From Textbook Causality to Regulatory Evidence

Building the Evaluation Substrate for Clinical AI

Yuan (Emily) Xue

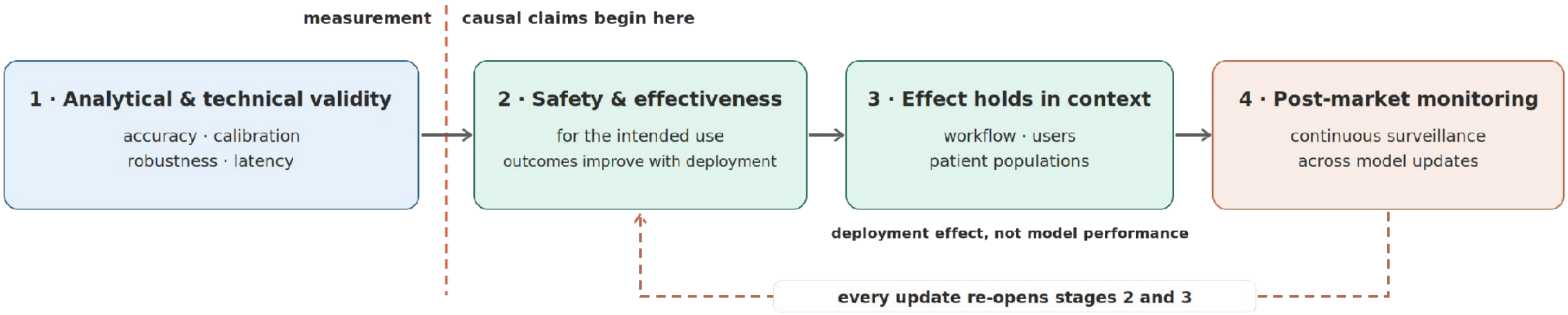
Head of Enterprise AI, Scale AI

Ex-Google · Gemini / Medical Brain / Cloud Agent Benchmark
Ex-Vanderbilt · Associate Professor of Computer Science

GOAL

Regulatory Evidence for Next-Generation Clinical AI

The traditional evidence model assumes a stable intervention. The new generation of clinical AI does not.



The regulatory question becomes

What evidence shows that deploying this AI system improves clinical decisions and patient outcomes, while controlling safety and risk?

WHERE THE FIELD IS

Current Benchmarks: Task Reliability, Not Deployment Evidence

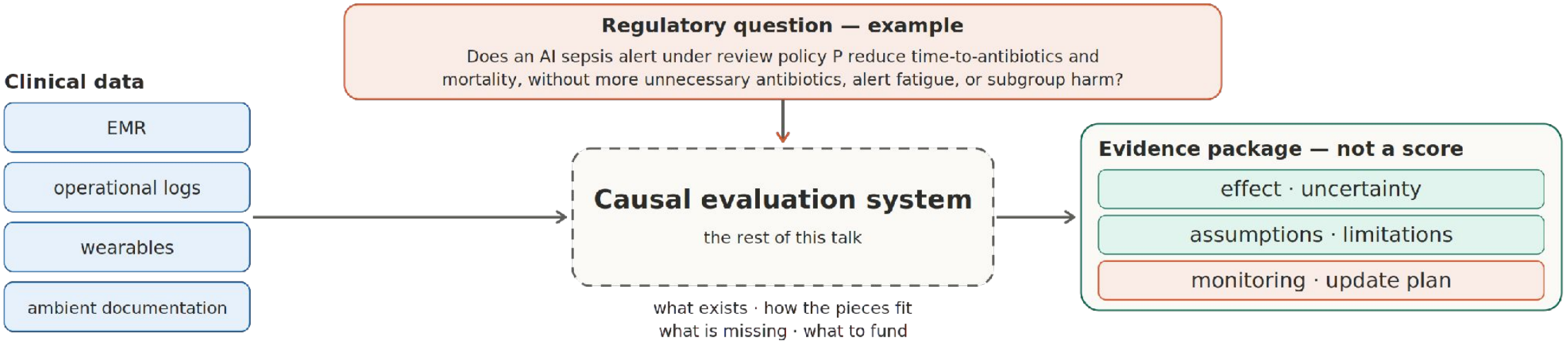
Existing benchmarks largely test whether an AI system can perform a clinical task reliably in a controlled evaluation. They generally do not establish whether deploying that system is safe and clinically effective within a real healthcare workflow.

