RealICU

Do LLM Agents Understand Long-Context ICU Data?

A Benchmark Beyond Behavior Imitation

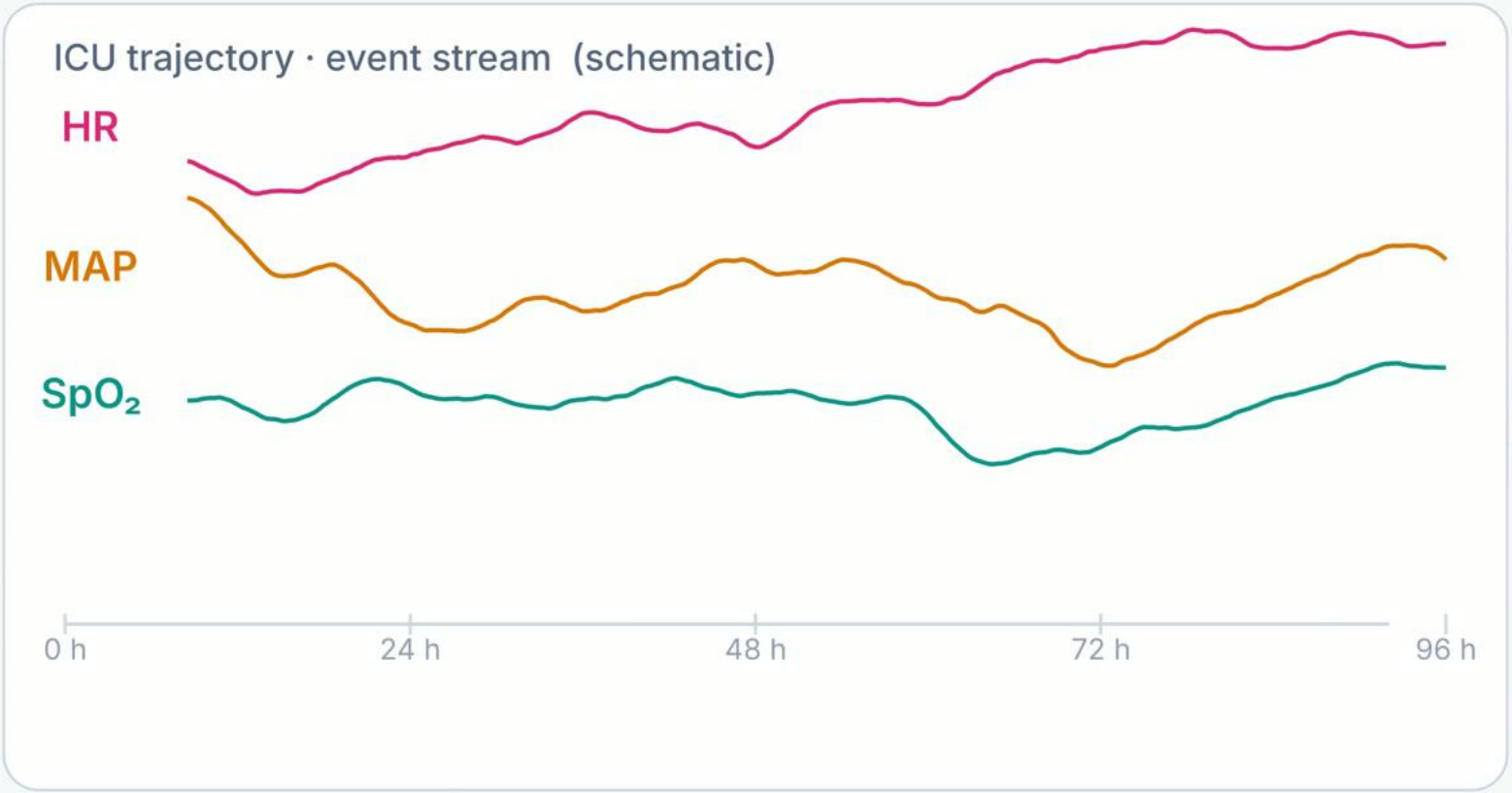
NeurIPS 2026

Chengzhi Shen · Weixiang Shen · Tobias Susetzky · Chen Chen · Jun Li
Yuyuan Liu · Xuepeng Zhang · Zhenyu Gong · Daniel Rueckert · Jiazhen Pan

