EpiWorld: Grounding LLM Policy Agents in Epidemiological World Models

Zeeshan Memon[1], Yiqi Su[2], Kai Shu[1], Naren Ramakrishnan[2], Liang Zhao[1]

[1]Department of Computer Science, Emory University, Atlanta, GA, USA
[2]Department of Computer Science, Virginia Tech, Alexandria, VA, USA
Correspondence: zeeshan.memon@emory.edu

Abstract

Epidemic intervention policies are textual artefacts that human decision-makers interpret, justify, and revise through natural language, making large language models a natural candidate for epidemic policy reasoning. A naive LLM, however, lacks the epidemic dynamics needed to project intervention consequences, the quantitative surveillance signals required to assess severity, and the institutional constraints that define admissible actions. We present EpiWorld, a closed-loop framework that grounds an LLM policy actor in a learned action-conditioned epidemiological world model and a tiered skill library of public-health protocols, surveillance tools, and adaptive lessons accumulated through after-action analysis. Given a candidate intervention, the world model predicts regional epidemic evolution and enables fast counterfactual rollouts that provide feedback for policy selection and refinement. Outcomes of simulated futures are distilled into reusable lessons while protocol constraints remain fixed, allowing the decision process to improve without sacrificing interpretability or controllability. We evaluate both the world model and the end-to-end framework on retrospective COVID-19 and Influenza datasets: the world model achieves the best out-of-distribution Peak-MAE among all forecasting baselines, and the closed-loop framework reduces cumulative hospitalisation by up to 59% across datasets and by an average of $\sim$16% across six LLM backbones, outperforming reinforcement-learning and optimal-control policy baselines.

1 Introduction

Many policy reasoning problems are language-mediated: the relevant knowledge, constraints, and decisions are expressed through natural-language artefacts such as statutes, regulations, advisories, and operational guidance. Epidemic response is one such setting. Public-health authorities communicate interventions through textual policies while simultaneously integrating surveillance observations, forecasting signals, and previously enacted decisions under uncertainty. Because policies influence future observations and must themselves be justified and communicated through language, epidemic response forms a natural testbed for studying whether large language models can act as policy reasoners when grounded in quantitative decision environments.

Epidemic response is fundamentally a problem of decision making over counterfactual futures. Public-health authorities must not only forecast outcomes under current policies, but also reason about what would happen under alternative interventions and which actions should be taken next to optimise public-health objectives such as hospital burden and mortality (Flaxman et al., 2020; Hale et al., 2021). During an evolving outbreak, these decisions are made repeatedly across interacting regions under delayed clinical observations (Gibson et al., 2023), shifting behavioural responses (Williams et al., 2023; Park et al., 2023), and policy-endogenous dynamics, where observed trajectories are themselves consequences of prior interventions (Memon et al., 2026). As a result, epidemic policy optimisation requires models that can evaluate alternative intervention trajectories rather than forecast a single future.

Existing computational tools only partially address this challenge. Statistical and neural epidemic forecasters can predict future hospitalisations, or deaths under observed covariates (Shi et al., 2022; Wan et al., 2025; Du et al., 2025), but they are not designed to answer counterfactual policy questions: they estimate what is likely to happen, not what would happen under alternative interventions. Causal-inference methods estimate retrospective intervention effects (Flaxman et al., 2020), yet struggle with co-occurring policies, delayed outcomes, and prospective multi-step decision making. Mechanistic simulators and reinforcement-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 11885–11906
October 24-29, 2026 ©2026 Association for Computational Linguistics

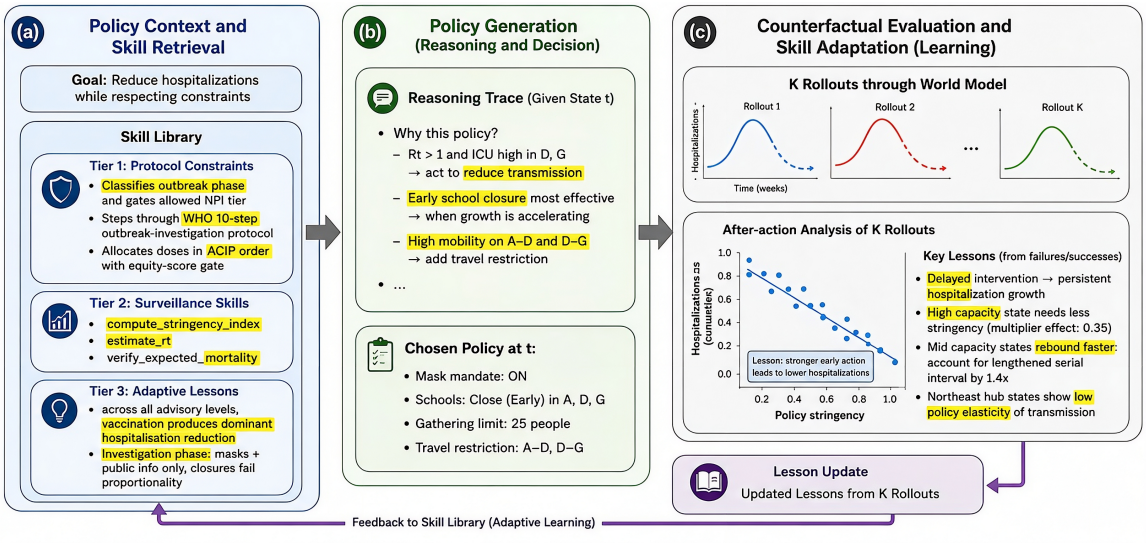

Figure 1: EpiWorld framework. (a) The policy actor retrieves protocol constraints, surveillance skills, and adaptive lessons. (b) The actor generates interventions and reasoning traces. (c) Counterfactual rollouts evaluate candidate policies and produce updated lessons. Yellow highlights denote retrieved evidence and activated skills.

learning approaches support policy optimisation in principle (Williams et al., 2023; Chopra et al., 2024; Bampa et al., 2022; Song et al., 2020; Mai et al., 2023), but depend on predefined dynamics or simulation environments that are difficult to adapt, inspect, and revise. Conversely, large language models operate over textual protocols, advisories, and accumulated lessons, yet lack grounded epidemic dynamics for evaluating the downstream consequences of their recommendations.

The missing object is therefore a grounded language-mediated decision environment: one that allows an LLM policy actor to reason over textual guidance while testing proposed actions against learned policy-conditioned epidemic dynamics. In machine learning, this object is known as a *world model* (Hafner et al., 2021, 2023; Micheli et al., 2022; Hansen et al., 2022): a learned dynamics model that supports planning and policy learning through imagined rollouts (Hafner et al., 2019, 2020). World models have shown strong performance in games and robotics, where agents improve decisions by simulating future consequences before acting. Epidemic world modelling (Memon et al., 2026), however, is not a direct transplant of these settings: epidemic dynamics evolve across interacting regions, delayed clinical signals, and intervention-dependent behavioural responses that alter the observed trajectory itself (Flaxman et al., 2020; Kang et al., 2020). Moreover, epidemic interventions are textual and institutional objects rather than unconstrained actions—they must be justified, communicated, and constrained under public-health protocols (Hale et al., 2021).

We introduce EpiWorld, a closed-loop framework for epidemiological policy optimisation with LLM agents (Figure 1). EpiWorld couples an action-conditioned regional world model with a protocol-guided policy actor. As illustrated in Figure 1(a), the actor first constructs policy context through a tiered epidemiological skill library, consisting of protocol constraints, surveillance tools, and adaptive lessons accumulated from prior experience, mirroring the components of the standard public-health monitoring and evaluation cycle (World Health Organization, 2018). Conditioned on this context, the actor generates interventions and explicit reasoning traces over the current epidemic state (Figure 1(b)). Candidate policies are then evaluated through the regional world model, which predicts how epidemic trajectories evolve under interventions and enables fast counterfactual rollout across interacting regions (Figure 1(c)). Outcomes of simulated futures are retrospectively analysed to update the lesson library and improve future policy generation. Thus, the LLM is not an unconstrained recommender but a language-facing policy reasoner whose proposals remain protocol-constrained and grounded through counterfactual epidemic rollouts, and can assist public-health authorities as a decision-support layer whose recommendations remain subject to their expert approval. Our contributions are as follows:

- We formulate epidemic policy optimisation as

grounded, language-mediated, closed-loop decision making unifying forecasting, counterfactual reasoning, and prospective intervention optimisation within a single rollout-based framework.
- We propose an epidemiological world model for regional epidemic dynamics that integrates delay-aware temporal encoding, graph-structured regional coupling, residual next-state prediction, and policy-conditioned latent modulation.
- We introduce a protocol-guided LLM policy actor that operationalises public-health protocols, surveillance estimators, and after-action lessons as an interpretable and human-intervenable epidemic decision mechanism.

Together, these results suggest a new direction for epidemic AI systems: not merely predicting outbreaks, and not merely prompting LLMs to recommend interventions, but using LLMs as the connective layer between textual policy reasoning and epidemiologically grounded imagined futures.

2 Related Work

Forecasters and LLM proposers. Mechanistic Bayesian models dominate classical epidemic forecasting (Flaxman et al., 2020; Gibson et al., 2023; Qian et al., 2020), and neural extensions replace hand-coded compartments with latent ODEs and graph-CDE encoders (Shi et al., 2022; Wan et al., 2025). LLMs have since entered the loop as trend forecasters (Du et al., 2025; Yu et al., 2026), policy proposers inside mechanistic environments (Aoki and Ghaffarzadegan, 2026; Mao et al., 2025; Shi et al., 2026), agentic simulator builders (Datta et al., 2026a), grounded narrative generators over scenario projections (Datta et al., 2026b), which communicate projections rather than optimise interventions, and generative behavioural agents (Williams et al., 2023; Hou et al., 2025; Park et al., 2023). These methods either fold interventions into covariate weightings or evaluate policies inside a fixed mechanistic simulator. We retain the LLM as the policy proposer and close the loop through a learned world model.

Policy optimisation and skill learning. Optimal control over compartmental or metapopulation dynamics yields analytic policies under assumed dynamics (Alvarez et al., 2021; Acemoglu et al., 2021; Bertsimas et al., 2022), and deep RL trains policies on simulated epidemic environments (Bampa et al., 2022; Song et al., 2020) but transfers poorly across disease or period. LLM-agent work reframes policy improvement as skill accumulation (Wang et al., 2023; Shinn et al., 2023; Xia et al., 2026) and prompt optimisation (Agrawal et al., 2025; Khattab et al., 2023; Yang et al., 2023), with LLM-driven simulators closing the loop (Chopra et al., 2024; Xi et al., 2025). We instantiate this direction for epidemic intervention, accumulating epidemic-specific skills against a learned world model.

World models. World models learn latent dynamics that support policy learning through in-model rollouts (Hafner et al., 2019, 2020), refined via discrete representations, transformer backbones, and model-predictive integration (Hafner et al., 2021, 2023; Micheli et al., 2022; Hansen et al., 2022). This literature is almost exclusively developed in robotics and games, with recent extensions to structured scientific domains such as actionable molecular optimisation (Qiao et al., 2026). Memon et al. (2026) formalise the epidemic case as a controlled, partially-observed system with policy-endogenous observations. We instantiate that framing as an action-conditioned recurrent state-space world model and use it as the rollout substrate for our policy.