Layer	Examples	Evaluates	Gap
Medical QA	MultiMedQA, HealthBench	Medical knowledge	Not workflow or outcome impact
Simulated clinical interaction	AgentClinic, Clinical Environment Simulator	Sequential behavior in synthetic settings	Simulated outcomes, not real deployment effect
EHR-grounded / agentic tasks	MedHELM, MedAlign, EHRSHOT, EHRSQL, EHRAgent, EHR-Complex, MedAgentBench, FHIR-AgentBench	Task completion over clinical records and tools	Not causal deployment evidence
Causal reasoning / causal agents	CLadder, CausalBench (Zhou; Wang), CounterBench, Causal-Copilot	Causal reasoning or causal-analysis workflow	Synthetic and clinical-regulatory impact
Deployment effect	**Regulatory causal-evaluation substrate**	**Effect of AI deployment on decisions, outcomes, safety**	**Needs new data + benchmark + simulation**

Most current benchmarks primarily measure capability and task reliability stage 1.
Regulatory evidence starts when we ask for deployment effect.

FOCUS OF THIS TALK

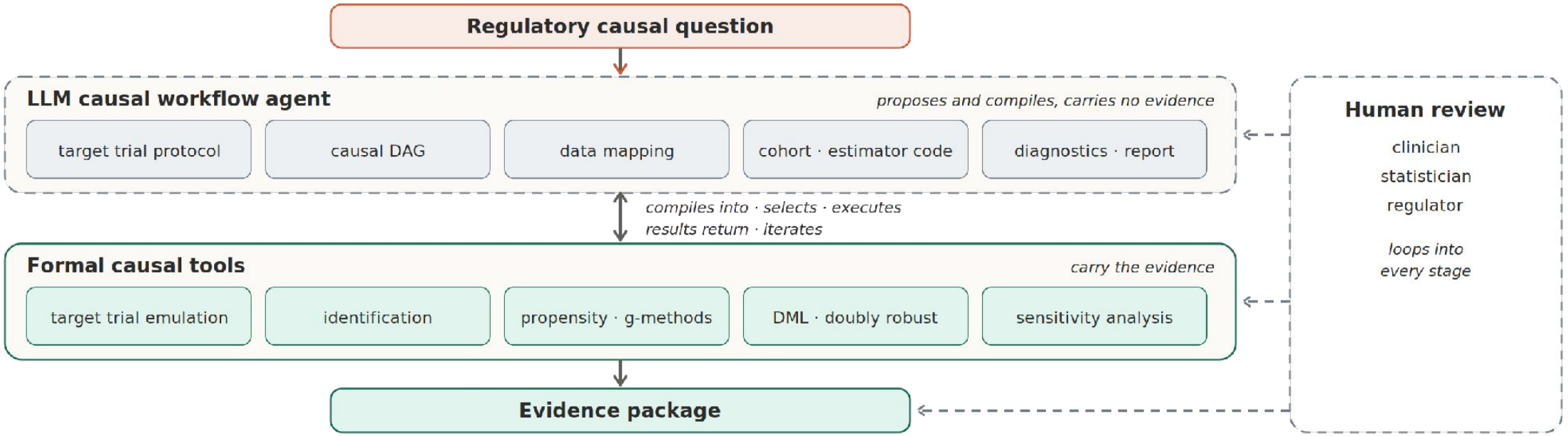
Causal Evaluation System



What should this system look like, and what must we build to make it credible for regulatory use?