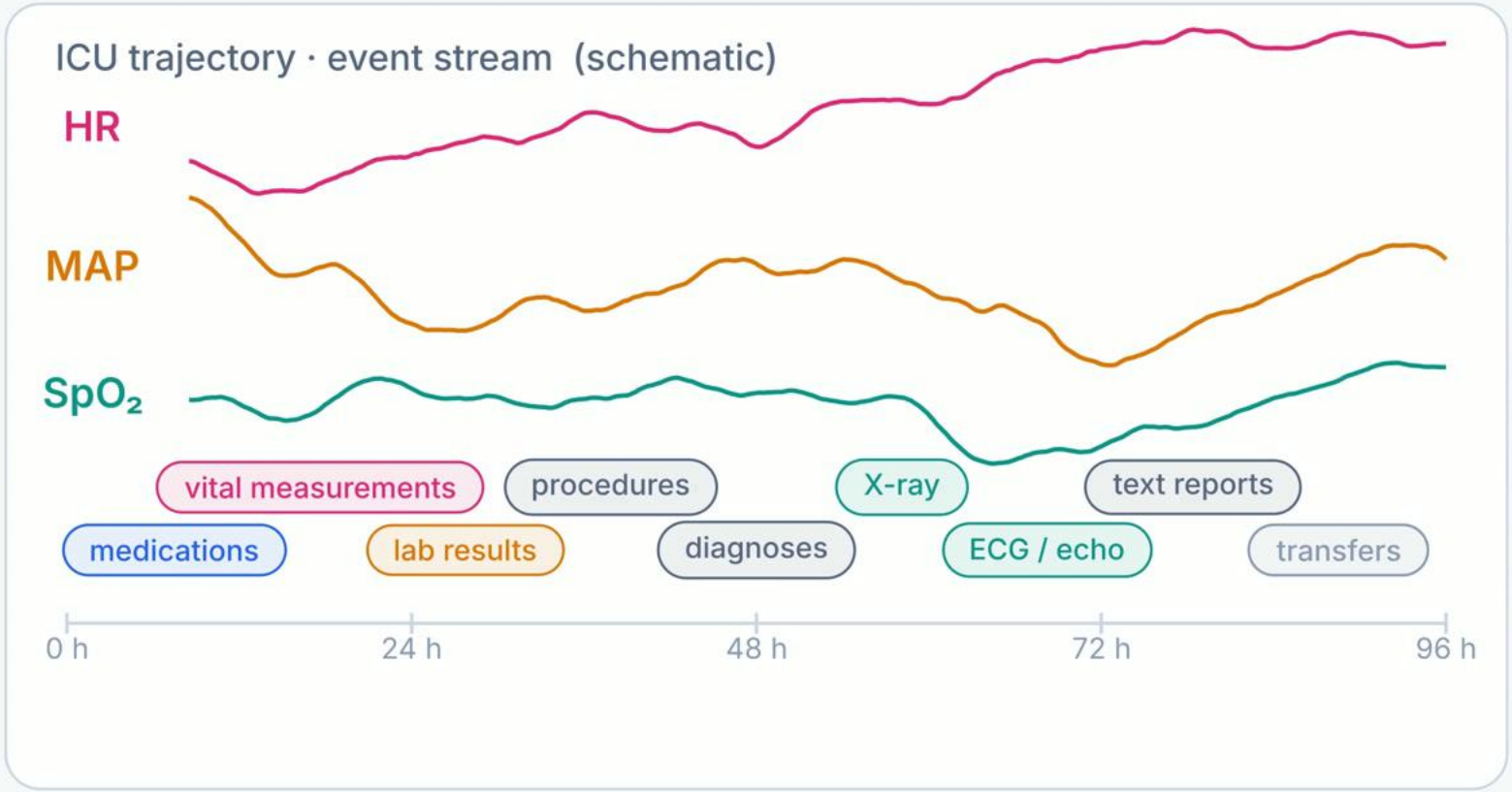
Technical University of Munich · TUM University Hospital · LMU Munich · University of Sheffield · University of Oxford
Zhongshan Hospital Fudan University · Sun Yat-sen University Cancer Center · Imperial College London · MCML · relAI

01 THE ICU PROBLEM

An ICU stay is a long, dense, evolving stream of data



An ICU stay is a long, dense, evolving stream of data



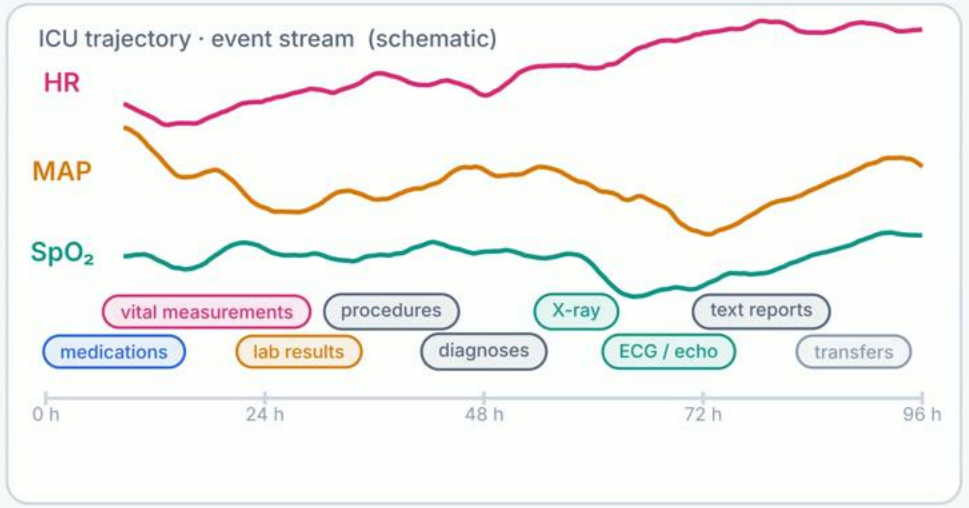


hour 96

Physicians must reassess again and again,
under time pressure.
A decision now can shape outcomes hours or days later.

What must an AI co-pilot deliver, at every moment?