3 Method

3.1 Problem Formulation

We consider epidemic policy optimisation over a set of interacting regions. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a regional interaction graph, where each node $i \in \mathcal{V}$ represents a geographic unit and each edge $(i, j) \in \mathcal{E}$ encodes epidemiologically relevant coupling, such as mobility, spatial adjacency, commuting flow, or correlated transmission pressure. Each region is associated with static attributes $\mathbf{s}_i$ describing demographic, socioeconomic, healthcare-capacity, and vulnerability indicators, and dynamic epidemic observations $\mathbf{x}_i^t$, including infections, hospitalisations, mortality, vaccination uptake, testing intensity, and disease severity signals. Aggregating across regions gives the global state $\mathbf{X}^t = \{\mathbf{x}_i^t\}_{i\in\mathcal{V}}$ and static context $\mathbf{S} = \{\mathbf{s}_i\}_{i\in\mathcal{V}}$.

At each decision step t, a policy maker selects interventions $\mathbf{a}^t = \{\mathbf{a}_i^t\}_{i\in\mathcal{V}}$ with $\mathbf{a}_i^t \in \mathcal{A}$, where $\mathcal{A}$ may include non-pharmaceutical interventions, vaccination allocation, school policies, testing strategies, mobility restrictions, or other public-health actions. The epidemic evolves according to unknown policy-conditioned dynamics:

$$\mathbf{X}^{t+1} \sim P\big(\cdot \mid \mathbf{X}^t, \mathbf{S}, \mathbf{a}^t, \mathcal{G}\big)\,. \tag{1}$$

Given a policy objective g, such as minimising cumulative hospitalisations, reducing peak healthcare burden, satisfying intervention budgets, or enforcing equity constraints, the goal is to find a goal-conditioned decision policy $\pi(\mathbf{a}^t \mid \mathbf{X}^t, \mathbf{S}, \mathcal{G}, g)$ that maximises expected utility over horizon H:

$$\pi^\star = \arg\max_{\pi} \; \mathbb{E}_{\tau\sim(\pi,P)} \left[J_g(\tau) \right], \tag{2}$$

where $\tau = (\mathbf{X}^0, \mathbf{a}^0, \mathbf{X}^1, \mathbf{a}^1, \ldots, \mathbf{X}^H)$ denotes the resulting trajectory. This formulation is intentionally method-agnostic and encompasses existing epidemic decision-making paradigms. Human public-health operators incorporate contextual knowledge and operational constraints, but decisions may be subjective and difficult to scale consistently. Mechanistic simulators and optimal-control approaches support rich counterfactual analysis but often depend on hand-designed assumptions and become computationally expensive for large-scale policy search. Reinforcement-learning approaches can optimise long-horizon interventions but typically require simulator access and may offer limited semantic feedback or controllability for domain experts. Surrogate forecasting models are computationally efficient but are primarily optimised for prediction rather than iterative policy evaluation. These limitations motivate the need for models that can combine expressive epidemic dynamics, efficient counterfactual evaluation, and scalable policy exploration.

3.2 EpiWorld Framework

EpiWorld addresses the epidemic policy optimisation problem in Eq. 2 through two coupled components: (i) an epidemiological world model that approximates the unknown policy-conditioned dynamics and enables counterfactual rollout, and (ii) a protocol-guided policy actor that generates interventions from predicted epidemic trajectories. Formally, EpiWorld introduces a world model that learns unknown epidemic dynamics

$$f_\theta : (\mathbf{X}^t, \mathbf{S}, \mathbf{a}^t, \mathcal{G}) \to \hat{\mathbf{X}}^{t+1}, \qquad f_\theta \approx P, \tag{3}$$

which predicts policy-conditioned epidemic evolution and supports multi-step counterfactual rollout. Policy optimisation is then performed through interaction with the world model:

$$\pi^\star \in \arg\max_{\pi} \; \mathbb{E}_{\hat{\tau}\sim(\pi,f_\theta)} \left[J_g(\hat{\tau}) \right], \tag{4}$$

where $\hat{\tau} = (\hat{\mathbf{X}}^1, \hat{\mathbf{a}}^1, \ldots, \hat{\mathbf{X}}^H)$ denotes a simulated trajectory. The closed loop replaces the unknown P in Eq. 2 with the learned f_θ, enabling forecasting, counterfactual evaluation, and policy optimisation within a unified framework.

3.3 Graph-Structured Epidemiological World Model

Generic world models are typically designed for homogeneous sequential prediction and do not explicitly account for delayed disease progression, intervention feedback, or structured transmission across interacting regions. To address these limitations, we specialise the world-model paradigm for epidemic policy reasoning through three epidemiological inductive biases: delay-aware temporal encoding, policy-conditioned latent modulation, and graph-structured regional coupling. The world model introduced in Eq. 3 adopts an encoder–dynamics architecture trained through multi-objective supervision in both latent and observation spaces. Epidemic observables evolve asynchronously: infections precede reported cases, cases precede hospitalisations, and hospitalisations precede mortality signals. To capture both short-term recency and delayed cross-channel effects, we encode each regional trajectory through two complementary temporal branches. For node $i \in \mathcal{V}$ with $\mathbf{x}_i = \mathbf{X}_i^{t-W:t}$, where W denotes the temporal observation window,

$$\mathbf{e}_i^{\mathrm{r}} = \mathrm{GRU}_\theta(\mathbf{x}_i), \; \mathbf{e}_i^{\mathrm{c}} = \mathrm{CDE}_\theta(\mathrm{DelayMix}_\theta(\mathbf{x}_i)), \tag{5}$$

where $\mathrm{DelayMix}$ is a causal cross-channel mixing module and CDE models continuous temporal evolution along the induced path. A learnable scalar gate γ balances the contribution of the two views, $\mathbf{e}_i = [\mathbf{e}_i^{\mathrm{r}}, \; \sigma(\gamma)\mathbf{e}_i^{\mathrm{c}}]$, with σ the logistic function. Actions modify how epidemic trajectories evolve rather than acting as ordinary covariates. The temporal representation is combined with static regional context and projected into a policy-conditioned latent representation $\mathbf{z}_i^t$, which serves as the internal state for epidemic dynamics modelling. We condition regional latent states through feature-wise modulation,

$$\begin{aligned} \mathbf{z}_i^t &= \left(\mathbf{1} + \boldsymbol{\beta}_i^t\right) \odot \mathrm{Proj}_\theta([\phi_s(\mathbf{s}_i), \mathbf{e}_i]) + \boldsymbol{\alpha}_i^t, \\ (\boldsymbol{\alpha}_i^t, \boldsymbol{\beta}_i^t) &= \psi_\theta(\mathbf{a}_i^t). \end{aligned} \tag{6}$$

Here $\mathbf{z}_i^t, \boldsymbol{\alpha}_i^t, \boldsymbol{\beta}_i^t \in \mathbb{R}^{d_z}$, and $\psi_\theta : \mathcal{A} \to \mathbb{R}^{2d_z}$ is a small MLP that maps the per-region action to FiLM-style scale and shift parameters. This conditioning allows interventions to reshape every la-

tent channel rather than entering as additive inputs. Stacking over regions yields $\mathbf{z}^t = (\mathbf{z}_i^t)_{i\in\mathcal{V}}$, which initialises $\mathbf{Z}^{(0)}$. Epidemic evolution exhibits strong inter-region dependency through mobility, adjacency, and correlated behavioural patterns. To model regional transmission while preventing information leakage across unrelated locations, we apply adjacency-masked graph attention:

$$\mathbf{Z}^{(\ell+1)} = \mathrm{FFN}\Big(\mathrm{Attn}(\mathbf{Z}^{(\ell)};\mathcal{G})\Big),$$
$$[\mathrm{Attn}(\mathbf{Z};\mathcal{G})]_{ij} \propto \mathbb{1}[(i,j)\in\mathcal{E}]\exp\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}}\right). \tag{7}$$

After L layers, the model predicts residual updates,

$$\hat{\mathbf{z}}^{\,t+1} = \mathbf{Z}^{(L)}, \qquad \hat{\mathbf{X}}^{\,t+1} = \mathbf{X}^t + \mathrm{Head}_\theta(\hat{\mathbf{z}}^{\,t+1}), \tag{8}$$

which stabilises learning for evolving epidemic trajectories while preserving responsiveness to transmission shocks and interventions. The world model is trained through complementary supervision objectives. A latent transition objective aligns predicted latent dynamics with encoded future representations,

$$\mathcal{L}_{\mathrm{lat}} = \mathbb{E}_{\mathcal{D}_{\mathrm{hist}}}\left[\left\|\hat{\mathbf{z}}^{\,t+1} - \mathrm{Enc}_\theta(\mathbf{X}^{\,t+1-W:t+1},\mathbf{S},\mathbf{a}^{\,t})\right\|^2\right], \tag{9}$$

Here Enc_θ denotes the same encoder used to construct the latent state $\mathbf{z}^t$ in Eq. 6, applied to the next-step window to form the prediction target; inference uses the predictor alone. An observation objective additionally anchors predictions to measured epidemic outcomes,

$$\mathcal{L}_{\mathrm{obs}} = \mathbb{E}_{\mathcal{D}_{\mathrm{hist}}}\left[\mathcal{L}_{\mathrm{rec}}(\hat{\mathbf{X}}^{\,t+1},\mathbf{X}^{\,t+1})\right], \tag{10}$$

where $\mathcal{L}_{\mathrm{rec}}$ is a per-channel Huber loss for robustness to reporting noise and count outliers. The final objective is $\mathcal{L} = \mathcal{L}_{\mathrm{lat}} + \lambda\mathcal{L}_{\mathrm{obs}}$. Together, delay-aware temporal modelling, policy-conditioned latent modulation, and graph-constrained regional interaction transform generic world-model reasoning into an epidemic-specialised dynamics model for counterfactual policy optimisation.

3.4 Protocol-Guided LLM Policy Actor

Predicted trajectories alone are insufficient for policy making: decisions must remain protocol-compliant, draw on quantitative surveillance, and adapt to behavioural feedback from prior responses. EpiWorld therefore implements the policy $\pi(\cdot \mid g)$ as a protocol-guided LLM actor equipped with an explicit skill library. The actor produces interventions through a tiered skill hierarchy,

$$\pi(\mathbf{a}^t \mid \mathbf{X}^t,\mathbf{S},\mathcal{G},g,\mathcal{T}), \qquad \mathcal{T} = \mathcal{T}_1\cup\mathcal{T}_2\cup\mathcal{T}_3, \tag{11}$$

where the three tiers correspond to Protocol Constraints, Surveillance Skills, and Adaptive Lessons. Each tier reflects a distinct mode of public-health knowledge, regulatory protocols, quantitative surveillance, and operational learning, grounded respectively in the International Health Regulations (World Health Organization, 2008), the Joint External Evaluation tool (World Health Organization, 2022), and WHO After-Action Review guidance (World Health Organization, 2019); our contribution exposes each layer as a skillset for agents.

Protocol Constraints ($\mathcal{T}_1$). Protocol Constraints are deterministic procedures codified by international and national public-health authorities: pandemic-interval classification, ACIP vaccine-allocation ordering, NPI proportionality validation, and IHR notification thresholds. Each skill $\kappa\in\mathcal{T}_1$ returns a constraint predicate $c_\kappa(\mathbf{a},\mathbf{X}^t,g)\in\{0,1\}$. The resulting admissible action envelope is

$$\mathcal{A}_g^t = \left\{\mathbf{a}\in\mathcal{A} \;\middle|\; c_\kappa(\mathbf{a},\mathbf{X}^t,g)=1,\ \forall\kappa\in\mathcal{T}_1\right\}, \tag{12}$$

which is computed once per decision step and remains fixed during policy generation.

