-style protocol system	patient-level protocol suffices
L	Implicit graph learning (learned graph, GAT)	explicit mechanism cards are replaceable

Table A3: The twelve baselines (A to L) and the assumption each one rules out. The three decisive contrasts are communication (independent organs vs. PhysioCircuit), composition (voting vs. PhysioCircuit), and topology (random graph vs. PhysioCircuit).

trajectory labels, and also authors the expected direction and lag window used to score every intervention. The mechanism team writes cards from pathophysiology literature and organ-interaction reviews without seeing the truth artifact or the scoring labels. A third-party coordinator audits the split, resolves disagreements, and records the audit trail. Knowledge overlap is estimated by two independent raters and remains below the 30% protocol bound in all systems (21.7%, 19.4%, and 23.1%). Because the expected direction and lag window are still authored inside the author team, three independent clinicians separately adjudicated the same expectations under blinded conditions on the 78 mechanism counterfactual pairs; on the pairs reaching majority consensus, their expected direction agrees with the truth-team label on 95.1% of cases (§A.5.6).

The independence probe injects K medically wrong rules into the truth generator, freezes the modified generator, and regenerates a probe set. For example, the probe may assert that bile reflux improves the gastric mucosal barrier. If intervention responses follow the injected rule, the model is copying the truth generator; if they preserve the medically correct direction, the cards provide independent content. We use $K = 18$ in the reported probe and require at least 80% medically correct responses. A truth-retrieval baseline follows the injected rules on 92.3% of cases, confirming that the probe detects truth-module mimicry.

A.5.2 Input-level counterfactual protocol for edge-less systems

For systems that do not expose a named organ-to-organ edge, we administer the same intervention descriptor as a controlled input-level counterfactual. The protocol holds the patient case, model, temperature, output limit, and retrieval scope fixed. The intervention descriptor is a natural-language sentence that states the mechanism, its expected direction, and its admissible lag window (e.g., "Assume that bile no longer refluxes from the gallbladder to the stomach; the expected gastric effect is improvement within 0–2 steps."). The model then produces a future trajectory under this descriptor, and the same DA / TA / joint IDAR scoring is applied.

Three operational definitions fix comparability. First, *non-response* is defined as the affected organ's predicted trajectory being identical to its baseline run (no directional change in any step). Second, *first-change time* is read from the model's step-by-step rollout; models that emit the full future sequence in a single generation are scored by reading the predicted first step that differs from baseline. Third, the same $N = 180/210/190$ intervention instances are used for both edge-IDAR and input-IDAR in each system, so denominators are matched.

Stratum	Share	Design purpose
Simple or near-stable	30%	single-organ competence
Cross-organ conflict	40%	mechanism arbitration
Turning point and decompensation	20%	macro transitions
Intermittent and non-monotonic	10%	long-range structure

Table A4: Stratification of the blind test set, identical in all three systems. Conflict cases carry the weight of the evaluation because they are where mechanism arbitration should matter most.

A PhysioCircuit prompt-only control runs the same LLM and uses the same card knowledge but disables typed messages, graph propagation, and arbitration; the intervention descriptor is injected directly into the single-prompt context. This control separates the contribution of the graph channel from those of model knowledge and card text.

A.5.3 SYNTHETIC CASES

Each system contributes 120 cases (20 development and 100 blind test) with five to eight future time points and three robustness variants: missing observations, noisy values, and reordered descriptions. Three physicians review each truth independently; disagreements are resolved by majority vote. Agreement by Fleiss κ is 0.74, 0.71, and 0.73 for digestive, pneumonia, and cardiovascular cases. Fingerprints over mechanism, severity, temporal pattern, evidence, intervals, and symptoms remove near-duplicates; fingerprints are computed within each system so that cases from different systems cannot collide.

The blind-test strata are identical across systems (Table A4). Gold truth, design cards, expert notes, and future time points are excluded from prompts; retrieval uses stable mechanisms and current patient memory only. Predicted nodes are marked `predicted=true`.

A.5.4 TREATMENT EVENTS AND TREATMENT COUNTERFACTUALS

The treatment layer complements edge interventions with physiological inputs. Each system contributes 480 pharmacological, surgical, or supportive events, four per case. Every event binds to a card through `treatment_bindings`, specifies the expected mechanism direction, and defines a lag window. We score both direction agreement and whether the response appears within that window. The layer also contains 360 treatment counterfactual pairs, 120 per system; each pair holds the case fixed, applies one treatment in the treated branch, withholds it in the control branch, and records a field identifying the single changed factor so that silent multi-variable edits cannot pass.

A.5.5 COUNTERFACTUAL PAIRS

Each mechanism counterfactual changes one key factor and stays within its system. The digestive set has 30 pairs across five families: high-fat diet with partial biliary obstruction; bile reflux with improving fat absorption; gastric hyperacidity with later loss of intestinal compensation; transient obstruction followed by resolution; and post-cholecystectomy bile insufficiency with gastrointestinal feedback. Pneumonia has 24 pairs across airway-clearance failure, alveolar flooding with vascular loading, systemic inflammatory drive, and renal retention with uraemic airway injury. Cardiovascular-sepsis has 24 pairs across vasoplegia with myocardial depression, renal and hepatic coupling, coagulopathy with glycocalyx shedding, and recovery with restitution. Every pair records the changed factor, and no pair crosses a system boundary.