4 physician-motivated tasks · distilled from > 30 clinicians

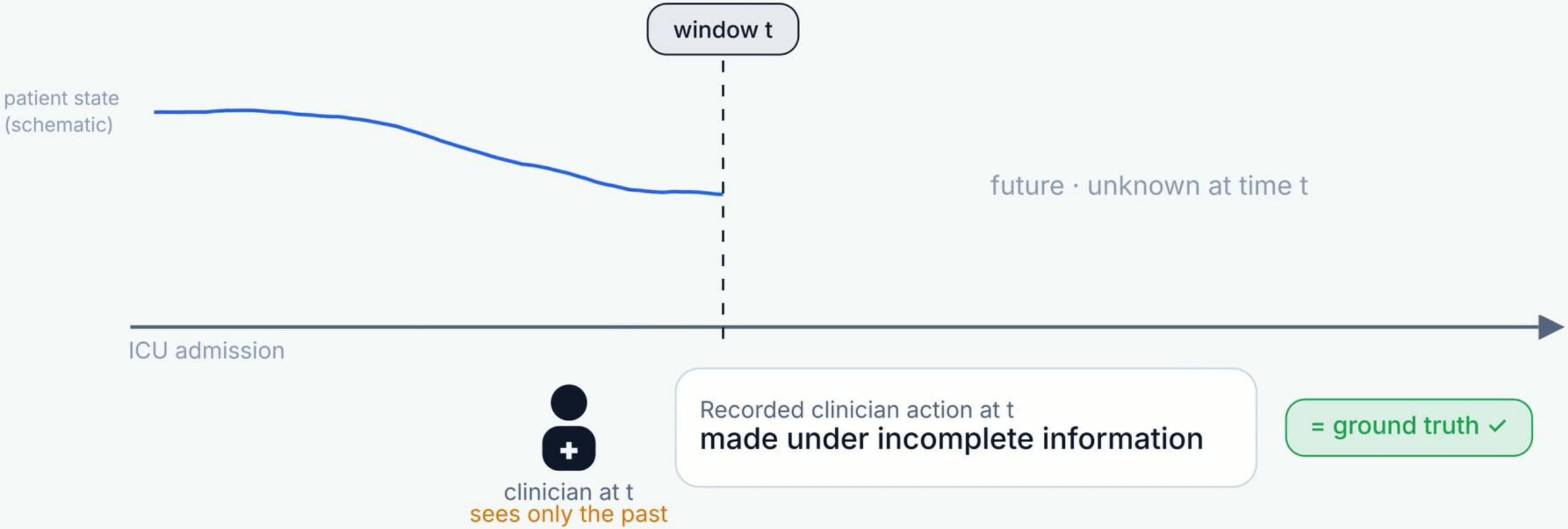


AI co-pilot

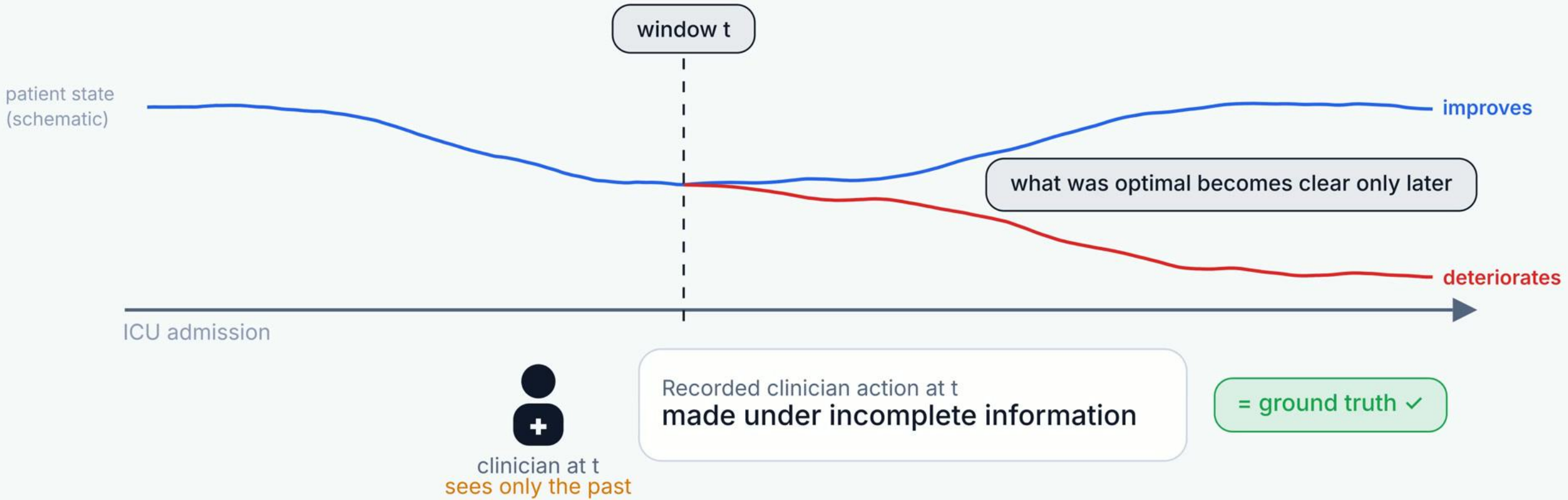
- **Patient Status**
  e.g. deteriorating
- **Acute Problems**
  e.g. refractory intracranial hypertension
- **Recommended Actions**
  e.g. therapeutic temperature management
- **Red Flags**
  e.g. routine hyperosmolar therapy at Na^+ 155

examples: physician labels of one RealICU-Gold window (Appendix D.1)