Surveillance Skills ($\mathcal{T}_2$). Surveillance Skills are analytical tools epidemiologists use: R_t estimators (Cori et al., 2013) quantify transmission acceleration, severity ratios (CFR, IFR) project hospital load, the OxCGRT stringency index (Hale et al., 2021) aggregates NPI levels into a scalar, and the ATSDR Social Vulnerability Index flags vulnerable populations. Each skill $\kappa\in\mathcal{T}_2$ contains parameters $\phi_\kappa\in B_\kappa$ and returns quantitative summaries

$$\hat{y}_\kappa = \kappa(\mathbf{X}^t,\mathbf{S},\mathcal{G};\phi_\kappa), \tag{13}$$

which provide interpretable evidence for policy generation under the admissible action set $\mathcal{A}_g^t$.

Adaptive Lessons ($\mathcal{T}_3$). Adaptive Lessons store reusable lessons accumulated from previous world-model rollouts. This mirrors the After-Action Review framework used by WHO and FEMA to convert past outbreaks into reusable operational knowledge (World Health Organization, 2019). Each lesson is represented as $\ell = (\tau_\ell,\rho_\ell,\boldsymbol{\delta}_\ell)$, where τ_ℓ

denotes behavioural and regional tags, $\rho_\ell \in [0, 1]$ is a confidence score, and $\boldsymbol{\delta}_\ell$ is a parameter adjustment targeting Surveillance Skill parameters. At decision step t, the actor extracts tags from the current context, $\tau(\mathbf{X}^t, \mathbf{S})$, retrieves the top-$k$ matching lessons $L^t \subset \mathcal{T}_3$, and refines surveillance parameters through,

$$\phi'_\kappa = \mathrm{clip}_{B_\kappa}\left(\phi_\kappa + \sum_{\ell \in L^t} \rho_\ell \boldsymbol{\delta}_\ell^{(\kappa)}\right), \tag{14}$$

allowing lessons accumulated in one region to transfer to another with a similar behavioural signature within the same infection regime. Lessons are calibrated to the regime they are distilled from and are not transferable across arbitrary regimes or world models (§5.2). Adaptive Lessons can refine Surveillance Skill estimates but cannot expand the Protocol-Constraint admissible action envelope. At inference, the actor applies $\mathcal{T}_1$ to fix $\mathcal{A}_g^t$, retrieves and applies $\mathcal{T}_3$ lessons to refine $\mathcal{T}_2$ parameters, and conditions on the resulting $\mathcal{T}_2$ estimates to propose an action within $\mathcal{A}_g^t$. The examples of these skills are shown in Appendix E; constraint-satisfaction rates over real-US decisions are reported in §5.2.

3.5 Closed-Loop Skill Adaptation

The protocol-guided actor improves through an after-action adaptation cycle that updates the Adaptive-Lesson bank from simulated epidemic rollouts, leaving Protocol Constraints and Surveillance Skill structures untouched and inference cost unchanged. After every K closed-loop rollout episodes $\{\hat{\tau}^{(k)}\}_{k=1}^K$ generated by the world model, a teacher module performs retrospective analysis over trajectory outcomes. The teacher focuses on policy-relevant events including missed objectives, delayed interventions, epidemic rebounds, capacity-limit violations, and instability in intervention timing. Each analysed episode is distilled into an Adaptive Lesson defined in Section 3.4. The lesson bank is updated through additive insertion with supersession,

$$\mathcal{T}_3 \leftarrow (\mathcal{T}_3 \setminus \mathcal{S}(\ell)) \cup \{\ell\}, \tag{15}$$

where $\mathcal{S}(\ell)$ denotes lessons retired due to contradiction, obsolescence, or lower confidence under newly observed evidence. Unlike parameter optimisation in conventional policy learning, this update preserves accumulated behavioural memory while allowing outdated guidance to be replaced. Each lesson is indexed by the tag set $\tau(\mathbf{X}^t, \mathbf{S})$ of the rollout from which it was distilled. The adaptation loop operates at a slower time scale than policy generation: the actor selects interventions at each rollout step, while lesson extraction and bank updates occur periodically every K episodes. This separation keeps online policy inference unchanged while allowing the decision process to improve through accumulated epidemic experience. The selection criteria, confidence weighting, supersession rule and retrieval ranking are specified in Appendix L.

4 Experiments

4.1 Data

We evaluate EpiWorld on three datasets. **Real US COVID-19**: state-level weekly surveillance for 51 US states (Dec 2020 – Feb 2022) from HHS Protect (hospitalisations) and JHU CSSE (cases, deaths), spanning the Alpha, Delta, and Omicron waves; interventions follow the 13-dim OxCGRT NPI grid (Hale et al., 2021) (full grid in Appendix D). **COVID-19 SEIR**: mechanistic per-state SEIR(A) generator with policy-conditioned transmission and GeoDS mobility coupling (Kang et al., 2020); real US CUSP stay-at-home data calibrates simulator policy windows; the action space is a 5-tier intensity scalar over 50 US states and 183 days. **Influenza SEIR**: SEIRV variant with seasonal forcing, waning immunity, and a vaccination compartment; per-state parameters are calibrated against CDC FluView weekly ILI (Chakraborty et al., 2018) so trajectories match real US state seasonality; policy is a 2-dim action (advisory intensity, vaccination tier), distinct from COVID's 13-dim grid, testing cross-pathogen generality.

4.2 Baselines and Evaluation

We compare against two families of baselines.
World-model forecasting. Three architectural families of epidemic forecasters: Bayesian compartmental (Compartmental-GP (Qian et al., 2020), MechBayes (Gibson et al., 2023), and epidemia (Flaxman et al., 2020)), latent ODE (PAN-cODE (Shi et al., 2022)), and neural CDE (EARTH (Wan et al., 2025)). Each is evaluated in both a policy-blind variant (closest to the published model) and a policy-conditioned variant (+ *Policy*) that ingests per-week policy intensity through the same control interface as EpiWorld, ensuring like-for-like comparison. In Table 1, + *Policy* denotes the policy-conditioned variant in

Method	*COVID-19 SEIR*		*Influenza SEIR*		*COVID-19 Real US (Omicron)*	
	In-Dist. Peak / MAE	OOD Peak / MAE	In-Dist. Peak / MAE	OOD Peak / MAE	In-Dist. Peak / MAE	OOD Peak / MAE
Compartmental-GP	0.062 / 0.051	0.558 / 0.533	0.354 / 0.331	1.139 / 1.084	0.332 / 0.178	1.102 / 1.068
MechBayes (naive)	0.730 / 0.700	0.860 / 0.831	4.630 / 4.560	1.750 / 1.681	2.550 / 2.480	0.980 / 0.930
MechBayes + Policy	0.470 / 0.448	0.490 / 0.461	2.970 / 2.905	0.980 / 0.929	1.700 / 1.620	0.630 / 0.590
epidemia Mode A (naive)	0.680 / 0.648	0.750 / 0.709	4.300 / 4.185	1.500 / 1.431	2.400 / 2.310	0.910 / 0.860
epidemia Mode B (pretrained)	0.210 / 0.194	0.250 / 0.228	0.770 / 0.723	0.500 / 0.460	0.430 / 0.390	0.310 / 0.280
PAN-cODE†	0.105 / 0.084	0.462 / 0.504	0.607 / 0.548	0.956 / 1.016	0.553 / 0.366	0.912 / 0.903
PAN-cODE + Policy†	0.105 / 0.091	0.427 / 0.448	0.602 / 0.592	0.869 / 0.901	0.531 / 0.319	0.848 / 0.806
EARTH	0.067 / 0.065	0.226 / 0.251	0.383 / 0.423	0.466 / 0.507	0.356 / 0.239	0.468 / 0.271
EARTH + Policy	**0.069 / 0.066**	**0.224 / 0.224**	**0.397 / 0.432**	**0.451 / 0.452**	**0.361 / 0.225**	**0.451 / 0.224**
EpiWorld	**0.063 / 0.061**	**0.210 / 0.237**	**0.361 / 0.398**	**0.435 / 0.479**	**0.339 / 0.211**	**0.414 / 0.237**

Table 1: World-model forecast performance (12-week autoregressive rollout, log1p per-100k). Each cell reports *Peak-MAE / MAE*; lower is better in both. †PAN-cODE forecasts hospitalisations and deaths only.

Method	*COVID-19 Real US* (Omicron)	*COVID-19 SEIR*	*Influenza SEIR*
	Hosp. Δ (8w) $\downarrow$	Hosp. Δ (8w) $\downarrow$	Hosp. Δ (8w) $\downarrow$
EpiPolicy-RL	-22.2%	-8.2%	-12.4%
ADIOS (GGGP)	-33.4%	-12.5%	-18.6%
EpidRLearn	$\mathbf{-37.5\%}$	$\mathbf{-13.9\%}$	$\mathbf{-20.9\%}$
EpiWorld (Base LLM)	-46.0%	-17.1%	-25.6%
EpiWorld (with Static Skills)	-48.0%	-17.8%	-26.7%
EpiWorld (with Skill Evolution)	$\mathbf{-59.3\%}$	$\mathbf{-22.0\%}$	$\mathbf{-33.0\%}$

Table 2: Policy optimisation performance over 8-week rollouts, reported as cumulative hospitalisation reduction against the reference historical policy (lower is better). For scale, the reference historical policy yields 406 cumulative hospitalisations per 100k over the 8-week horizon against 1,544 under no intervention. Policy-alignment diagnostics are reported separately in Appendix J.

each family; epidemia's *Mode A* fits per replicate while *Mode B* pre-fits a global policy effect on the training corpus. Per-baseline architectural details and adaptations to our corpora are in Appendix B.

Policy optimisation. Three intervention-planning families: PPO/actor-critic over the EpiPolicy simulator (EpiPolicy-RL (Mai et al., 2023)), grammar-guided genetic programming (ADIOS (Wolpers et al., 2026)), and DQN with prioritised replay on a contact-based SEIR (EpidRLearn (Bampa et al., 2022)). All policy baselines are trained and rolled out against the same world model EpiWorld's policy actor uses, on the same evaluation harness. Codebooks, reward functions, and training hyperparameters are in Appendix B.

Evaluation. Each dataset reports three observables per state per week (COVID: hospitalisations, cases, deaths per 100k; Influenza: weighted ILI, flu hospitalisations, flu mortality per 100k), with hospitalisations as the primary policy-relevant target. Forecast quality is reported as mean absolute error (MAE) and peak-MAE on the 12-week autoregressive rollout. Policy quality is measured by a goal-aligned trajectory score: cumulative hospitalisation reduction under the proposed policy relative to the reference historical policy. A policy-alignment diagnostic (activation- and exact-match rate of proposed actions against historical actions) is reported in Appendix J as a plausibility check on the proposed action distribution rather than as a quality metric. Three OOD splits are evaluated (temporal, state-wise, intervention-OOD); see Appendix A for the mechanism behind each.