A.5.6 EXTERNAL CLINICAL ANCHORING OF COUNTERFACTUAL EXPECTATIONS

The expected direction and lag window used to score IDAR are authored by the truth team. To check that these expectations are not artifacts of the authors' own frozen rules, the three external clinicians from §B.1.2 independently adjudicated all 78 mechanism counterfactual pairs. These clinicians were not members of the truth team and wrote none of the generator rules. For each pair, they saw only the de-identified case and the single changed factor; the truth-team label, the PhysioCircuit output,

Stratum	N	Direction Fleiss κ	Window ($\pm$1 step)	Majority consensus	PhysioCircuit on consensus
Non-conflict	48	0.79	85.4%	87.5%	93.0%
Conflict	30	0.51	53.3%	63.3%	89.5%
All pairs	78	0.67	73.1%	78.2%	91.8%

Table A5: External clinical anchoring of counterfactual expectations. Three clinicians independent of the truth team judged expected direction and lag window under blinded conditions for all 78 mechanism counterfactual pairs. Direction agreement is Fleiss κ over three raters; window agreement allows a one-step tolerance; majority consensus requires a majority of the three raters to agree on both direction and window. On consensus pairs the clinicians' majority direction agrees with the truth-team label at 95.1%. The last column is PhysioCircuit edge-IDAR restricted to the consensus subset.

and the mechanism card text were hidden. They judged two quantities: the expected direction of the affected organ (worsen, improve, or no change) and the expected lag window in steps.

Table A5 reports the result. Direction agreement across the three raters, measured by Fleiss κ, is 0.67 overall, 0.79 on non-conflict pairs, and 0.51 on conflict pairs. Lag-window agreement within one step is 73.1% overall. A majority of the three raters agree on both direction and window for 61 of 78 pairs (78.2%); on conflict pairs, where two organ pathways pull in opposite directions, majority consensus falls to 19 of 30 (63.3%). On the 61 consensus pairs, the clinicians' majority expected direction agrees with the truth-team label on 58 of 61 (95.1%), confirming that the scoring expectations are clinically reasonable rather than artifacts of the generator. Of the 17 pairs without majority consensus, nine disagree only on the lag window while agreeing on direction, five split on direction, and three involve competing multi-pathway responses for which the direction itself is unclear. These are intrinsically ambiguous cases, not scoring disagreements.

Restricting PhysioCircuit's edge-IDAR to the 61 consensus pairs gives 91.8%, close to the 92.7% headline. Where independent clinicians can agree on the expected response, PhysioCircuit recovers it; where clinicians split, the expected label is itself under-specified.

A.5.7 Real longitudinal pillar

Because real records lack mechanism-level truth, the real pillar evaluates two practice-checkable tasks: event prediction, asking whether an organ will require support, deteriorate, or recover in the coming window; and mechanism-direction consistency after a documented treatment change. It uses 600 MIMIC-IV ICU admissions (Johnson et al., 2023), 200 per system, selected for complete records and documented treatment changes. Each admission contributes ten treatment-change observations, yielding 2,000 direction observations per system. Scores are stratified by treatment family and include a lag-window check. This pillar anchors the synthetic intervention results in observed clinical data.

B Additional Experiments and Analyses

B.1 Secondary experiments

B.1.1 Treatment response and response timing

Table A6 reports the treatment layer. Across 1,440 synthetic events, mechanism direction accuracy is 98.1%, 97.5%, and 98.5% for the three systems. On 360 treatment counterfactual pairs, response direction exceeds 90% and lag-window agreement is 88.3%, 86.7%, and 87.5%. In the 6,000 real treatment-change observations, direction consistency is 73.8%, 71.3%, and 74.7%; timing is 70.9%, 69.0%, and 71.6%. Pharmacological changes are most consistent in all systems, surgical changes are weakest on direction, and supportive care is weakest on timing. Supportive interventions often act on several mechanisms at once, so a single-card lag window only partly describes the observed response. The synthetic layer therefore provides a controlled temporal test, while the real layer exposes the additional variability present in practice. The ablations below isolate which components support these responses.

Panel A: synthetic treatment layer

System	Events	Mechanism direction	Pairs	Response direction	Lag in window
Digestive	480	98.1%	120	91.7%	88.3%
Pneumonia	480	97.5%	120	90.8%	86.7%
Cardiovascular-sepsis	480	98.5%	120	91.7%	87.5%

Panel B: real treatment-change observations (MIMIC-IV)

System	Family	Observations	Direction consistency	Lag in window
Digestive	pharmacological	900	76.0%	72.2%
	supportive	500	72.4%	67.6%
	surgical	600	71.7%	71.7%
Pneumonia	pharmacological	900	74.0%	70.0%
	supportive	500	70.4%	67.2%
	surgical	600	68.0%	69.0%
Cardiovascular	pharmacological	900	77.0%	73.7%
	supportive	500	73.2%	67.0%
	surgical	600	72.5%	72.3%

Table A6: Treatment response and response timing. Panel A scores synthetic treatment events and treatment counterfactual pairs against the mechanism cards. Panel B stratifies the real treatment-change observations by treatment family and adds the timing check. Drug changes are the most consistent family, surgical changes the least consistent on direction, and supportive changes the weakest on timing in all three systems.