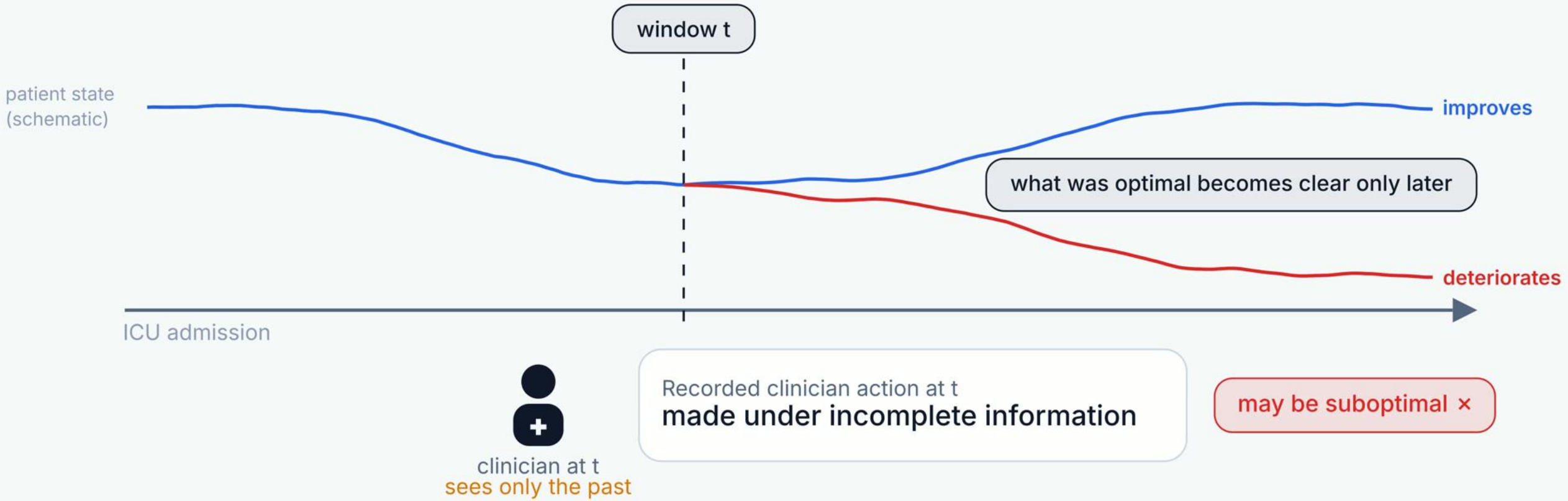
Existing benchmarks treat recorded clinician actions as ground truth



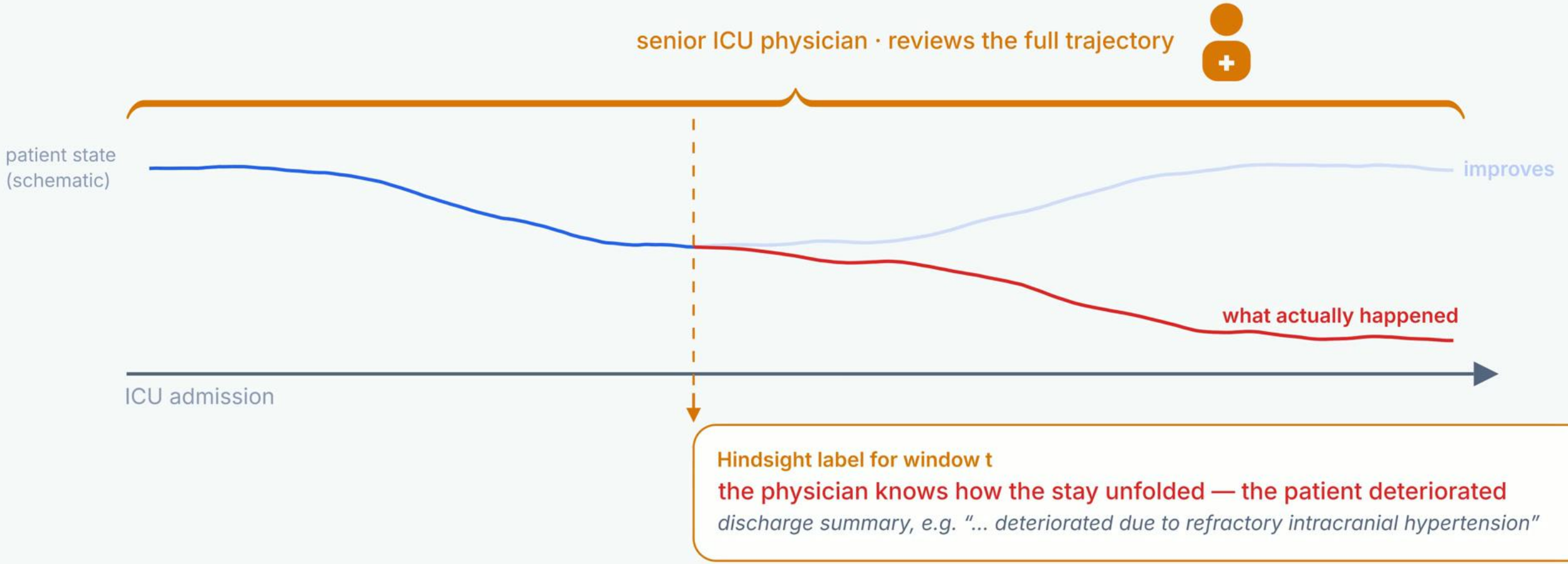
Existing benchmarks treat recorded clinician actions as ground truth