5 Results

5.1 World Model Performance

Table 1 shows world-model performance. EpiWorld outperforms all baselines on OOD Peak-MAE across the three datasets, with gains of $\sim 6\%$, $\sim 4\%$, and $\sim 8\%$ on COVID-19 SEIR, Influenza SEIR, and Real US (Omicron) respectively (versus the closest competitor, EARTH + Policy). The largest gain comes on the real-data Omicron setting, where unseen-variant dynamics expose the limits of baselines that rely on fixed mechanistic structure; the smallest comes on Influenza SEIR, where the simulator's explicit seasonal forcing and the simpler 2-dim action space narrow the gap. The

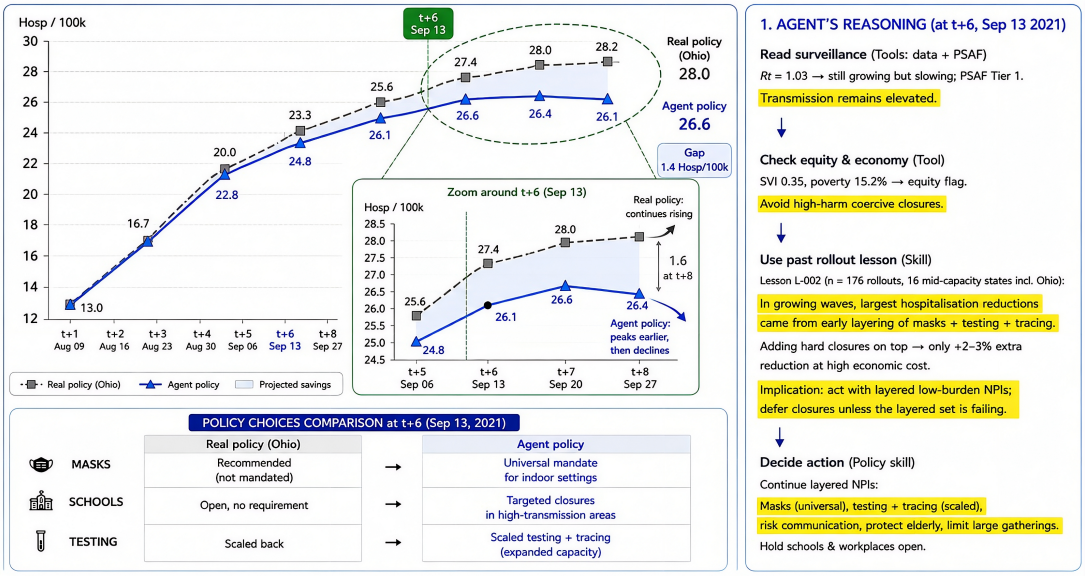

Figure 2: Case study: Ohio during the Delta wave. Highlighted text indicates retrieved key reasoning steps.

component ablation is repeated on influenza in Appendix I, where the same components remain load-bearing. The Bayesian compartmental family is hampered by rigid SEIR assumptions and improves only when policy is made explicit (MechBayes + Policy roughly halves OOD MAE), while the latent-ODE family lacks graph structure and trails uniformly. The neural-CDE EARTH is the closest competitor through graph coupling and reactive dynamics. EpiWorld outperforms both families by combining graph coupling with policy-conditioned latent modulation and delay-aware multi-channel decoding, giving stronger transfer to unseen dynamics. On COVID-19 SEIR state-OOD the Peak-MAE advantage over EARTH + Policy holds at paired Wilcoxon $p = 0.039$ across 50 held-out replicates.

A five-member ensemble of independently seeded world models supplies a decision-time reliability signal: ensemble spread is strongly rank-correlated with realised forecast error (Spearman $\rho = 0.83$, $p \approx 0$), identifying unreliable forecasts before the error is observable. The ensemble is under-dispersed — nominal 95% intervals cover $\sim$68% of outcomes overall and 73% on the hospitalisation channel — so spread should be read as a relative reliability signal rather than a calibrated interval.

5.2 Policy Optimisation

Table 2 reports closed-loop policy performance. EpiWorld with full Skill Evolution outperforms the RL baselines on all three datasets, achieving a $\sim$ 58% further reduction in cumulative hospitalisation relative to the strongest baseline EpidRLearn (-59.3% vs. -37.5% on COVID-19 Real US). The Base LLM already improves on every RL baseline before any after-action signal, indicating that protocol-conditioned proposals alone carry part of the gain. Static skills add modest gains ($\sim$1–2 points); the Skill Evolution cycle contributes the largest single component (4–11 points), peaking on the real-data Omicron setting. The same pattern holds in the mean per-100k hospitalisation metric and on the different flu NPI grid (Appendix I).

Method	In-model	Indep. sim	Ratio
EpiPolicy-RL	-8.2	-8.04	0.98
ADIOS (GGGP)	-12.5	-12.13	0.97
EpidRLearn	-13.9	-14.73	1.06
EpiWorld (Base LLM)	-17.1	-18.30	1.07
EpiWorld (Static Skills)	-17.8	-19.58	1.10
EpiWorld (Skill Evol.)	$\mathbf{-22.0}$	$\mathbf{-25.52}$	1.16

Table 3: Independent-simulator validation, COVID-19 SEIR state-OOD. Cumulative hospitalisation reduction (%) against the reference historical policy, scored inside the learned world model f_θ and re-executed in the mechanistic simulator that generated the corpus. Ratio is off-model over in-model. Real US is observational data and admits no ground-truth generator.

Scoring every policy inside the world model it was optimised against could reward exploiting the idiosyncrasies of f_θ rather than genuine policy quality. We therefore re-executed all six methods in the mechanistic SEIR simulator that generated the corpus, which reproduces the released trajectories at maximum absolute error 0. Table 3 reports both evaluations. The ordering is preserved exactly

Model	Base LLM	w/ Static Skills	w/ Skill Evolution	$\Delta\%$
Proprietary				
GPT-4o	0.1334	0.1371	0.0936	$\mathbf{-29.8}\%$
GPT-5-mini	0.1252	0.1468	0.1172	-6.4%
GPT-5	0.1413	0.1468	0.1395	-1.3%
Open-weights				
Qwen2.5-7B	0.0884	0.0875	0.0692	-21.6%
DeepSeek-R1	0.1332	0.1370	0.1174	-11.9%
Llama-3.3-70B	0.1243	0.1247	0.0924	-25.7%

Table 4: Cross-backbone ablation on COVID-19 SEIR (state-OOD). Cumulative hospitalisation per 100k under the proposed policy over the 8-week post-intervention horizon (lower is better). $\Delta\%$ is the relative reduction of the full Skill Evolution configuration over Base LLM.

across all six methods, each landing within $\times 0.97$–1.16 of its in-model value, so world-model fidelity does not alter the relative ranking. The margin over the strongest baseline, EpidRLearn, is larger on independent ground truth ($+73\%$) than in-model ($+58\%$), and the ratio rises monotonically with configuration quality: f_θ understates rather than inflates the advantage of the closed-loop policy.

Because Tier-1 constraints bound the action set the actor reasons over, their effect is measurable in the proposals themselves. Across 2,640 real-US decisions the envelope admits at most hand hygiene in 42% of decisions, mild measures in 21% and a closure in 37%, and never a full stay-at-home order. 59% of proposed actions fall within it; of the 41% that do not, 39 points are the actor holding masking and testing in place while a wave declines and the envelope has narrowed, at almost no burden. The remaining 2 points are the only unsafe departures and are the decisions surfaced for human approval.

Table 4 isolates where the gain comes from. The full skill cycle reduces hospitalisation across all six backbones (average $\sim 16\%$), while static skills alone leave Base-LLM performance roughly unchanged: lessons are the load-bearing component. GPT-4o achieves the largest reduction at -30%; the much smaller Qwen2.5-7B still reaches -22%, indicating that the gain is not limited by model scale only.

Replacing the actor with a fixed severity-proportional rule that selects from the same skill library recovers only 59% of the Base LLM's reduction at matched intervention intensity (Appendix K): the gain comes from timing and targeting rather than from stronger measures.

Adaptive Lessons generalise across states within an infection regime, but are calibrated to the regime they are distilled from. Mined on {California, New York} and applied to disjoint held-out OOD states on the independent simulator, they reduce cumulative hospitalisation by a further 9.0% over the Base LLM (12/12 seeds), and by 4.3% over the tools-only configuration (11/12). Across regimes the sign reverses: a bank distilled under Omicron, where the marginal return on additional stringency is small, shrinks the expected effect size ($\delta = -0.234$); applied to a first wave in a fully susceptible population ($R_0 \approx 1.7$–2.9) that shrinkage under-responds and degrades outcomes by 12.1% (Appendix I). We therefore mine a separate lesson bank per infection setting; lessons are not universally transferable across arbitrary world models. The same regime-dependence appears within a single setting: on states whose baseline error is already low (FL, NJ, CA) the multiplicative correction over-applies and per-state MAE regresses (Appendix I), so the correction is most useful where there is substantial error to remove. The world-model design ablation is reported in Appendix I.

5.3 Case Study: Ohio, Delta Wave

We analyse one held-out decision point to show how retrieved skills inform policy. Under Ohio's historical interventions, the world model reproduces the observed trajectory (MAE$=$ 1.56 per 100k). When the agent intervenes at $t{+}6$, the trajectory bends downward, reducing peak hospital burden by 1.68 per 100k ($\approx 6\%$ at horizon). The proposed policy is not generated from hindsight labels: the agent converges on layered mitigation (masking, testing, tracing, targeted protection, and risk communication) while avoiding broad closures. This broadly aligns with retrospective Delta-wave analyses (Brauner et al., 2021; Haug et al., 2020; Sachs et al., 2022; Talic et al., 2021), and the retrieved lesson L-002 prioritises layered low-burden NPIs over stringency escalation, consistent with public-health reviews reporting diminishing returns from blanket closures (Bjørnskov, 2021).

6 Conclusion

We presented EpiWorld, a closed-loop framework that grounds LLM policy agents in a learned epidemiological world model and protocol-guided skills to generate interpretable policies through counterfactual evaluation. Across all epidemic settings, EpiWorld improved out-of-distribution forecasting and closed-loop policy outcomes, suggesting that LLMs can serve as useful public-health decision-support tools under human oversight.

Limitations

EpiWorld is intended to assist public-health decision-makers with learned policy proposals and counterfactual rollouts, not to automate them; final decisions and accountability remain with human operators. The evaluations reported here follow the standard world-model protocol of forecast accuracy on out-of-distribution states and counterfactual policy rollouts in the learned dynamics. Re-executing every policy in an independent mechanistic simulator preserves the relative ranking (§5.2), so the comparison does not depend on the fidelity of f_θ alone; both evaluations remain simulator-grounded and retrospective rather than prospective, and cannot substitute for evaluation in a live deployment. In a real setting the system would face behavioural fatigue, surveillance heterogeneity, and the political and economic constraints that fall outside the present goal space and the world model's training support, and counterfactuals over action combinations absent from historical data rely on extrapolation. Adaptive Lessons are calibrated to the infection regime they are distilled from and do not transfer across regimes, so a bank must be mined for each setting. The five-member ensemble reported in §5.1 ranks forecast reliability well but is under-dispersed, so its intervals should not be read as calibrated. Operational use would require dedicated validation studies, equity audits, and continuous human oversight.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) under Grant Nos. 2432418, 2414115, 2403312, 2318831, and 2521631; the National Institutes of Health (NIH) under Grant Nos. R01AG089806 and R01CA297856; and a Sony Research Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, NIH, or other funding organizations.

This work is supported in part by US National Science Foundation grants IIS-2312794 and DBI-2412389. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsor(s).

References

Daron Acemoglu, Victor Chernozhukov, Iván Werning, and Michael D Whinston. 2021. Optimal targeted lockdowns in a multigroup sir model. *American Economic Review: Insights*, 3(4):487–502.

Agency for Toxic Substances and Disease Registry. 2022. CDC/ATSDR Social Vulnerability Index 2020 Database, U.S. https://www.atsdr.cdc.gov/place-health/php/svi/svi-data-documentation-download.html. Accessed 2026-08-30.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, and 1 others. 2025. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*.

Fernando Alvarez, David Argente, and Francesco Lippi. 2021. A simple planning problem for covid-19 lockdown, testing, and tracing. *American Economic Review: Insights*, 3(3):367–382.

Goshi Aoki and Navid Ghaffarzadegan. 2026. Ai agents as policymakers in simulated epidemics. *arXiv preprint arXiv:2601.04245*.