Method	Likert mean	Likert std	Exact rater agreement	Fleiss κ
PhysioCircuit	**4.31**	0.65	88.6%	0.77
Implicit graph learning (L)	4.00	0.72	82.3%	0.76
Global single agent	3.54	0.83	74.3%	0.74

Table A7: Clinical expert evaluation. Three clinicians rate 80 trajectories per method under blinded conditions on a five-point scale (240 ratings per method). Trajectories are drawn from the digestive system so that raters compare methods under a fixed physiological setting. Exact rater agreement is the fraction of rater pairs that give the same score to the same trajectory. Fleiss κ is computed on the binary judgement of whether a trajectory is clinically acceptable (rating of 4 or above).

B.1.2 Clinical expert trajectory evaluation

Expert readers assess whether a predicted course is clinically reasonable. Three clinicians rated 80 digestive trajectories from each method under blinded conditions on a five-point Likert scale (Table A7). PhysioCircuit scores 4.31 (standard deviation 0.65), compared with 4.00 (0.72) for implicit graph learning and 3.54 (0.83) for the global agent. Exact agreement is 88.6%, 82.3%, and 74.3%; Fleiss κ for binary acceptability is 0.77, 0.76, and 0.74. The difference between PhysioCircuit and implicit graph learning is modest in accuracy but more pronounced in clinical ratings and rater agreement. Each PhysioCircuit trajectory also names the mechanisms that produced it, so a reviewer can trace a rating to a card.

B.1.3 Policy-level counterfactual

Single-edge interventions test graph organization; a clinical decision also asks what happens when treatment is given now or later. We built 60 policy pairs, 20 per system, by applying the indicated treatment at the earliest eligible step or two to four steps later. The expected macro response is derived from the same frozen generator rules that score all other interventions, not from the cards the model reads. The early branch matches the expected response in 91.7% of pairs and resolves in 2.40 steps, versus 48.3% and 4.70 steps for delayed treatment (Table A8). Matched real groups have comparable APACHE II scores (Apache, 1985) (18.17 vs. 18.90; 95% bootstrap interval for the difference $[-2.30, 0.77]$), reducing a baseline-severity explanation. This experiment tests a decision-shaped counterfactual through both response direction and response timing.

Strategy	Pairs	Expected macro response	Steps to resolution	APACHE II
Early intervention	60	91.7%	2.40	18.17
Delayed intervention	60	48.3%	4.70	18.90

Table A8: Policy-level counterfactual. Each pair holds the case fixed and moves the treatment in time. Expected macro response is the fraction of pairs whose macro outcome matches the clinically expected change. The APACHE II difference between the matched real groups is -0.73 with a 95% bootstrap interval of $[-2.30, 0.77]$.

B.1.4 A CLINICAL WHAT-IF WALKTHROUGH

Case `PC-D-014` (the post-cholecystectomy example of Appendix A.1) illustrates the audit path. In the unperturbed run, the macro course is BASELINE, LOAD_ACCUMULATION, CONFLICT_EMERGES, TRANSITION, and DECOMPENSATION. Deleting only the `bile_reflux_injury` edge changes the gastric sequence from stable, worsen, worsen, worsen, worsen to stable, worsen, improve, improve, improve. The macro course becomes BASELINE, LOAD_ACCUMULATION, COMPENSATION, RECOVERY, RESIDUAL_STATE. All other inputs remain fixed, so the example isolates the effect of one editable mechanism.

B.2 SECONDARY RESULTS AND ABLATIONS

The main text reports headline numbers; all secondary results, the backbone sweep, the per-intervention-type breakdown, component ablations, cross-system transfer, the real longitudinal pillar, downstream retention, calibration, and cost are detailed below.

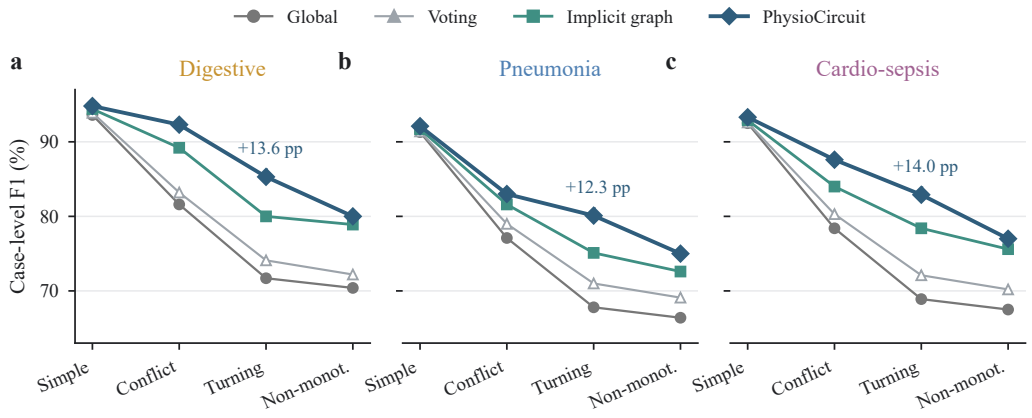


Figure 6: Performance by mechanism-complexity stratum. Direction accuracy is shown for global prediction, optimal voting, implicit graph learning, and PhysioCircuit in the digestive (**a**), pneumonia (**b**), and cardiovascular-sepsis (**c**) systems. The annotated difference is the PhysioCircuit gain over the global single agent on turning-point cases. The gain is smallest on simple cases and largest on turning-point cases.