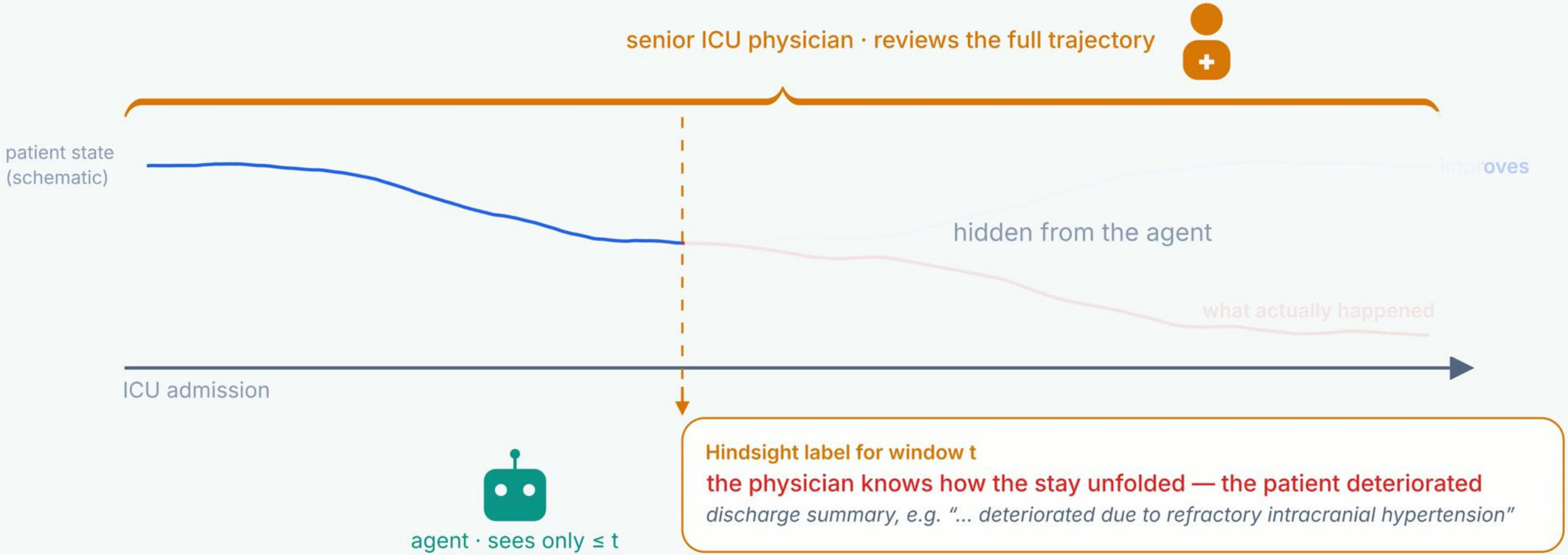
Recorded ≠ correct: scoring against these labels rewards imitation



RealICU: senior ICU physicians label each window in hindsight



The agent sees the past. The label knows the future.



From MIMIC-IV to hindsight-labeled windows



From MIMIC-IV to hindsight-labeled windows



RealICU-Gold

930 windows

senior ICU physicians · consensus labels

hindsight review of the full trajectory

inter-rater reliability 0.826 – 0.985

From MIMIC-IV to hindsight-labeled windows



ReallCU-Gold

930 windows

senior ICU physicians · consensus labels

hindsight review of the full trajectory

inter-rater reliability 0.826 – 0.985

ReallCU-Scale

11,862 windows · 13×

Oracle: an LLM hindsight labeler

validated against physician consensus

F1 vs. Gold 0.895 – 0.987 across the four tasks

Four tasks, scored against the hindsight labels

TASK	WHAT THE AGENT MUST OUTPUT	METRIC
Patient Status	improving / stable / deteriorating	Accuracy · F1
Acute Problems	problems that need active management now	Hit@5 · Recall@5
Recommended Actions	actions likely to help within 1 hour	Hit@5 · Recall@5
Red Flags	high-risk actions to avoid	HRR@5 (lower is better)

Four tasks, scored against the hindsight labels

TASK	WHAT THE AGENT MUST OUTPUT	METRIC
Patient Status	improving / stable / deteriorating	Accuracy · F1
Acute Problems	problems that need active management now	Hit@5 · Recall@5
Recommended Actions	actions likely to help within 1 hour	Hit@5 · Recall@5
Red Flags	high-risk actions to avoid	HRR@5 (lower is better)