Maria Bampa, Tobias Fasth, Sindri Magnusson, and Panagiotis Papapetrou. 2022. Epidrlearn: Learning intervention strategies for epidemics with reinforcement learning. In *International Conference on Artificial Intelligence in Medicine (AIME)*, pages 189–199. Springer.

Dimitris Bertsimas, Vassilis Digalakis Jr, Alexander Jacquillat, Michael Lingzhi Li, and Alessandro Previero. 2022. Where to locate covid-19 mass vaccination facilities? *Naval Research Logistics (NRL)*, 69(2):179–200.

Christian Bjørnskov. 2021. Did lockdown work? An economist's cross-country comparison. *CESifo Economic Studies*, 67(3):318–331.

Jan M Brauner, Sören Mindermann, Mrinank Sharma, David Johnston, John Salvatier, Tomáš Gavenčiak, Anna B Stephenson, Gavin Leech, George Altman, Vladimir Mikulik, and 1 others. 2021. Inferring the effectiveness of government interventions against COVID-19. *Science*, 371(6531):eabd9338.

Prithwish Chakraborty, Bryan Lewis, Stephen Eubank, John S Brownstein, Madhav Marathe, and Naren Ramakrishnan. 2018. What to know before forecasting the flu. *PLoS Computational Biology*, 14(10):e1005964.

Ayush Chopra, Shashank Kumar, Nurullah Giray-Kuru, Ramesh Raskar, and Arnau Quera-Bofarull. 2024. On the limits of agency in agent-based models. *arXiv preprint arXiv:2409.10568*.

Anne Cori, Neil M Ferguson, Christophe Fraser, and Simon Cauchemez. 2013. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512.

Rituparna Datta, Zihan Guan, Baltazar Espinoza, Yiqi Su, Priya Pitre, Srini Venkatramanan, Naren Ramakrishnan, and Anil Vullikanti. 2026a. Agentic framework for epidemiological modeling. *arXiv preprint arXiv:2602.00299*.

Rituparna Datta, Srini Venkatramanan, Bryan L. Lewis, Yiqi Su, Harry Hochheiser, Lucie Contamin, Parantapa Bhattacharya, Naren Ramakrishnan, and Anil Vullikanti. 2026b. EpiNarrate: Agentic generation of grounded narratives from epidemiological scenario projections. *arXiv preprint arXiv:2607.15544*.

Hongru Du, Yang Zhao, Jianan Zhao, Shaochong Xu, Xihong Lin, Yiran Chen, Lauren M Gardner, and Hao Frank Yang. 2025. Advancing real-time infectious disease forecasting using large language models. *Nature Computational Science*, 5(6):467–480.

Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, and 1 others. 2020. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261.

Graham C Gibson, Nicholas G Reich, and Daniel Sheldon. 2023. Real-time mechanistic Bayesian forecasts of COVID-19 mortality. *The Annals of Applied Statistics*, 17(3):1801.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2020. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. 2019. Learning latent dynamics for planning from pixels. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pages 2555–2565.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. 2021. Mastering Atari with discrete world models. In *International Conference on Learning Representations (ICLR)*.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, and 1 others. 2021. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, 5(4):529–538.

Nicklas Hansen, Xiaolong Wang, and Hao Su. 2022. Temporal difference learning for model predictive control. In *Proceedings of the International Conference on Machine Learning (ICML)*. ArXiv:2203.04955.

Nils Haug, Lukas Geyrhofer, Alessandro Londei, Elma Dervic, Amelie Desvars-Larrive, Vittorio Loreto, Beate Pinior, Stefan Thurner, and Peter Klimek. 2020. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nature Human Behaviour*, 4:1303–1312.

Abe Bohan Hou, Hongru Du, Yichen Wang, Jingyu Zhang, Zixiao Wang, Paul Pu Liang, Daniel Khashabi, Lauren Gardner, and Tianxing He. 2025. Can a society of generative agents simulate human behavior and inform public health policy? a case study on vaccine hesitancy. *arXiv preprint arXiv:2503.09639*.

Yuhao Kang, Song Gao, Yunlei Liang, Mingxiao Li, Jinmeng Rao, and Jake Kruse. 2020. Multiscale dynamic human mobility flow dataset in the U.S. during the COVID-19 epidemic. *Scientific Data*, 7(1):390.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, and 1 others. 2023. DSPy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Anh Mai, Nikunj Gupta, Azza Abouzied, and Dennis Shasha. 2023. Planning multiple epidemic interventions with reinforcement learning. *arXiv preprint arXiv:2301.12802*.

Kangkun Mao, Fang Xu, Jinru Ding, Yidong Jiang, Yujun Yao, Yirong Chen, Junming Liu, Xiaoqin Wu, Qian Wu, Xiaoyan Huang, and 1 others. 2025. Epiplanagent: Agentic automated epidemic response planning. *arXiv preprint arXiv:2512.10313*.

Zeeshan Memon, Yiqi Su, Christo Kurisummoottil Thomas, Walid Saad, Liang Zhao, and Naren Ramakrishnan. 2026. Toward world models for epidemiology. *arXiv preprint arXiv:2604.09519*.

Vincent Micheli, Eloi Alonso, and François Fleuret. 2022. Transformers are sample-efficient world models. *arXiv preprint arXiv:2209.00588*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–22.

Zhaozhi Qian, Ahmed M Alaa, and Mihaela van der Schaar. 2020. When and how to lift the lockdown? Global COVID-19 scenario analysis and policy assessment using compartmental Gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 10729–10740.

Yang Qiao, Bo Pan, Hao-Wei Pang, Peter Zhiping Zhang, Liying Zhang, and Liang Zhao. 2026. MolWorld: Molecule world models for actionable molecular optimization. *arXiv preprint arXiv:2605.08954*.

Jeffrey D. Sachs, Salim S Abdool Karim, Lara Aknin, Joseph Allen, Kirsten Brosbøl, Francesca Colombo, and 1 others. 2022. The Lancet Commission on lessons for the future from the COVID-19 pandemic. *The Lancet*, 400(10359):1224–1280.

Ruian Shi, Haoran Zhang, and Quaid Morris. 2022. PAN-cODE: COVID-19 forecasting using conditional latent ODEs. *Journal of the American Medical Informatics Association*, 29(12):2089–2095.

Ziyi Shi, Xusen Guo, Hongliang Lu, Mingxing Peng, Haotian Wang, Zheng Zhu, Zhenning Li, Yuxuan Liang, Xinhu Zheng, and Hai Yang. 2026. Coordinated pandemic control with large language model agents as policymaking assistants. *arXiv preprint arXiv:2601.09264*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 8634–8652.

Sirui Song, Zefang Zong, Yong Li, Xue Liu, and Yang Yu. 2020. Reinforced epidemic control: Saving both lives and economy. *arXiv preprint arXiv:2008.01257*.

Stella Talic, Shivangi Shah, Holly Wild, Danijela Gasevic, Ashika Maharaj, Zanfina Ademi, Xue Li, Wei Xu, Ines Mesa-Eguiagaray, Jasmin Rostron, Evropi Theodoratou, Xiaomeng Zhang, Ashmika Motee, Danny Liew, and Dragan Ilic. 2021. Effectiveness of public health measures in reducing the incidence of covid-19, SARS-CoV-2 transmission, and covid-19 mortality: systematic review and meta-analysis. *BMJ*, 375:e068302.

U.S. Census Bureau. American Community Survey. `https://www.census.gov/programs-surveys/acs`. Accessed 2026-08-30.

Guancheng Wan, Zewen Liu, Xiaojun Shan, Max S. Y. Lau, B. Aditya Prakash, and Wei Jin. 2025. EARTH: Epidemiology-aware neural ODE with continuous disease transmission graph. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

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

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

Anja Wolpers, Johannes Ponge, and Adelinde M Uhrmacher. 2026. Optimizing interventions for agent-based infectious disease simulations. *arXiv preprint arXiv:2604.02016*.

World Health Organization. 2008. *International Health Regulations (2005)*, 2nd edition. World Health Organization, Geneva. Accessed 2026-08-30.

World Health Organization. 2018. International Health Regulations (2005): IHR Monitoring and Evaluation Framework. Technical Report WHO/WHE/CPI/2018.51, World Health Organization, Geneva. Accessed 2026-08-30.

World Health Organization. 2019. Guidance for After Action Review (AAR). Technical Report WHO/WHE/CPI/2019.4, World Health Organization, Geneva. Accessed 2026-08-30.

World Health Organization. 2022. *Joint External Evaluation Tool: International Health Regulations (2005)*, 3rd edition. World Health Organization, Geneva. Accessed 2026-08-30.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, and 1 others. 2025. Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning. *arXiv preprint arXiv:2509.08755*.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, and 1 others. 2026. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Xiao Yu, Baolin Peng, Ruize Xu, Yelong Shen, Pengcheng He, Suman Nath, Nikhil Singh, Jiangfeng Gao, and Zhou Yu. 2026. Reinforcement world model learning for llm-based agents. *arXiv preprint arXiv:2602.05842*.

A Datasets and Splits

Real US COVID-19. State-level weekly surveillance for 51 US states (50 + DC) over Dec 2020 – Feb 2022, with interventions from the 13-dimensional OxCGRT NPI grid (Hale et al., 2021). Hospitalisations come from HHS Protect, cases and deaths from JHU CSSE, mobility from GeoDS (Kang et al., 2020), and static demographics inherited from PandemicLLM (Du et al., 2025), which assembles them from the American Community Survey (U.S. Census Bureau), the Commonwealth Fund state health-system scorecard, and

2020 FEC election returns, together with the ATSDR/CDC Social Vulnerability Index (Agency for Toxic Substances and Disease Registry, 2022). The world model is trained on Dec 2020 – Sep 2021 (Alpha and Delta waves) and evaluated on the held-out Jan 2022 Omicron wave plus eight subsequent weeks, giving a temporal-OOD setting.

COVID-19 SEIR (corpus `scenarios_v2`). A per-state SEIR(A) generator with GeoDS mobility coupling, run under policy windows derived from real US CUSP stay-at-home data. The action grid is a five-tier intensity scalar $\{0, 0.25, 0.5, 0.75, 1.0\}$. Training comprises 30 scenarios $\times$ 5 replicates (150 trajectories) over 40 train states; the in-distribution test set is 10 scenarios $\times$ 5 replicates on the same train states; the state-OOD test set is 10 scenarios $\times$ 5 replicates on 10 held-out states (DE, KS, MI, MN, MS, NV, NJ, NC, SC, WA). Each replicate spans 183 days.

Influenza SEIR (corpus `flu_scenarios_v1`). A per-state SEIRV generator with seasonal $\beta(t)$ forcing, waning immunity ($R \to S$, $V \to S$), and a vaccination compartment. Per-state (β_0, α_{season}, wILI scale) are calibrated against CDC FluView weekly ILI for the 2017–18 season (Chakraborty et al., 2018); FluView and FluVaxView data parameterise the simulator and do not enter evaluation directly. The action grid is two-dimensional (advisory intensity, vaccination tier). Training and state-OOD test splits mirror COVID-19 SEIR (30 training scenarios, 10 OOD scenarios, 5 replicates each, 10 held-out states). Each replicate spans 240 days.