REFERENCE ARCHITECTURE

LLM: Knowledge Extractor and Workflow Orchestrator

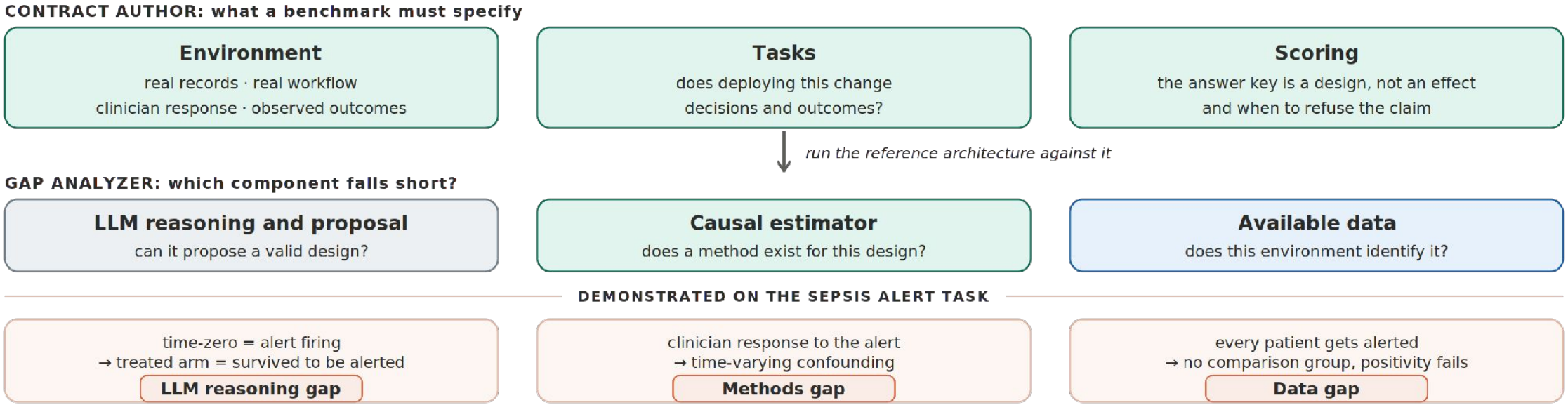


LLM – Not as Causal Authority
Auditable artifacts and formal estimators carry the evidence.

THE FIRST FOUNDATION

Benchmark: Contract Author and Gap Analyzer

The architecture is a reference. It does not tell us which components work and which are weak. That is what a benchmark is for.

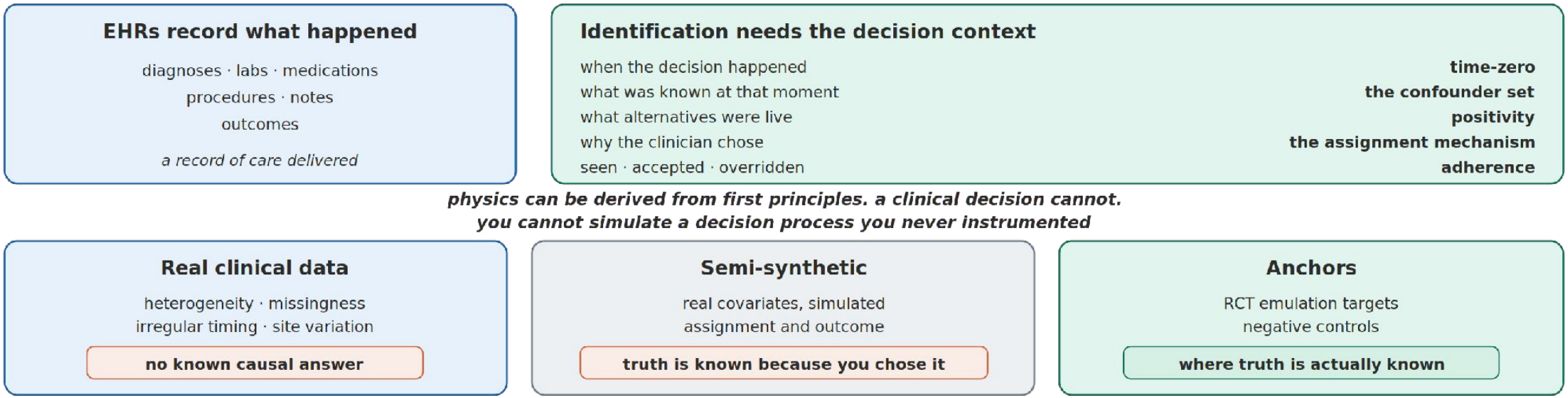


The benchmark does not rank systems. It localizes the gap.