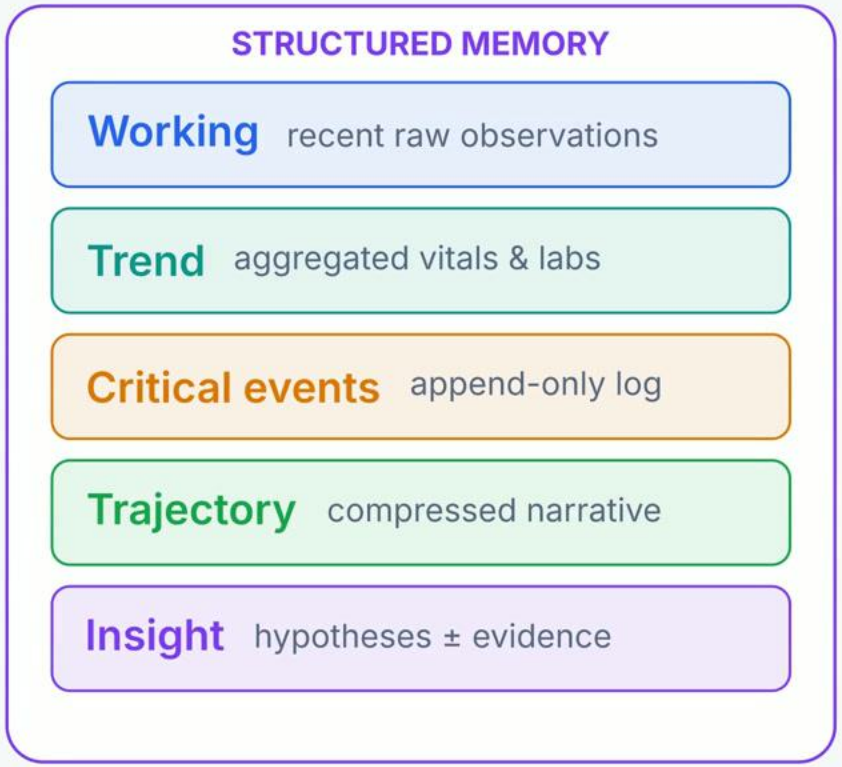
$$\mathrm{HRR@}k = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}[\, a_i \in \text{Red Flags}\,]$$

fraction of the top-k recommended actions that match a red flag
averaged per stay, then across stays · k = 5 in the paper

free-text matching
PubMedBERT embeddings

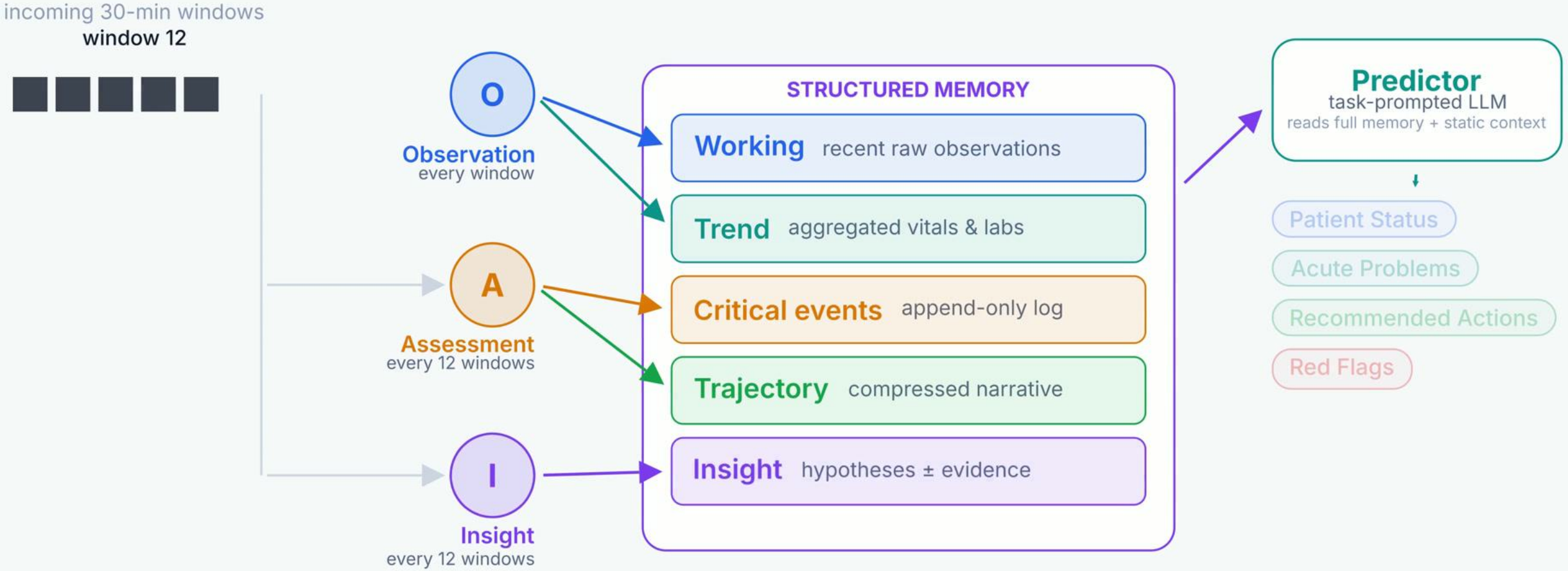
ICU-Evo: memory as a structured belief state