OOD splits. **Temporal OOD** (Real US COVID-19): training and test windows occupy different epidemic waves – training uses Dec 2020 – Sep 2021 (Alpha + Delta), while the test window is the Jan 2022 Omicron wave plus eight subsequent weeks; the Omicron-era dynamics (immune escape, faster generation interval, milder severity) are absent from training. **State-wise OOD** (COVID-19 SEIR and Influenza SEIR): a disjoint set of 10 states is held out from training entirely; evaluation runs only on the held-out states under the same disease and policy distributions as training. **Intervention-OOD** (Influenza SEIR): test trajectories are generated under policy configurations outside the training distribution, with three sub-splits covering unseen advisory intensities (`advisory OOD`), unseen vaccination caps (`vaccination OOD`), and both combined (`combined OOD`).

Legacy two-panel description. For reference, Table 5 contrasts the COVID-19 surveillance panel with an Influenza-like Illness (ILI) dataset preserved from an earlier two-panel framing. The ILI data was used only as an observation-model sanity check and does not enter any reported result.

Targets. Each dataset reports three observables per state per week, all $\log(1+x)$-transformed (metrics reported in the same space). *COVID-19*: hospitalisations, cases, and deaths per 100k. *Influenza*: weighted ILI %, lab-confirmed flu hospitalisations per 100k, and flu mortality per 100k. Hospitalisations per 100k is the primary policy-relevant target optimised by the closed-loop split. Closed-loop policy rollouts run for 8 weeks per state.

Preprocessing. Counts are converted to per-100k rates using ACS state populations; negative weekly deltas are clipped to zero; targets and mobility flows are $\log(1 + x)$-transformed; static indicators are z-scored on the train split. Missing weeks (median 0.4% on COVID-19; 1.7% on Influenza) are linearly interpolated within state with a missingness-indicator channel. Daily GeoDS flows are aggregated to weekly OD counts and row-normalised per origin to give the mobility matrix $\mathcal{G}$.

B Baselines: Adaptations and Hyperparameters

B.1 World-Model Forecasting Baselines

Compartmental-GP (Qian et al., 2020). NumPyro re-implementation of the published methodology. The published model is a hierarchical two-layer Gaussian process with a compartmental SEIR prior mean function, conditioned on per-country policy indicators. With our 3-week warmup, the full hierarchical GP is over-parameterised, so we approximate its temporal smoothness with an AR(1) random walk on $\log R_t$ with mean reversion. Two variants: `cgp_naive` (AR(1) only, no policy term) and `cgp_policy` (AR(1) + global linear policy effect, $\log R_t = \text{AR}(1) + \beta \cdot \text{intensity}_t$).

MechBayes (Gibson et al., 2023). NumPyro re-implementation of the published Bayesian SEIR model. `mechbayes_naive`: log-β follows a per-state Gaussian random walk during

	COVID-19	Influenza-like Illness (ILI)
Used for	forecast + policy	observation model only
Period (ISO weeks)	2020-W01 to 2022-W52	2012-W30 to 2013-W40 (PAHO), 2010-W40 to 2015-W20 (CDC)
Panels / records	51 states $\times$ 156 weeks (7,956)	15 country panels $\times$ $\sim$130 weeks each
Cases / hosp / deaths	JHU CSSE / HHS Protect / JHU CSSE	ILINet wILI / n/a / n/a
Action grid	13-dim OxCGRT (Hale et al., 2021)	none
Special structure	per-week point estimates	per-week revision history (initial $\rightarrow$ final)
Mobility, statics	GeoDS (Kang et al., 2020), ACS, SVI (U.S. Census Bureau; Agency for Toxic Substances and Disease Registry, 2022)	Chakraborty et al. (2018) (PAHO, CDC); no static covariates

Table 5: Legacy two-panel dataset description preserved for reference. The ILI panel does not enter any reported result.

warmup, frozen post-warmup (closest to the published non-parametric-β formulation, policy-blind). `mechbayes_policy`: $\beta_t = \beta_0(1 - 0.7 \cdot \text{intensity}_t)$, matching the simulator's β-policy mechanism.

epidemia (Flaxman et al., 2020). NumPyro re-implementation of the published semi-mechanistic Bayesian renewal model. `epidemia_modeA`: per-replicate fit. α (intercept) and β (policy slope) sampled with priors and inferred from the 3-week warmup of each test replicate. `epidemia_modeB`: hierarchical transfer. α_{state} and β_{global} pre-fit on the training corpus via least-squares regression of weekly empirical log-growth on policy intensity; at test time only initial conditions and observation rates are sampled per replicate.

PAN-cODE (Shi et al., 2022). Latent ODE family: GRU-ODE encoder $\rightarrow$ reparametrisation $\rightarrow$ Neural ODE $\rightarrow$ autoregressive decoder. Per-week policy intensity aliased into PAN-cODE's stringency-index slot. Two variants: `uncond` (encoder/ODE/decoder all policy-blind) and `+ Policy` (per-week intensity concatenated to encoder input, latent z_0, and decoder input at every step). Hyperparameters: `latent_dim=32`, `n_units=50`, `n_layer=3`, `lr=1e-3`, batch 128, ELBO with 3 MC samples, 80 epochs.

EARTH (Wan et al., 2025). Neural CDE over interpolated regional observations with multi-view spatial coupling. The published model uses a static geographic adjacency and a DTW-derived semantic adjacency to inject inter-region transmission pressure into the CDE vector field. Lookback shortened from the published 8 weeks to our 3 weeks for compatibility with the EpiWorld harness. The policy-conditioned variant adds a 1-D policy encoder (inspired by EpiWorld's PolicyEncoder) whose output modulates the CDE vector field at each integration step.

B.2 Policy-Optimisation Baselines

EpiPolicy-RL (Mai et al., 2023). The published method uses PPO and SAC over the EpiPolicy compartmental simulator with a multi-objective reward combining health and intervention cost. We re-implement the PPO actor against the EpiWorld world model (the same dynamics our policy actor sees), using a Categorical actor with a 5-tier intensity codebook for COVID-19 SEIR, the 13-dim OxCGRT action for Real US COVID-19, and the 2-dim (advisory, vaccination tier) action for Influenza SEIR.

ADIOS (Wolpers et al., 2026). Grammar-guided genetic programming for intervention discovery. The published method evolves interventions as parse trees over a context-free grammar and evaluates them inside the GEMS agent-based simulator. We swap the simulator for the EpiWorld world model and adapt the grammar to each dataset's action space (13-D OxCGRT, 5-tier intensity scalar, or 2-tier advisory$\times$vaccination). The fitness function combines goal achievement with a stringency penalty. Population 50, 50 generations, tournament selection, 10% mutation rate.

EpidRLearn (Bampa et al., 2022). PPO (on-policy actor–critic) over a contact-based, age-structured C-SEIR simulator in the original paper, with a one-hidden-layer MLP (128 units, $\tanh$, separate policy and value heads) and discount γ=0.99. We re-implement the PPO actor against the EpiWorld world model for apples-to-apples fairness, keeping the same policy/value architecture and discount, and use the 5-tier intensity codebook and reward function shared with EpiPolicy-RL.

Training protocol (all policy baselines). 200K environment steps against the world-model rollout harness, Adam at 10^{-3}, discount $\gamma = 0.95$. Final policies are evaluated on the held-out splits described in Section 4.1 and Appendix A.

C Feature Specification

Static covariates. 17 per-state indicators ($\mathbf{s}_i \in \mathbb{R}^{17}$), identical across all datasets and z-scored on the train split (Table 6).

Feature	Description / source
Population	ACS
under_20	age-bracket fraction
over_65	age-bracket fraction
White	race/ethnicity fraction
Black	race/ethnicity fraction
medicaid_coverage	fraction
medicare_coverage	fraction
uninsured_percent	ACS
healthcare_utilization	composite index
poor_health_status	fraction
adults_at_high_risk	fraction
poverty_rate	ACS
social_vulnerability_index	CDC/ATSDR SVI
healthcare_access_quality	composite index
older_at_high_risk	fraction
dem_percent	political lean
rep_percent	political lean

Table 6: Static per-state covariates ($\mathbf{s}_i \in \mathbb{R}^{17}$).

Dynamic covariates. 14 weekly channels per dataset ($\mathbf{X}_i^{t-T:t} \in \mathbb{R}^{14\times T}$), $\log(1+x)$-transformed for the counts and clipped at zero (Table 7). Real US panels use observed channels only; synthetic datasets additionally expose the SEIR(V)-style latent compartments emitted by the simulator.

Feature	Type	COVID	Flu
cases_reported	per 100k	✓	✓
hosp_reported	per 100k	✓	✓
death_reported	per 100k	✓	
incidence	per 100k	✓	✓
Rt	Cori estimator	✓	✓
policy	intensity scalar	✓	✓
vax_curve	cumulative cov.		✓
previous_infection_12w	derived	✓	✓
Dose1_Pop_Pct	fraction	✓	
Series_Complete_Pop_Pct	fraction	✓	
Latent $\{S, E, I, A, R\}$ (simulator-only, +5 ch.)		✓	✓

Table 7: Dynamic per-state covariates per dataset ($\mathbf{X}_i^{t-T:t} \in \mathbb{R}^{14\times T}$). Each dataset sums to 14 channels: 9 observed + 5 latent.

D Action-Space Specification

The three datasets use different action interfaces, matched to what each world model is conditioned on (Table 8). Real US COVID-19 uses the published 13-D OxCGRT (Hale et al., 2021) indicator panel for compatibility with historical NPI records. COVID-19 SEIR collapses this to a 5-tier intensity scalar, aligning the agent's decision with the 1-channel stringency the simulator reacts to. Influenza SEIR splits a 2-D continuous interface (advisory $\times$ vaccination cap) into four tiers per channel. The OxCGRT panel groups into four envelopes used by the Protocol-Constraint check (Section 3.4): **personal protective**: H1; **community measures**: C3M, C4M, H2, H3, H6M; **closures and movement**: C1M, C2M, C6M, C7M, C8NV, C8V; **vulnerable populations**: H8M.

E Skill Catalogue

Each tier's skills are listed in Tables 9–11. Tier ordering mirrors the Method section (Section 3.4).

The lesson schema, tag vocabulary, and a sample of the current lesson bank appear in Appendix G.

F Key Agent Prompts

G Lesson Schema, Tag Vocabulary, and Examples

Schema. An Adaptive Lesson is a record with the following fields:

- `id`: unique identifier (e.g., `L-001`).
- `applicable_tags`: list of behavioural tags drawn from the vocabulary below.
- `target_skill`: identifier of the Protocol-Constraint or Surveillance Skill whose refinable parameter is adjusted.
- `parameter_delta`: dict of additive parameter shifts (clamped to per-skill bounds).
- `regime`: the policy setting under which the cohort was observed.
- `confidence` $\in [0, 1]$: aggregated evidence strength across source states, computed as in Appendix L.
- `sign_agreement`: fraction of supporting observations agreeing in sign with the cohort median.
- `source_states`: states contributing observations (≥ 2 required).