B.2.1 A CONTROLLED BACKBONE SWEEP ACROSS ALL THREE SYSTEMS

All main systems use the same Qwen3 backbone (Yang et al., 2025), while Mistral-Large-3 and DeepSeek-V4-Pro (Xu et al., 2026) provide cross-backbone checks. We run every backbone on every mechanism graph with identical cases, prompts, and temperature. The full 3×3 sweep in Table A9 reports the observed between-backbone variation; these values are robustness checks, not a replacement for the topology controls below. The mechanism claim is based on matched graph-bearing interventions, not on comparison with systems for which IDAR is undefined.

Figure 6 shows the stratified breakdown. PhysioCircuit gains little on simple cases, where one organ dominates, and most on turning-point cases, where macro transitions require arbitration. Its gain over the global agent rises from 1.2, 0.8, and 0.8 points on simple cases to 13.6, 12.3, and 14.0 points of turning-point F1 on turning-point cases. On conflict cases, PhysioCircuit leads implicit

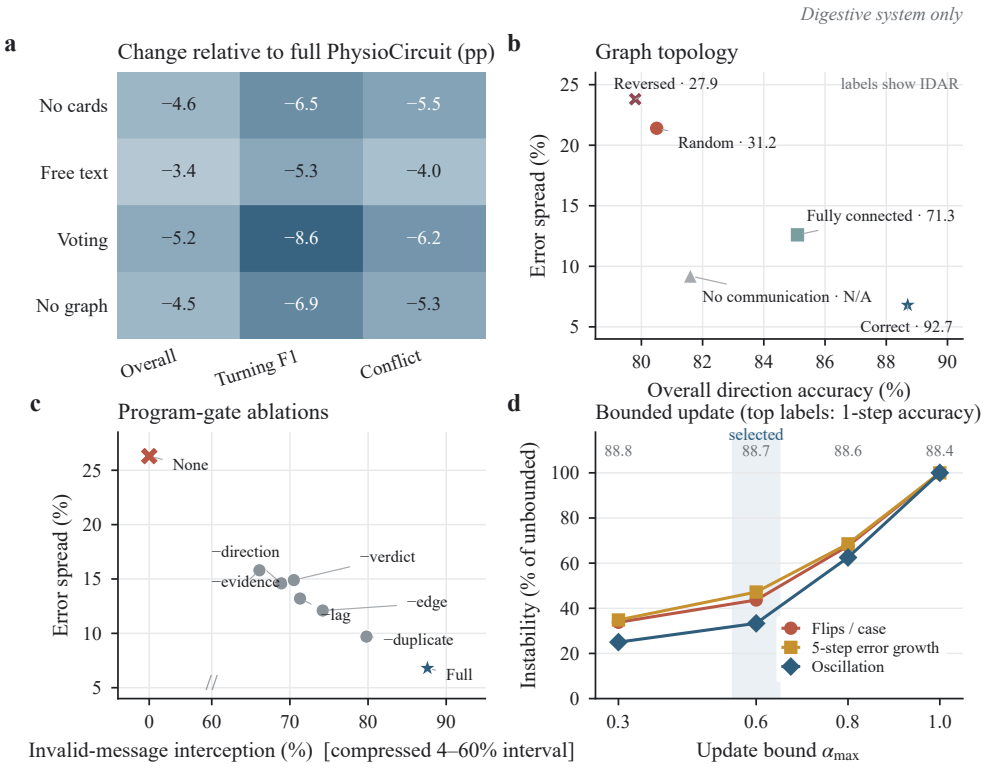


Figure 7: Component ablations on the digestive system. **a**, Change in overall accuracy, turning-point F1, and conflict accuracy relative to the full system. **b**, Overall accuracy versus error spread under graph-topology controls; edge-IDAR is reported for graph-bearing controls, while input-IDAR for the global agent is reported separately because it uses a controlled input-level counterfactual. **c**, Invalid-message interception versus error spread after removing gate categories. **d**, Flip rate, five-step error growth, and oscillation normalized to the unbounded update; numbers above the curves report one-step accuracy.