backbone-agnostic · every output must cite raw recorded events



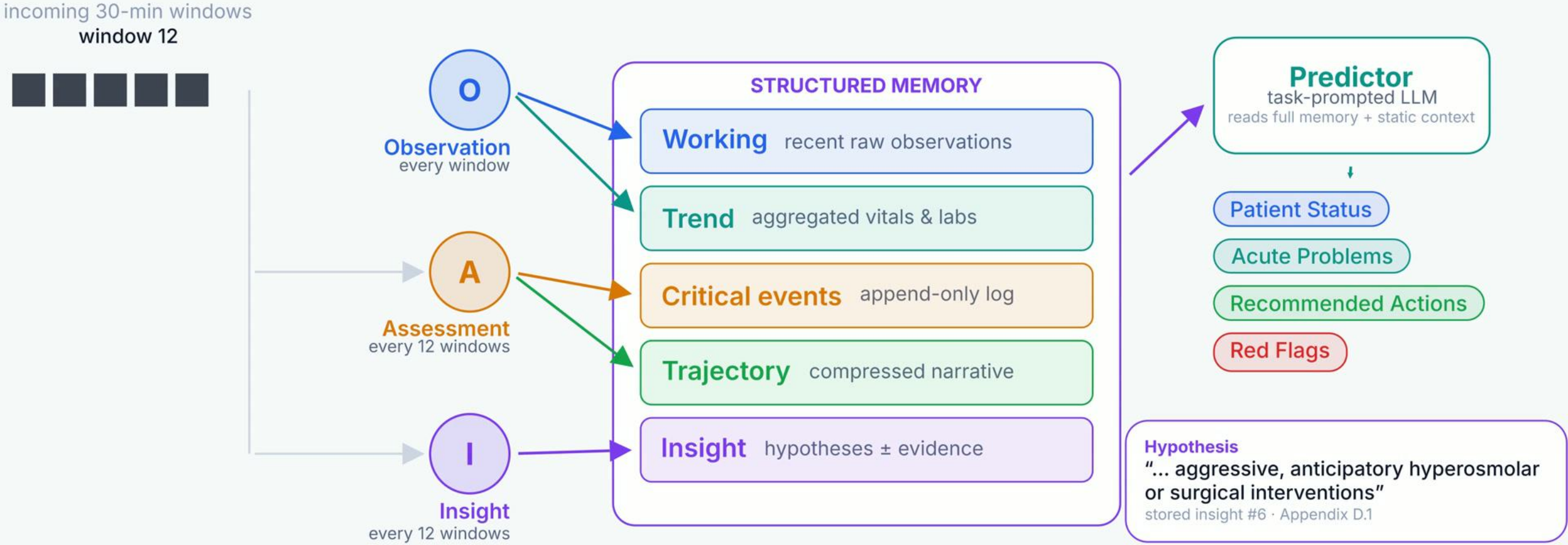
ICU-Evo: memory as a structured belief state

backbone-agnostic · every output must cite raw recorded events



ICU-Evo: memory as a structured belief state

backbone-agnostic · every output must cite raw recorded events



Structured memory consistently improves clinical reasoning

RealICU-Gold · Gemini-3.1-pro · all four memory modes (Table 2)



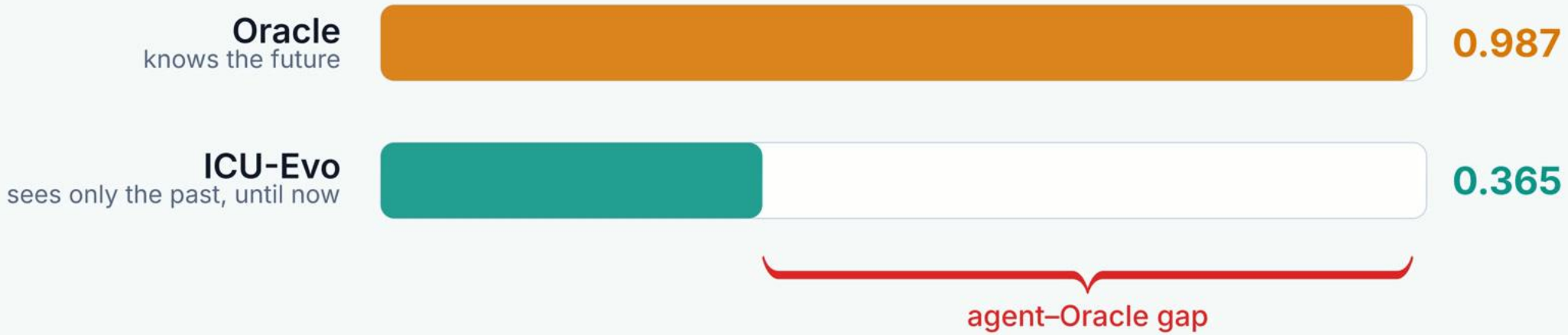
… and the gain holds across backbones

ReallCU-Gold · Hit@5 on Acute Problems and Recommended Actions · all memory modes (Table 2)