Dataset	Dimension	Levels	Description
Real US COVID-19: 13-D OxCGRT indicator panel, ordinal NPI levels (increasing stringency)			
	C1M_School closing	$\{0,1,2,3\}$	open / recommend / require some / require all
	C2M_Workplace closing	$\{0,1,2,3\}$	open / recommend / require some / require all
	C3M_Cancel public events	$\{0,1,2\}$	none / recommend / require
	C4M_Restrictions on gatherings	$\{0,1,2,3,4\}$	no limit $\rightarrow \leq$10 persons
	C6M_Stay at home requirements	$\{0,1,2,3\}$	none $\rightarrow$ minimal-exception lockdown
	C7M_Restrictions on internal movt.	$\{0,1,2\}$	none / recommend / restrict
	C8NV_Intl. travel (non-visitors)	$\{0,1,2,3,4\}$	open $\rightarrow$ total border closure
	C8V_Intl. travel (visitors)	$\{0,1,2,3,4\}$	open $\rightarrow$ total border closure
	H1_Public information campaigns	$\{0,1,2\}$	none / coordinated public info
	H2_Testing policy	$\{0,1,2,3\}$	none $\rightarrow$ open-public testing
	H3_Contact tracing	$\{0,1,2\}$	none / limited / comprehensive
	H6M_Facial Coverings	$\{0,1,2,3,4\}$	none $\rightarrow$ required in all public spaces
	H8M_Protection of elderly people	$\{0,1,2,3\}$	none $\rightarrow$ extensive support and isolation
COVID-19 SEIR: 5-tier intensity scalar (aligned with WM's 1-channel stringency input)			
	intensity_tier	5 tiers	NO_MEASURES $\rightarrow$ LIGHT $\rightarrow$ MODERATE $\rightarrow$ STRICT $\rightarrow$ FULL_LOCKDOWN
Influenza SEIR: 2-D advisory + vaccination, four tiers per channel			
	advisory_intensity	4 tiers	NONE (0.05), LIGHT (0.15), MODERATE (0.30), STRICT (0.50)
	vax_coverage_cap	4 tiers	LOW (0.20), MID (0.35), HIGH (0.45), AGGRESSIVE (0.55)

Table 8: Action-space specification across the three datasets. The agent selects a named tier; the world model receives the corresponding scalar intensity. COVID-19 SEIR tiers map uniformly to $\{0.00, 0.25, 0.50, 0.75, 1.00\}$; Influenza SEIR intensities are listed in-line.

Skill	Function	Anchoring authority
tier1_classify_pandemic_interval	Classifies the six WHO pandemic intervals (Investigation, Recognition, Initiation, Acceleration, Deceleration, Preparation) from (R_t, R_t CI, case trend, ICU, cluster detected); returns max_npi_tier_permitted, which gates downstream NPI choice.	WHO pandemic-interval framework
tier1_vaccine_allocation	Enforces the priority ordering 1A $\rightarrow$ 1B $\rightarrow$ 1C $\rightarrow$ 2 $\rightarrow$ 3; flags equity_score < 0.8 for human review; prohibits price-, insurance-, or market-based allocation.	ACIP (US) / WHO SAGE
tier1_check_npi_proportionality	Validates a proposed NPI bundle against the admissible tier envelope; requires legal authority for lockdown and movement restrictions; enforces a 14-day step-down minimum at $R_t < 0.9$.	ECDC proportionality test
tier1_who_outbreak_investigation	Enforces the WHO nine-step outbreak investigation ordering; blocks quantitative analysis before step 3 and control measures before step 6; mandates IHR notification at ≥ 5 confirmed cases.	WHO IHR (World Health Organization, 2008)

Table 9: Protocol Constraints ($\mathcal{T}_1$). Deterministic predicates returning $\{0, 1\}$ admissibility; computed once per decision step.

- n_observations: number of contributing rollout observations.
- pattern_observed: free-text description of the divergence prompting the lesson.
- recommendation: free-text statement of the action-level implication, used only for human inspection.
- retires, superseded_by: supersession metadata used to invalidate prior lessons whose sign conflicts with a newer one.

Tag vocabulary. Tag sets $\tau(\mathbf{X}^t, \mathbf{S})$ are drawn from the following namespaces:

Example lessons. Two regime-conditioned lessons from the current bank (one COVID, one influenza), each combining a cohort tag set with the active policy regime to produce an action-level recommendation:

H Experimental Details

World-model instantiation. The world model in Section 3.3 is realised as a dual-encoder Graph

Skill	Method (source)	Refinable parameters
`tier2_estimate_rt`	Bayesian renewal-equation estimator for R_t (Cori et al., 2013).	serial-interval mean (4.7 d), s.d. (2.9 d), window τ (1 w)
`tier2_verity_expected_mortality`	Age-stratified IFR with case-ascertainment correction.	ascertainment ratio (0.3)
`tier2_compute_stringency_index`	Nine-indicator OxCGRT normalisation to $[0, 100]$ (Hale et al., 2021).	—
`tier2_lookup_svi`	ATSDR Social Vulnerability Index → quartile and equity directives (Agency for Toxic Substances and Disease Registry, 2022).	—
`tier2_compute_growth_rate`	Log-linear OLS for exponential growth rate r, doubling/halving time, and fit R^2.	stable band ($\pm 5\%$/week)
`tier2_r_from_growth_rate`	Wallinga–Lipsitch closed-form R from growth rate r and generation interval.	—
`tier2_classify_psaf_tier`	CDC Pandemic Severity Assessment Framework → severity tier 1–5.	hospitalisation severity thresholds, R_t transmissibility bands

Table 10: Surveillance Skills ($\mathcal{T}_2$). Analytical estimators returning quantitative summaries; refinable parameters may be adjusted by Adaptive Lessons within per-skill bounds.

Skill	Signature	Description
`tier3_retrieve_lessons`	(τ, k)	Returns the top-k lessons ranked by tag overlap with the current context tags τ, weighted by lesson confidence ρ_ℓ.
`tier3_apply_lesson`	(ℓ, ϕ)	Merges the parameter delta $\boldsymbol{\delta}_\ell$ into base parameters ϕ, clamping to per-skill bounds B_κ; only keys in the refinable-parameter set are modulated.

Table 11: Adaptive Lessons ($\mathcal{T}_3$). Retrieve-and-apply operators over the lesson bank; structure is fixed, content evolves through after-action analysis.

Transformer. The recency branch is a GRU over a 3-week window; the delayed-effect branch is a Neural-CDE over the same 3-week window with a causal cross-channel mixer (DELAYMIX). A learnable scalar gate balances the two branches and their outputs are concatenated. Policy actions enter through feature-wise (FiLM-style) modulation of the latent state. Spatial coupling is a 4-layer adjacency-masked Graph Transformer. A residual head predicts the additive update to each observed channel (cases, hospitalisations, deaths per 100k for COVID-19; cases and hospitalisations per 100k for influenza). Training combines the latent-prediction and observation-reconstruction losses defined in Equations 9–10.

World-model training. We train with AdamW (learning rate $3{\times}10^{-4}$, weight decay 10^{-3}, 10% linear warmup, gradient clipping at 1.0), batch size 16, and 5 000 steps. Targets are $\log(1+x)$-transformed and standardised on the train split. Each model is trained on a single NVIDIA A100 (Perlmutter). The best checkpoint by in-distribution validation loss is used for all downstream evaluations.

Policy actor. The actor calls a chat-completion API for each backbone in Table 4 (GPT-4o, GPT-5-mini, GPT-5, Qwen2.5-7B-Instruct, DeepSeek-R1, Llama-3.3-70B-Instruct) at temperature 0.2. Outputs are parsed from structured JSON with a regex fallback to the `MODERATE` tier on parse failure. The action grid is chosen per dataset as specified in Table 8. Adaptive-Lesson retrieval injects the top-5 lessons by tag overlap (Section 3.4).

Closed-loop adaptation. Closed-loop rollouts run for an 8-week horizon. Skill adaptation runs after each rollout batch: the analyser proposes lessons from the batch, and the bank is updated through additive insertion with opposite-sign supersession (Equation 15). Protocol Constraints and Surveillance Skill structures are not modified during adaptation.

Evaluation. World-model accuracy is reported as Peak-MAE (absolute error at the hospitalisation peak per state, per 100k) and mean trajectory MAE on the OOD splits in Section 4.1. Policy performance is reported as cumulative hospitalisation per 100k over an 8-week post-intervention horizon, averaged over the held-out states. All policy methods

Protocol-Constraints system prompt (excerpt)

Role. You are an epidemiology policy agent operating under the Protocol Constraints derived from WHO, CDC, and ECDC codified guidelines. Protocol-Constraint outputs are non-overridable and anchor every downstream decision.

Ordering rules (no exceptions).

1. `tier1_classify_pandemic_interval` MUST be the first tool called in every session, before any other tool, and re-invoked whenever the epidemic state changes. Subsequent reasoning is gated by its outputs (`current_interval` , `max_npi_tier_permitted` , `Rt_status`).
2. `tier1_who_outbreak_investigation` steps 1–3 must complete before any quantitative analysis.

Adaptive-Lessons after-action analyser prompt (excerpt)

Role. You are the Adaptive-Lessons analyser for an epidemiology policy agent. Read a batch of rollout summaries plus the current lessons bank, and propose new or refined lessons describing cross-state behavioural patterns. Each lesson encodes a `parameter_delta` the agent will apply to a specific Protocol-Constraint or Surveillance-Skill refinable parameter.

Hard rules.

1. Additive memory. You do not overwrite the bank. Same-sign refinements update an existing lesson in place; opposite-sign refinements require `retires: [L-XXX]` referencing the old lesson id, with a justification in `pattern_observed` .

Reflection / prompt-evolution operator (excerpt)

Role. You are a prompt engineer optimising a system prompt for an epidemic policy advisor. The advisor recommends a 13-dimensional NPI policy per region per week. The recommendations are evaluated by a world model that predicts hospitalisation trends. The goal is to minimise hospitalisations per 100k.

Task.

1. Diagnose specific weaknesses in the current system prompt that may have led to suboptimal recommendations.
2. Propose an improved system prompt addressing those weaknesses, under 250 words.

Tag namespaces

```
compliance:      low, mid, high
rebound:         fast, slow, none
capacity:        low, mid, high
vax-uptake:      slow, mid, fast
vax-ceiling:     low, mid, high
mandate-latency: slow, mid, fast
svi:             low, moderate_low, moderate_high, high
spread:          hub, rural, border, mixed
political:       dem, rep, mixed
psaf:            tier1, tier2, tier3, tier4, tier5
```

are evaluated against the same world model and the same initial conditions to isolate the effect of the actor.

I Additional Experiments

Hospitalisation reduction (mean per-100k). Figure 3 reports the §5.2 ablation in the mean per-100k hospitalisation metric. Static skills shift mean hospitalisation by at most 2% on any dataset, while the after-action Skill Evolution cycle drives mean reduction to between 21% and 31%. The pattern is preserved on Influenza SEIR despite its different NPI grid (CDC PMGR rather than OxCGRT), indicating the gain is structural to the closed-loop skill cycle.

World-Model Design Ablation. Table 12 reports a per-component ablation of the EpiWorld world-model architecture on COVID-19 SEIR (Peak-

Example lessons (L-A: COVID, L-B: influenza)

```
{
  "id": "L-A",                                        // COVID-19 SEIR
  "applicable_tags":
      ["access:high", "lean:dem", "pop:large", "svi:high"],
  "regime":  "stringency = MODERATE",
  "target_skill":  "tier2_classify_psaf_tier",
  "confidence":  0.46,
  "source_states":  ["California", "New York"],
  "n_observations":  21,
  "sign_agreement":  0.86,
  "pattern_observed":
    "SVI:high dense urban cohorts respond to MODERATE
     stringency as a lower-SVI cohort would respond to
     LIGHT -- transmission keeps growing.",
  "recommendation":
    "Same lever, one tier up: stringency MODERATE -> STRICT."
}

{
  "id": "L-B",                                        // Influenza SEIR
  "applicable_tags":
      ["access:high", "lean:dem", "pop:large", "svi:low"],
  "regime":  "advisory = ADV_MODERATE,  vax = VAX_HIGH",
  "target_skill":  "tier2_classify_psaf_tier",
  "confidence":  0.46,
  "source_states":  ["Illinois", "Pennsylvania"],
  "n_observations":  11,
  "sign_agreement":  1.00,
  "pattern_observed":
    "Vaccination is already saturated (VAX_HIGH) but a
     moderate advisory still lets transmission climb --
     the vaccination lever cannot substitute for contact
     reduction at this density.",
  "recommendation":
    "Different lever: hold vax = VAX_HIGH, escalate
     advisory ADV_MODERATE -> ADV_STRICT."
}
```

MAE, log1p per-100k). Dropping any single component degrades both in-distribution and OOD Peak-MAE; the residual prediction head is load-bearing, and replacing it with absolute next-state prediction diverges. The largest single contributor to OOD generalisation is the multi-stream SIR-coupled head; the dual encoder (GRU + Neural-CDE) and the long lookback contribute next.