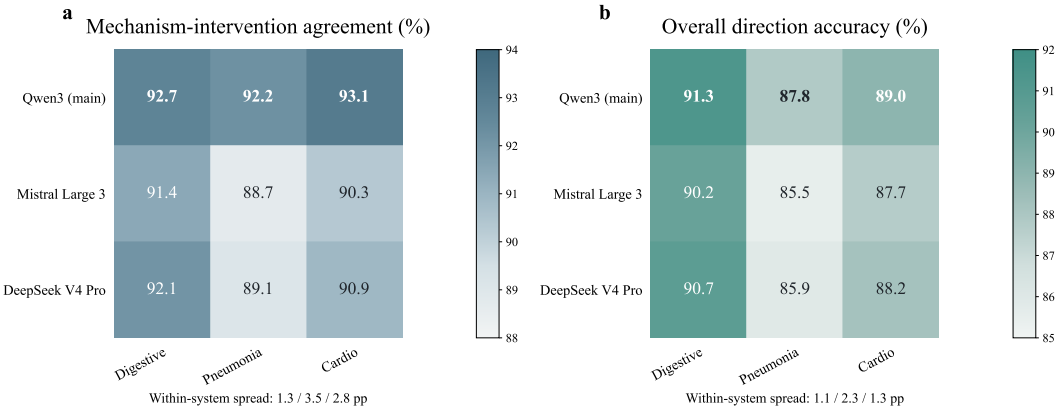


Figure 8: Between-backbone variation on IDAR across the three systems. Qwen3-235B-A22B-Instruct-2507 is the main backbone; Mistral-Large-3-675B-Instruct-2512 and DeepSeek-V4-Pro-0813 are run on the same cases, prompts, and temperature.

graph learning by 3.1 points on digestive (92.3 vs. 89.2) and 3.6 points on cardiovascular (87.6 vs. 84.0); its advantage widens further on turning-point and non-monotonic cases, where the answer depends on a mechanism transition instead of state co-occurrence. This is why accuracy stratification and intervention testing are both needed.

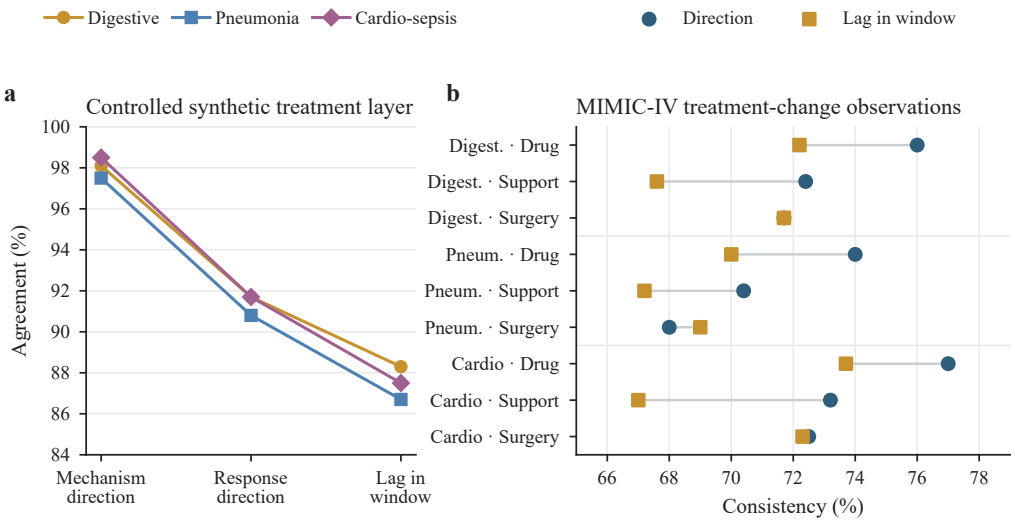


Figure 9: Treatment response and response timing. **a**, Direction consistency and lag-window agreement on synthetic treatment events. **b**, Real treatment-change observations stratified by treatment family. Drug changes are most consistent on direction; supportive care is weakest on timing in all three systems.

Backbone	**Digestive**	**Pneumonia**	**Cardio-sepsis**
PhysioCircuit edge-IDAR (%)			
Qwen3-235B-A22B-Instruct-2507 (main)	**92.7**	**92.2**	**93.1**
Mistral-Large-3-675B-Instruct-2512	91.4	88.7	90.3
DeepSeek-V4-Pro-0813	92.1	89.1	90.9
within-system spread (max − min)	1.3	3.5	2.8
Global agent input-IDAR (%)			
Qwen3-235B-A22B-Instruct-2507 (main)	35.4	33.7	36.0
Mistral-Large-3-675B-Instruct-2512	34.8	33.1	35.5
DeepSeek-V4-Pro-0813	35.9	34.2	36.6
within-system spread (max − min)	1.1	1.1	1.1
PhysioCircuit prompt-only input-IDAR (%)			
Qwen3-235B-A22B-Instruct-2507 (main)	62.1	59.8	63.4
Mistral-Large-3-675B-Instruct-2512	60.5	58.1	61.8
DeepSeek-V4-Pro-0813	61.4	59.0	62.7
within-system spread (max − min)	1.6	1.7	1.6
Overall direction accuracy (%)			
Qwen3-235B-A22B-Instruct-2507 (main)	**91.3**	**87.8**	**89.0**
Mistral-Large-3-675B-Instruct-2512	90.2	85.5	87.7
DeepSeek-V4-Pro-0813	90.7	85.9	88.2
within-system spread (max − min)	1.1	2.3	1.3

Table A9: Cross-backbone robustness sweep. The Qwen row is the main backbone; Mistral and DeepSeek are additional models run on the same cases, prompts, and temperature, with only the backbone changed. The ordering within each channel is stable across backbones: within edge-IDAR, the correct graph remains highest; within input-IDAR, prompt-only exceeds implicit graph learning and the global agent.