… but ReallCU remains unsolved

Patient Status · F1 on ReallCU-Gold

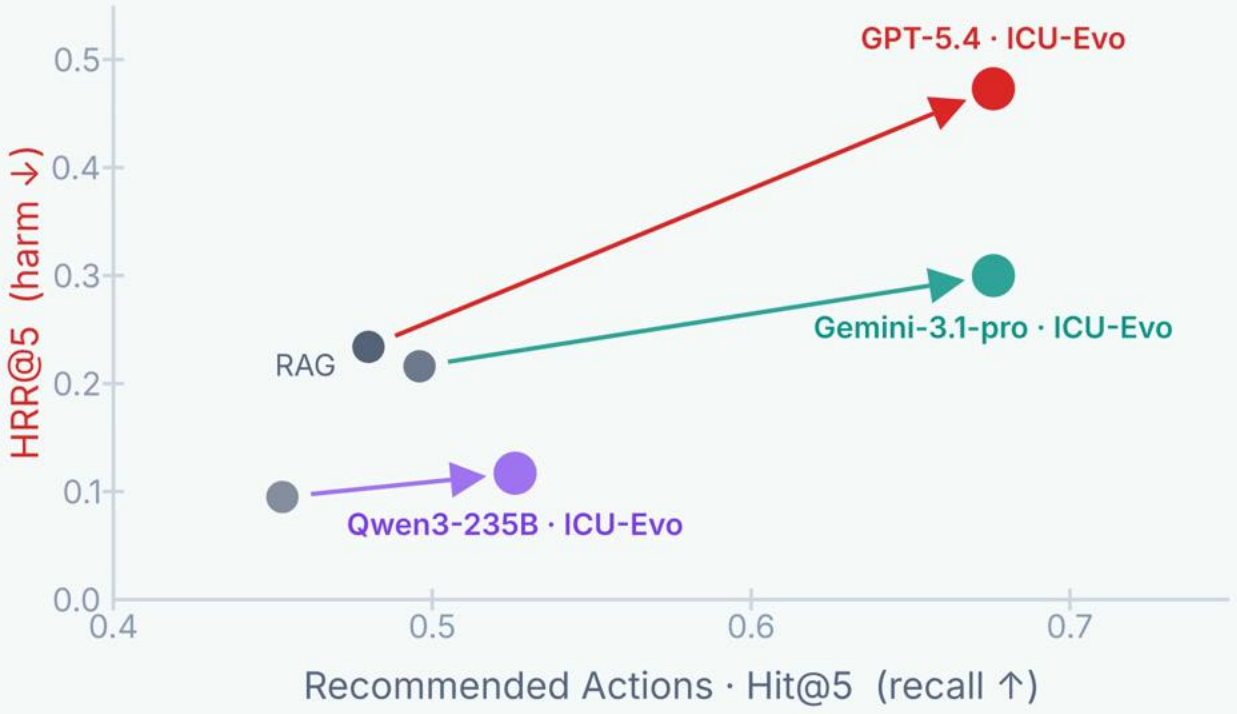


The gap lies in how agents integrate evidence over time.

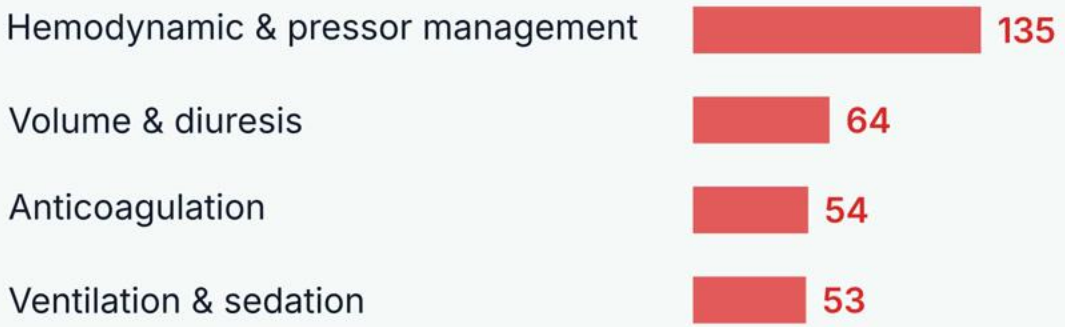
Longitudinal data calls for more comprehensive clinical reasoning.

Failure mode 1 · the recall–safety tradeoff

ReallCU-Gold · RAG → ICU-Evo for each backbone (Table 2) · flagged cases analysed in Sec. 5.5



394 flagged recommendations (Sec. 5.5)



Agents recognise part of a syndrome and propose the full treatment bundle before contraindications have been ruled out.

Failure mode 1 · recall–safety tradeoff

Appendix D.1 · ICU-Evo + Gemini-3.1-pro · ICU day 8

PATIENT / GOLD REVIEW

Severe SAH, ICP 40+ mmHg.

Na^+ 155, osm 325.

Avoid routine hyperosmolar therapy.

→

INSIGHT MEMORY #6

Anticipate severe ICP crises with aggressive hyperosmolar therapy.

No current sodium trend.

→

AGENT PREDICTION #2

Give mannitol or 3 % saline for sustained ICP above 20–22 mmHg.

RED FLAG

The agent recalls the treatment bundle but overlooks a patient-specific contraindication.

Failure mode 2 · anchoring bias

Appendix D.2 · ICU-Evo · ICU day 4

Anchoring bias the agent clings to its early assumptions and does not revise them as new evidence comes in

PATIENT / GOLD REVIEW

Eisenmenger syndrome.

Baseline SpO_2 85 – 93 %.

Status: **Stable.**

No new SpO_2 readings.

→

INSIGHT MEMORY #2

Refractory hypoxemia despite intensive oxygen therapy and diuresis.

Earlier hypothesis carried forward.

→

AGENT PREDICTION #4

Escalate supplemental oxygen to achieve SpO_2 > 95 %, despite the patient's baseline saturation of 85 – 93 %.

Gold review: avoid titrating oxygen to SpO_2 > 95 % in this patient.

RED FLAG

An earlier hypoxemia hypothesis continues to influence recommendations despite limited new evidence.

≈ 6 hypotheses per patient · 80 % carry anticipatory exceptions (Sec. 5.5)

Takeaways

1

Clinically grounded evaluation.

ReallCU evaluates four essential ICU co-pilot capabilities, defined with input from 30+ clinicians, focusing on clinical correctness based on hindsight.

2

A scalable benchmark.

ReallCU-Gold provides 930 physician-consensus windows, while ReallCU-Scale extends to 11,862 windows using a physician-validated LLM evaluator.

3

Key failure modes in memory-augmented LLM agents.

Despite improvements from ICU-Evo, recall–safety tradeoffs and anchoring bias remain critical challenges for reliable ICU decision support.

Project page
chengzhi-leo.github.io/ReallCU-Bench

Code
github.com/chengzhi-leo/ReallCU-Bench

Paper
arXiv:2605.13542