WHAT DOES NOT EXIST YET

Data Substrate: Ground Truth Under Realistic Conditions

A benchmark needs tasks where the answer is known. Real clinical data does not have them.

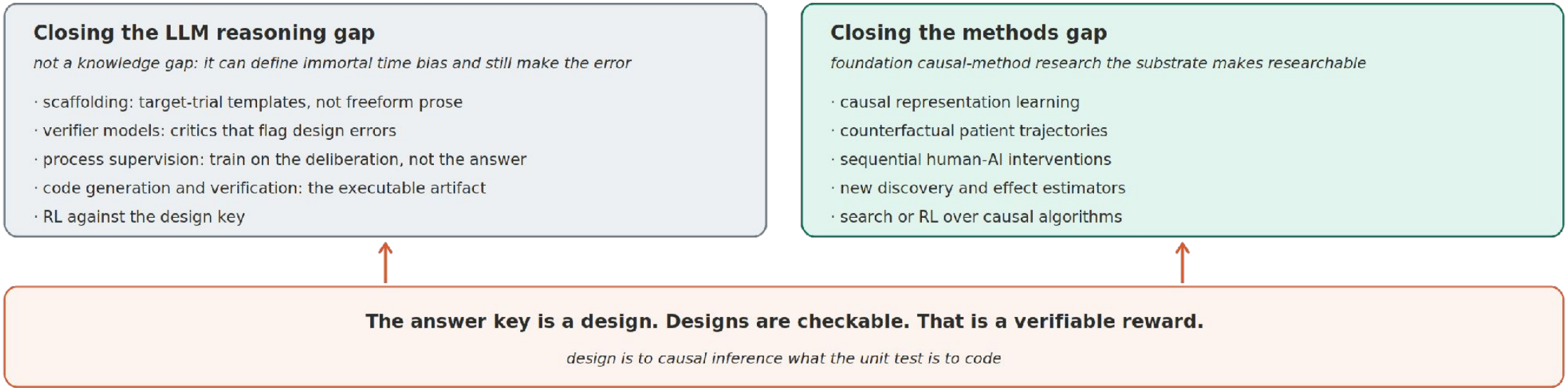


In-silico evidence works when the physics is known.
A clinical decision has to be instrumented.

WHAT THE BENCHMARK UNLOCKS

Methods: A Verifiable Goal for Causal Reasoning

The benchmark localized two technology gaps. It also supplies what closes them: a goal that can be checked.

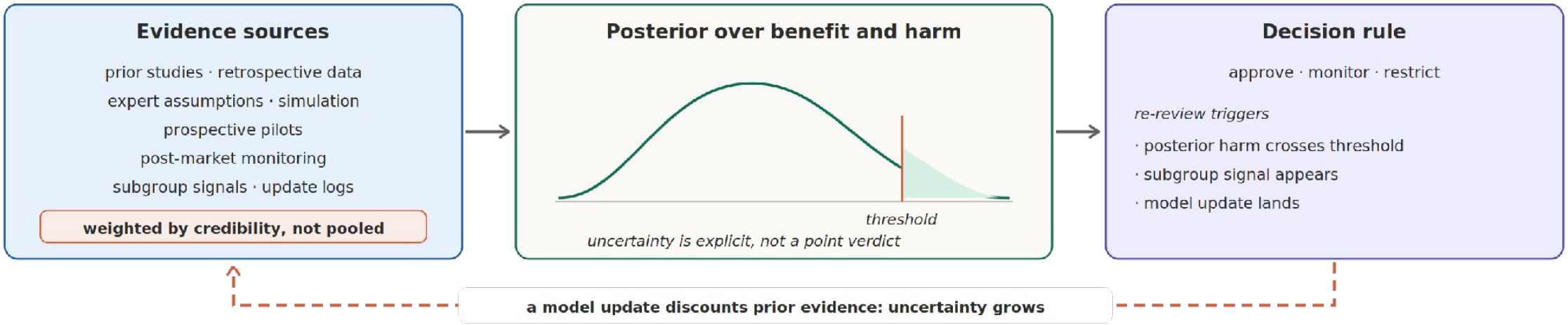


The benchmark finds the gap. The same key closes it.