B.2.2 Mechanism intervention and the independence probe

Figure 4(a) shows the matched topology comparison on the digestive system; the headline edge-IDAR and input-IDAR values are reported in Section 5. These controls all expose a graph-bearing intervention interface, so the comparison asks whether topology and direction determine the response. Table A10 summarizes both channels and all controls; Table A11 reports DA, TA, and joint

IDAR separately; Table A12 breaks down the six intervention operations; Table A13 extends the topology controls to all three systems.

Figure 4(b) reports PhysioCircuit's joint IDAR by operation. Delete, replace, reverse, extend-lag, inject-wrong-message, and mask-evidence interventions are evaluated under the same direction-and-window contract. The ordering is stable across systems.

Control	IDAR (%)	Intervention channel
Edge-IDAR (graph-bearing, digestive system)		
Correct expert graph	**92.7**	Named edges; all six operations applicable
Fully connected graph	71.3	Named edges; extra non-physiological links retained
Random graph	31.2	Named edges; topology randomized
Reversed graph	27.9	Named edges; directions reversed
Input-IDAR (edge-less, D/P/C)		
Global single agent	35.4 / 33.7 / 36.0	Controlled input-level counterfactual
Organ + optimal voting	38.2 / 36.5 / 39.1	Controlled input-level counterfactual
No communication	37.0 / 35.8 / 38.3	Controlled input-level counterfactual
Vital Trace-style	45.2 / 43.8 / 46.1	Controlled input-level counterfactual
Implicit graph learning	58.4 / 56.7 / 59.2	Controlled input-level counterfactual
PhysioCircuit prompt-only	62.1 / 59.8 / 63.4	Same LLM and cards; graph propagation disabled

Table A10: IDAR channels and reported values. Graph-bearing controls use edge-IDAR on named mechanism edits; these rows report the digestive system. Edge-less controls use input-IDAR on the same semantic counterfactual descriptors; their slash-separated values are digestive / pneumonia / cardiovascular-sepsis. The PhysioCircuit prompt-only row disables typed messages and graph propagation while retaining the same LLM and card knowledge, isolating the contribution of the graph channel from model knowledge.

Control	DA (%)	TA (%)	Joint IDAR (%)	95% CI
Edge-IDAR (digestive, $N = 180$ applicable interventions)				
Correct expert graph	94.1	95.2	**92.7**	[90.1, 94.8]
Fully connected graph	74.2	82.1	71.3	[67.5, 74.8]
Random graph	42.3	58.7	31.2	[27.6, 35.0]
Reversed graph	33.5	78.4	27.9	[24.4, 31.6]
Input-IDAR (digestive, $N = 180$ descriptors)				
Global single agent	44.6	72.3	35.4	[31.2, 39.8]
Optimal voting	47.1	74.0	38.2	[33.9, 42.7]
No communication	45.8	73.5	37.0	[32.7, 41.6]
Vital Trace-style	51.8	79.4	45.2	[40.8, 49.7]
Implicit graph learning	63.2	85.1	58.4	[54.1, 62.5]
PhysioCircuit prompt-only	67.3	82.6	62.1	[58.0, 66.0]

Table A11: Direction Agreement (DA), Timing Agreement (TA), and joint IDAR on the digestive system. DA and TA are reported separately to show whether failures arise from wrong direction or missed timing windows. 95% bootstrap intervals resample intervention instances. The reversed-graph row keeps TA above the random graph (78.4 vs. 58.7) but below the fully connected graph (82.1), while DA falls near the chance level of a three-way choice: the timing channel partially survives and the directional answer loses its information.

The independence probe injects 18 wrong truth rules. PhysioCircuit retains the medically correct direction on 84.6%, 81.5%, and 83.2% of probe interventions, follows the injected rules on 11.2%, 13.6%, and 12.0%, and produces no directional response on the remainder (4.2%, 4.9%, and 4.8%). Among the interventions that retain the correct direction, 88.1%, 86.4%, and 87.5% also fall within

Intervention operation	Digestive	Pneumonia	Cardio-sepsis
Delete mechanism	93.2	92.5	93.8
Replace mechanism card	91.8	91.0	92.3
Reverse mechanism direction	90.5	89.8	91.2
Extend lag window	92.1	91.5	92.7
Inject wrong message	94.0	93.3	94.4
Mask source evidence	94.5	93.9	95.0
Average	92.7	92.0	93.2

Table A12: PhysioCircuit edge-IDAR by intervention operation across the three systems. All six operations use the same direction-and-window scoring contract. Inject-wrong-message and mask-source-evidence are the most faithful because they force the receiving organ to arbitrate. Values are joint DA∧TA over all applicable interventions ($N = 180/210/190$).