Variant	In-Dist. Peak-MAE	OOD Peak-MAE
EpiWorld (full)	**0.339**	**0.414**
– dual encoder (CDE-only)	0.385	0.433
– dual encoder (GRU-only, 8w)	0.419	0.540
– long lookback (GRU 3w)	0.587	0.720
– DelayMixer (1-D causal mixer)	0.399	0.426
– multi-stream SIR-coupled head	0.554	0.734
– policy conditioning	0.421	0.585
– residual head (absolute)	diverges (≥ 8)	

Table 12: World-model component ablation on COVID-19 SEIR (Peak-MAE, log1p per-100k; lower better). Each row removes one component from the full model.

Influenza Component Ablation. To test whether the architectural choices are specific to COVID-19, we repeat the component ablation on the influenza corpus (two seeds per configuration, influenza state-OOD). The dual temporal encoder is load-bearing on influenza exactly as on COVID-19: removing the Neural-CDE branch raises overall MAE from 0.487 to 0.604. The DelayMixer gives the best influenza Peak-MAE (0.441 against 0.463 without it). The design is therefore not COVID-specific.

Scored head-to-head under an identical protocol, EpiWorld is lower than EARTH + Policy on influenza overall rollout MAE (0.541 against 0.581) and on the hospitalisation channel (0.491 against 0.525), with peak error competitive. At the policy level, the independent-simulator provenance gate passes (maximum absolute error 3.4×10^{-5}) and Skill Evolution beats the Base LLM off-model on

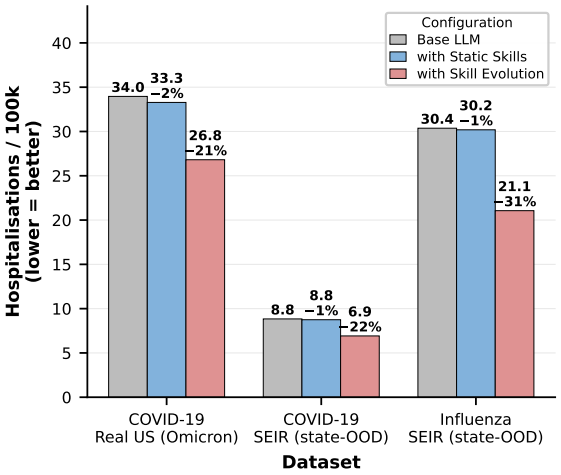

Figure 3: Hospitalisation reduction (mean per-100k) across three datasets and three skill configurations.

the final-week metric (-10.6%). The cumulative influenza margin is smaller than on COVID-19, consistent with the simpler 2-dimensional action space leaving less room for timing and targeting to matter.

Cross-Domain Negative Lesson Transfer. Lessons mined under one world model do not necessarily port to a different world model. We take the lesson bank `lessons_top20.json` mined under the real-US MECHANISMGRAPHWMREAL world model and apply it to the synthetic SCENARIOGRAPHWMV15_DUAL on five held-out OOD states. Pooled across states and replicates, the transferred lessons *degrade* performance by $+12.1\%$ (Table 13).

Held-out state	Base	+ Real-US lessons	Δ
KS	12.51	12.64	$+1.0\%$
MN	6.87	6.78	-1.4%
NV	57.74	74.79	$+29.5\%$
SC	11.12	10.29	-7.5%
WA	88.24	93.28	$+5.7\%$
Pooled	**35.30**	**39.56**	$+12.1\%$

Table 13: Cross-domain lesson transfer is negative. Lessons mined under the real-US world model are applied to the synthetic SEIR world model on five held-out OOD states. Mean hospitalisation per 100k; lower better.

The mechanism is bias-direction-specific. The real-US WM *over-predicts* Omicron hospitalisations, so L-002's shrinkage ($\delta = -0.234$) corrects the bias and helps. The synthetic v15_dual WM *under-predicts* on state-OOD, so the same shrinkage *amplifies* the under-prediction. Tier-3 lessons are therefore not universally portable: they must be mined under the target world model's own training distribution.

Per-State Breakdown. Table 14 reports per-state hospitalisation MAE for the 20 OOD-evaluated US states on the real-data Omicron split, complementary to the aggregate hospitalisation metric in Table 2. States are listed in decreasing order of baseline MAE; bold marks the configuration with the lowest MAE per state.

State	A (Base)	D (+ Skills)	F (+ Skill Evo.)
Arizona	23.34	25.03	**14.24**
Indiana	20.01	18.84	**10.86**
Wisconsin	19.16	17.06	**10.44**
Pennsylvania	15.49	15.62	**9.27**
Tennessee	15.24	16.18	**10.03**
Michigan	14.07	14.63	**6.28**
Ohio	13.25	13.14	**8.66**
Massachusetts	12.66	12.32	**9.19**
Missouri	11.75	11.77	**8.67**
Georgia	11.27	10.48	**9.03**
Texas	11.13	11.43	**8.92**
North Carolina	9.69	9.64	**7.89**
New York	8.94	9.35	**6.53**
Maryland	7.65	7.35	**4.86**
Virginia	7.43	8.27	**5.50**
Florida	7.08	**6.38**	9.32
New Jersey	6.43	**6.21**	8.18
California	5.70	**4.42**	8.44
Illinois	5.45	**5.20**	7.41
Washington	**3.24**	3.27	4.27

Table 14: Per-state hospitalisation MAE (per 100k) on the real-US Omicron OOD split. Skill Evolution improves 15 of 20 states and gives its largest reductions on high-baseline-MAE Midwest states (AZ, IN, WI). On the five lowest-baseline-MAE states (FL, NJ, CA, IL, WA) the multiplicative correction over-applies and MAE regresses; this regime-dependence is discussed in Appendix I.

J Policy-Alignment Diagnostics

Policy alignment measures how far proposed interventions stay within the distribution of actions taken in practice: the activation rate is the fraction of decisions on which the proposed and historical action agree on which measures are switched on, and the exact-match rate the fraction on which they agree on the full action bundle. The score is a plausibility diagnostic, not a measure of policy quality: the goal of the framework is to improve on the reference historical policy, so agreement with it is neither necessary nor sufficient for a good out-

come. Table 15 reports both rates for the methods in Table 2.

Method	*Real US* Act./Ex.	*COVID* Act./Ex.	*Flu* Act./Ex.
EpiPolicy-RL	43.1/22.3	92.3/10.4	38.1/17.2
ADIOS (GGGP)	31.2/33.4	90.1/15.5	27.6/25.8
EpidRLearn	31.9/33.7	90.4/15.8	28.2/26.0
EpiWorld (Base LLM)	66.1/50.3	97.2/23.3	58.5/38.8
EpiWorld (Static Skills)	69.2/53.8	97.7/25.0	61.2/41.5
EpiWorld (Skill Evol.)	74.7/57.3	98.9/26.7	66.1/44.2

Table 15: Policy-alignment diagnostics (activation / exact-match rate, %) for the methods in Table 2. Reported as a plausibility check on the proposed action distribution, not as a quality metric.

On COVID-19 SEIR every method scores between 90.1% and 98.9% activation while cumulative hospitalisation reduction ranges from 8.2% to 22.0%, so alignment does not discriminate between policies whose outcomes differ by a factor of 2.7, and exact-match never exceeds 60% in any cell.

K No-LLM Rule-Based Selector

To separate the contribution of LLM reasoning from that of the engineered skill library, we replace the policy actor with a fixed severity-proportional rule that selects from the same library: the Tier-2 R_t and severity estimators set the intervention level, and the Tier-1 constraints cap it exactly as they do for the LLM actor. No language model is involved at any point in the decision.

Selector	Cumulative reduction	Mean intensity
Rule-based (no LLM)	-21.4%	0.485
Base LLM	-36.5%	≈ 0.50
+ Skill Evolution	-39.3%	≈ 0.50

Table 16: No-LLM ablation, COVID-19 SEIR state-OOD, evaluated on the independent mechanistic simulator. All three selectors are scored on an identical set of rollouts, so the values are directly comparable within this table; they are computed on a matched subset and are not on the same scale as Table 3.

Removing the LLM roughly halves the reduction while the mean intervention intensity is essentially unchanged (0.485 against ≈ 0.50), so the LLM does not obtain its advantage by proposing more stringent measures. What it contributes is when and where a comparable intervention budget is spent. The Base LLM alone, without lessons or the after-action cycle, accounts for most of the gap; Skill Evolution adds a further 2.8 points.

L Teacher Module and Lesson Retrieval

This section specifies the after-action analyser of §3.5 and the retrieval operator of Eq. 14. The analyser learns which intervention levels produce which outcomes for which kinds of state, and stores that as adjustments to Tier-2 Surveillance-Skill parameters; Tier-1 Protocol Constraints are never modified.

Selection. Each closed-loop episode records how a state's hospitalisation outcome responded to the stringency the actor applied. Episodes are pooled into cohorts keyed by the state's profile (population, social vulnerability, political lean, healthcare access), the target skill, and the policy regime. A cohort yields a lesson only if it draws on at least two distinct states, which prevents single-state idiosyncrasies from entering the bank.

Confidence weighting. Each cohort emits a delta $\boldsymbol{\delta}_\ell$ on one refinable Tier-2 parameter, capturing how that cohort's response differs from the skill's default. Its confidence is

$$\rho_\ell = \underbrace{\left(1 - \frac{1}{1+n/5}\right)}_{\text{evidence}} \cdot \underbrace{a_\ell}_{\text{sign agreement}} \cdot \underbrace{\left(1 - \frac{1}{1+n_{\text{states}}}\right)}_{\text{breadth}}, \tag{16}$$

where n is the number of supporting observations, a_ℓ the fraction of those observations agreeing in sign with the cohort median, and n_{states} the number of contributing states. All three factors lie in $[0, 1]$, so $\rho_\ell \in [0, 1]$.

Merge rule. When a new observation batch refines an existing lesson, confidence is blended by evidence weight,

$$\rho \leftarrow \frac{\rho_{\text{old}} n_{\text{old}} + \rho_{\text{new}} n_{\text{new}}}{n_{\text{old}} + n_{\text{new}}}, \tag{17}$$

so ρ_ℓ is a running evidence-weighted mean rather than a value overwritten by the most recent episode.

Contradictions. Same-sign refinements merge in place via Eq. 17. An opposite-sign refinement is rejected unless it explicitly retires the prior lesson, which is then marked superseded and retained rather than deleted. The bank therefore cannot oscillate between opposite corrections, and the superseded record preserves the audit trail.

Retrieval ranking. At each decision the actor extracts the state's context tags and ranks candidate lessons by $n_{\text{overlap}} \times \rho_\ell$, with tags unweighted. The deltas of the top five lessons with at least two overlapping tags are added to the Tier-2 parameters and clipped to $[-0.8, 0.8]$.

Convergence. Every factor in Eq. 16 is bounded and saturating in its argument, and Eq. 17 is a convex combination, so ρ_ℓ converges as observations accumulate rather than drifting with episode order. Combined with the $\geq$ 2-state selection rule and the supersession policy, the bank is stable under continued adaptation.