LIVING EVIDENCE

Bayesian Methods: Living Evidence

The evidence chain said every update re-opens stages 2 and 3. This is the machinery that closes that loop.

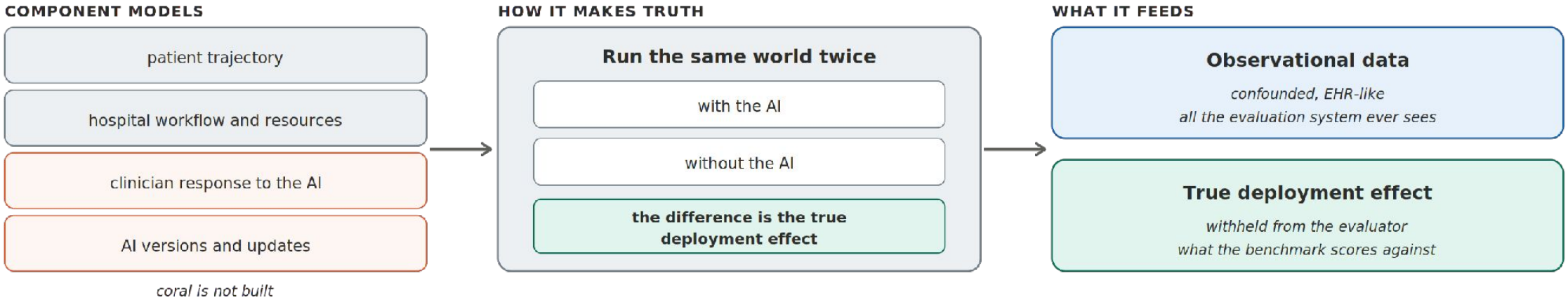


Causal models define what we estimate.
Bayesian methods update confidence as evidence accumulates.

WHERE THE SUBSTRATE LEADS

Simulation: Clinical AI Wind Tunnel

Semi-synthetic data gives a known effect for a static exposure. Deployment is not static: the alert fires, a clinician responds, the state moves, the model updates. To know the truth for that, you have to run it.



The AI never touches the patient. The effect flows through the clinician's response.

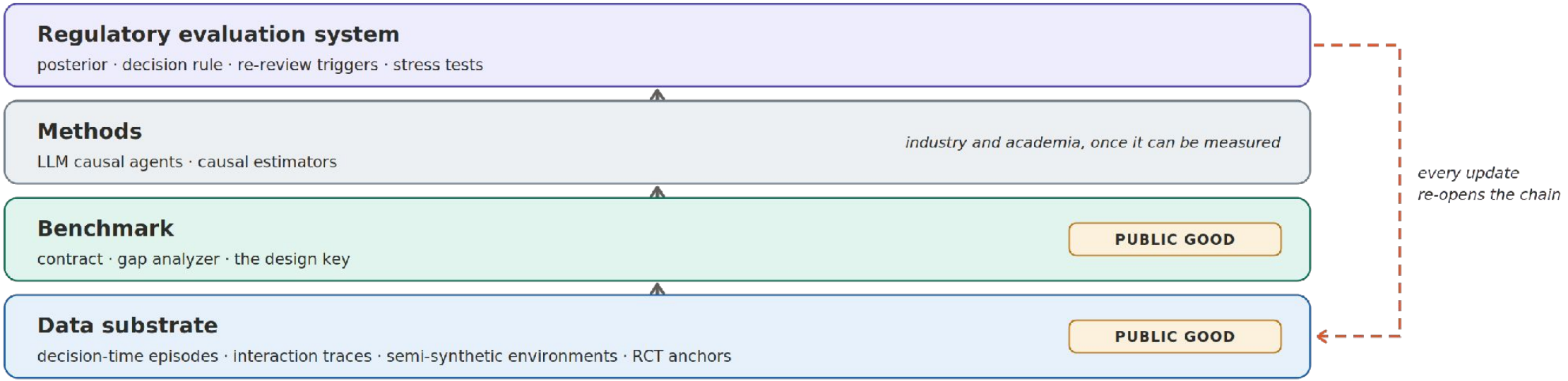
the least developed component is the one the causal question depends on, and it needs the decision context