Topology control	Digestive	Pneumonia	Cardio-sepsis
Correct expert graph	**92.7**	**92.2**	**93.1**
Fully connected graph	71.3	69.8	72.5
Random graph	31.2	29.8	32.4
Reversed graph	27.9	26.5	28.7

Table A13: Edge-IDAR for graph-bearing topology controls across all three systems. The correct-graph lead over the fully connected graph is stable (20.6–22.4 points), and random and reversed graphs remain near the lower bound in every system.

the card-defined lag window, yielding joint probe IDAR of 74.5%, 70.4%, and 72.8%. The truth-retrieval control follows the injected rules on 92.3% of cases. Source overlap remains below 30% in all systems. Treatment responses are evaluated separately because treatments enter through organ evidence.

B.2.3 Ablation analysis

We ablate mechanism knowledge, graph topology, program gates, and update magnitude on the digestive system; the same ordering holds qualitatively in the other systems. This design separates knowledge, communication, topology, robustness, and long-horizon stability.

Figure 7(a) shows that removing cards costs 4.6 points (91.3 to 86.7), replacing typed messages with free text costs 3.4 points (91.3 to 87.9), and optimal voting over organ states costs 7.8 points (83.5). Adding compute without the graph reaches 84.2 and recovers only part of the gap. These controls separate the contribution of mechanism knowledge, typed communication, macro arbitration, and topology.

Figure 7(b) shows a 10.8-point loss with a random graph (91.3 vs. 80.5, paired CI [8.9, 12.7], $p < 0.001$), 79.8 accuracy with reversed edges, and 81.6 without communication. A fully connected graph reaches 85.1, recovering part of the gap but remaining 6.2 points below PhysioCircuit. The corresponding edge-IDAR and input-IDAR values are reported in Section 5.

Figure 7(c) shows the fourteen gates as six atomic checks, edge existence, direction, lag, evidence, duplicate detection, and verdict. Removing the four functional groups: (i) edge existence + duplicate detection, (ii) direction + lag, (iii) evidence, (iv) verdict. This lowers invalid-message interception from 87.6% to 74.2%, 68.9%, 66.1%, and 70.5%, respectively. Removing all gates increases error spread from 6.8 to 26.3. The gates therefore constrain error propagation, not only message format.

Figure 7(d) shows that one-step accuracy is nearly flat across the bound, 88.8, 88.7, 88.6, and 88.4 at $\alpha_{\max} = 0.3$, 0.6, 0.8, and 1.0, so accuracy does not choose the operating point. The choice rests on long-run behavior: at 0.6, five-step error growth is 0.42 versus 0.89 at 1.0, and direction flips fall from 7.1 to 3.1 per case. A smaller bound of 0.3 under-revises conflict cases, where conflict accuracy drops to 87.5 against 92.3 at 0.6, while a bound of 1.0 accumulates error across steps. We set the operating point at 0.6.

Together, the ablations assign distinct roles to cards, typed messages, topology, gates, and bounded updates: the cards carry knowledge, typed messages support arbitration, topology determines prop-

agation, gates limit error spread, and bounded updates stabilize rollouts. Each removal degrades accuracy, intervention, robustness, or stability. The ablations are centered on the digestive system, so transfer across systems tests whether this decomposition remains useful beyond the source graph.

B.2.4 TRANSFER ACROSS SYSTEMS

Figure 5(a) compares zero-shot graph reuse with configuration-only transfer. Rewritten pneumonia and cardiovascular graphs require 127 and 154 configuration lines and reach 87.5% and 88.7%, 3.8 and 2.6 points below the digestive source. Zero-shot losses range from 3.9 to 11.2 points and track organ overlap: the digestive graph shares no organ with either six-organ system, while the pneumonia and cardiovascular graphs share heart and kidney structure. Panel B shows larger reuse losses for implicit graph learning (7.8–17.6 points vs. 3.9–11.2 for PhysioCircuit); PhysioCircuit leads in all six directions by 1.8–7.2 points. The result supports mechanism cards as reusable relations, while also motivating validation on records outside the generated benchmark.

B.2.5 THE REAL LONGITUDINAL PILLAR

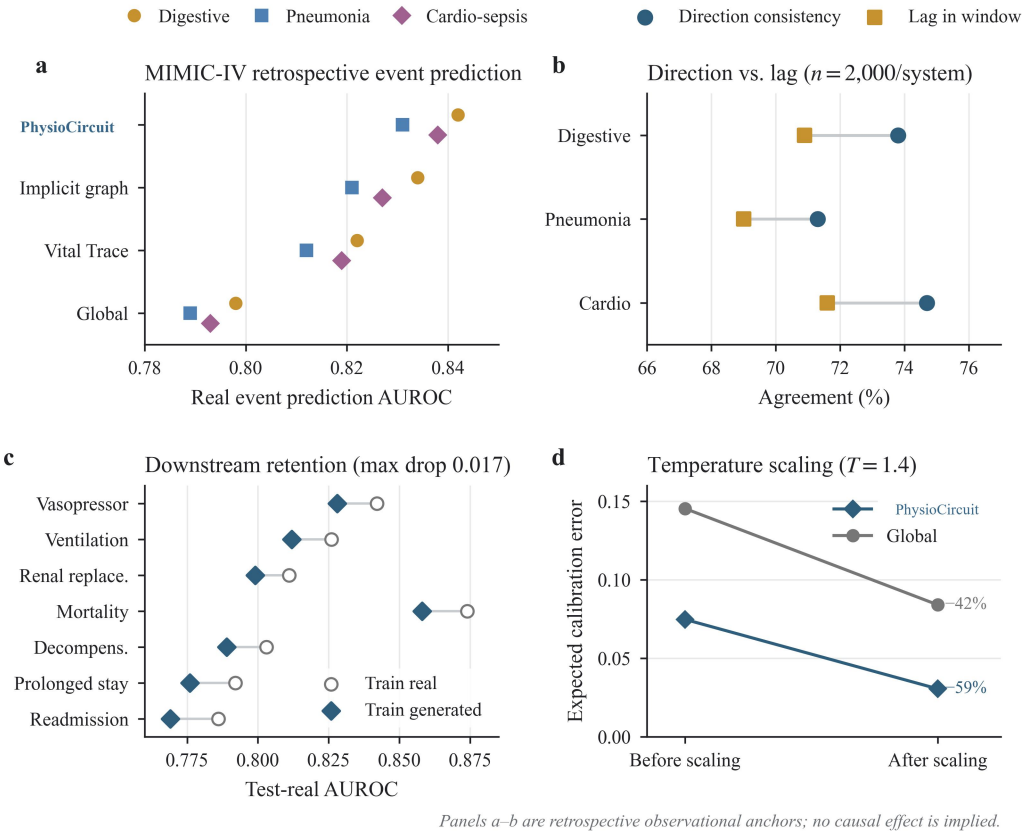


Figure 10: Real-data and downstream validation. **a**, Retrospective MIMIC-IV event-prediction AUROC. **b**, Direction consistency and lag-window agreement on 2,000 treatment-change observations per system. **c**, AUROC of models trained on real versus generated trajectories and evaluated on the same real test set. **d**, Overall expected calibration error before and after temperature scaling.

On 600 ICU admissions, PhysioCircuit reaches AUROC 0.842, 0.831, and 0.838 for real-event prediction, higher than implicit graph learning (0.834, 0.821, 0.827), Vital Trace-style reasoning (0.822, 0.812, 0.819), and the global agent (0.798, 0.789, 0.793) (Figure 10(a,b)). Mechanism-direction consistency is 73.8%, 71.3%, and 74.7%, with timing at 70.9%, 69.0%, and 71.6%. These real scores sit below synthetic IDAR because co-interventions go unrecorded and timestamps are coarser, yet the ordering holds.

Item	Organ-Agents	PhysioCircuit
Training stages	2 (supervised fine-tuning, then reinforcement-guided coordination)	0
Physiological systems covered	9	3
Backbone models	one fixed backbone	one main + two cross-backbone checks
Retraining on a new system	required	not required
Cost of a new system	full training run	127 and 154 configuration lines, plus 14 and 12 mechanism cards
Synthetic cases per system	not applicable, trained on real cohorts	120
Tokens per case	not reported	41.7K for three organs, 42.4K for six organs

Table A14: Cost comparison. The Organ-Agents column quotes the figures reported in its own paper; the PhysioCircuit column reports measured values from this work.

B.2.6 Downstream task retention

Figure 10(c) evaluates seven ICU tasks with admission-disjoint train, validation, and test splits (300/100/200). The validation split is used only to select the temperature-scaling parameter T; no admission appears in more than one split. Models trained on generated or real trajectories are evaluated on the same real test set, holding the target distribution fixed. The average AUROC drop is 0.015 and the largest is 0.017 on unplanned readmission; ICU mortality drops 0.016 from an upper bound of 0.874. The generated trajectories retain task-relevant information, and the expert evaluation supplies the clinical check.

B.2.7 Uncertainty calibration

Figure 10(d) reports ECE over 3,984 macro predictions using post-hoc temperature scaling (Guo et al., 2017). Temperature scaling with $T = 1.4$ reduces PhysioCircuit ECE from 0.0747 to 0.0305, a 59% reduction; the global agent changes from 0.1453 to 0.0840. Baseline and recovery phases start near calibration (0.0049 and 0.0057), while conflict-emerges and decompensation reach 0.0141 and 0.0153. Residual overconfidence therefore concentrates where organ evidence conflicts and the framework must arbitrate. Calibration shows both the benefit of post-hoc correction and the phase-specific risk that remains.

B.2.8 Cost and the zero-training comparison

Table A14 compares the training burden and extension cost. The Organ-Agents figures are quoted from its report: nine systems, two training stages, one fixed backbone, and retraining when the system changes. PhysioCircuit uses no training, adds systems through mechanism cards and configuration, and costs 41.7K–42.4K tokens per case in the measured settings. Adding pneumonia and cardiovascular-sepsis requires 14 and 12 cards plus 127 and 154 configuration lines. Gates intercept 87.6% of invalid messages, retain 82.4% of valid messages, and reduce effective-correction cost by 41.3% relative to free-text coordination, which spends 71.0K tokens per case on the same task.