Reality runs one arm. The simulator runs both.

THE OPPORTUNITY

A Regulatory-Grade Evaluation Substrate

The methods will come from industry and academia once they can be measured. The two layers underneath will not come from anyone.



The substrate comes first. It makes causal evaluation measurable, and then trainable.

APPENDIX

References: The Benchmark Landscape

Cited in the Examples column, slide 3.

Medical knowledge / static QA

MultiMedQA Singhal, K. et al. Large language models encode clinical knowledge. Nature 620, 172–180 (2023).

HealthBench Arora, R. K. et al. HealthBench: Evaluating Large Language Models Towards Improved Human Health. arXiv:2505.08775 (2025).

Simulated clinical interaction

AgentClinic Schmidgall, S. et al. arXiv:2405.07960 (2024); AgentClinic: a multimodal benchmark for tool-using clinical AI agents. npj Digit. Med. 9, 499 (2026).

Clinical Environment Simulator Luo, L. et al. A clinical environment simulator for dynamic AI evaluation. Nat. Med. 32, 820–827 (2026).

EHR-grounded / agentic tasks

MedHELM Bedi, S. et al. arXiv:2505.23802 (2025); Holistic evaluation of large language models for medical tasks with MedHELM. Nat. Med. (2026).

MedAlign Fleming, S. L. et al. MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. AAAI (2024).

EHRSHOT Wornow, M. et al. EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. NeurIPS Datasets & Benchmarks (2023).

EHRSQL Lee, G., Kweon, S., Bae, S. & Choi, E. Overview of the EHRSQL 2024 Shared Task on Reliable Text-to-SQL Modeling on EHRs. ClinicalNLP (2024).

EHRAgent Shi, W. et al. EHRAgent: Code Empowers LLMs for Few-shot Complex Tabular Reasoning on EHRs. EMNLP, 22315–22339 (2024).

EHR-Complex Qiao, Y. et al. EHR-Complex: Benchmarking Medical Agents for Complex Clinical Reasoning. arXiv:2606.23301 (2026).

MedAgentBench Jiang, Y. et al. MedAgentBench: A Virtual EHR Environment to Benchmark Medical LLM Agents. NEJM AI 2(9), AIdbp2500144 (2025).

FHIR-AgentBench Lee, G. et al. FHIR-AgentBench: Benchmarking LLM Agents for Realistic Interoperable EHR Question Answering. arXiv:2509.19319 (2025).

Causal reasoning / causal agents

CLadder Jin, Z. et al. CLadder: Assessing Causal Reasoning in Language Models. NeurIPS (2023).

CausalBench (causal learning) Zhou, Y. et al. CausalBench: A Comprehensive Benchmark for Causal Learning Capability of LLMs. arXiv:2404.06349 (2024).

CausalBench (causal reasoning) Wang, Z. CausalBench: A Comprehensive Benchmark for Evaluating Causal Reasoning Capabilities of LLMs. SIGHAN-10, 143–151 (2024).

CounterBench Chen, Y. et al. CounterBench: Evaluating and Improving Counterfactual Reasoning in Large Language Models. arXiv:2502.11008 (2025); AAAI (2026).

Causal-Copilot Wang, X. et al. Causal-Copilot: An Autonomous Causal Analysis Agent. arXiv:2504.13263 (2025).