Where Mechanism Enters: A Survey of Mechanism-Coupled Foundation Models for Epidemic Intelligence

Anonymous ACL submission

Abstract

As foundation models transition from standalone research tools to components of compound systems, understanding how they couple with structured domain engines becomes a pressing research mission that existing component-level taxonomies do not address: at what point in the model lifecycle should such an engine enter? This question, which we term *mechanism coupling*, determines what evaluation can verify and what deployments become tractable. We study mechanism coupling through epidemic intelligence, a domain where modeling decisions carry direct consequences for public-health response and whose mechanistic engines are formalized. Four coupling patterns emerge across two subfamilies: forecasting systems, where mechanism enters (i) in the training data or (ii) in the architecture; and simulation systems, where mechanism enters (iii) behind a verifier or (iv) in the runtime environment of an embedded agent. Through this case study, we make three contributions. First, we introduce a lifecycle taxonomy organized by when a structured engine enters the system. Second, we analyze what each coupling position makes verifiable and what it leaves opaque. Third, we connect deployment priorities to coupling choices. Together, these contributions frame mechanism-coupled systems not only as predictive tools, but also as instruments for mechanistic discovery in support of accountable public-health decisions.

1 Introduction

Epidemic intelligence (EI) is the systematic early identification, verification, and assessment of public-health hazards (Paquet et al., 2006). The COVID-19 pandemic underscored how signal fragmentation constrains real-time response, intensifying demand for methods that integrate heterogeneous data streams (Brownstein et al., 2023; Morgan et al., 2021).

EI draws on several modeling traditions. Compartmental models in the Kermack–McKendrick lineage (Kermack and McKendrick, 1927) gave the field its first quantitative backbone, formalizing transmission dynamics through differential equations; for instance, $\dot{S} = -\beta SI/N$ dictates that susceptible individuals leave the compartment at a rate proportional to their encounters with the infected fraction, scaled by a transmission rate β carrying direct epidemiological meaning. Agent-based simulators extended this framework to individual-level contact networks (Eubank et al., 2004; Ferguson et al., 2020); machine learning opened a data-driven branch for forecasting from heterogeneous signal (Brownstein et al., 2009). Yet public-health decisions, whether to restrict movement, allocate vaccines, or close schools, require reasoning traceable to factors that domain experts can inspect and challenge (Lau et al., 2026); purely data-driven models cannot provide this traceability. Foundation models, from large language models to domain-specific forecasters sharing their architectural paradigm, have recently joined this lineage in structurally different ways: some pretrain on simulator output and deploy without it (Dudley et al., 2025b), some embed mechanism into the model architecture (Liu et al., 2026), some emit symbolic representations checked by external verifiers (Datta et al., 2026), and some operate as agents inside running simulators (Chopra et al., 2025).

These systems share a structural decision that shapes what evaluation can verify, which failure modes appear, and what deployments become tractable: where in the model lifecycle a structured domain engine (a mechanistic simulator, verifier, or formal constraint system) enters. The same decision recurs wherever foundation models meet domain engines: execution-trace pretraining in code generation (Meta FAIR, 2025), formal verifiers in theorem proving (Hubert et al., 2026), and runtime simulators in social science (Piao et al., 2026).

Prior surveys organize the epidemic-intelligence literature by task, method, or AI role (Kraemer et al., 2025; Ye et al., 2025; Jiang et al., 2026; Lau et al., 2026); we organize it along this structural decision, which we term **mechanism coupling**, the point at which a structured engine enters the foundation-model lifecycle. We argue that recent advances in mechanistic interpretability open a bidirectional paradigm for mechanism-coupled epidemic intelligence: extracting the mechanistic concepts that coupled models implicitly encode, then using those discoveries to refine the coupling itself (§6.1). Such a cycle would let domain experts trace, challenge, and revise the factors driving a forecast, strengthening accountability for public-health decisions.

Three questions organize this survey. Where in the model lifecycle does the engine enter? What does each entry point make verifiable, and what does it leave opaque? What design principles follow from these distinctions? By answering them, we contribute (i) a lifecycle taxonomy organized by when mechanism enters, (ii) an analysis of what each coupling position makes verifiable and what it leaves opaque, and (iii) a bidirectional interpretability paradigm positioning mechanism-coupled systems as instruments for mechanistic discovery. §2 surfaces the epidemic-intelligence challenges that motivate mechanism-coupled capabilities; §3 develops the taxonomy; §4 traces the training and inference paradigms that realize each coupling pattern; §5 audits evaluation practice against it; §6 develops future directions and identifies open challenges.

2 Mechanism-Coupled Capabilities for Epidemic Intelligence

Epidemic-intelligence tasks share structural difficulties that classical pipelines and standalone language models struggle with: heterogeneous data integration across multilingual outbreak text, surveillance time series, mobility tensors, and genomic streams (Brownstein et al., 2023; Kraemer et al., 2025); underlying mechanistic dynamics that surface pattern-matching alone obscures; cross-disease generalization under sparse historical data (Lau et al., 2026); calibrated counterfactual reasoning under intervention (Ferguson et al., 2020); and real-time policy-relevant reasoning under delayed or partial surveillance (Morgan et al., 2021).

These challenges share a structural property: each requires a foundation model that does not merely extrapolate from historical signal, but reasons over a mechanism. In epidemic intelligence, this requirement is not merely technical: decisions affect population-level outcomes, and the reasoning behind a forecast must be traceable to mechanistic factors that epidemiologists can audit (Lau et al., 2026). A new paradigm has emerged in which foundation models (general-purpose language models and domain-specific forecasters alike) are coupled to mechanistic substrates, which we define as structured components encoding domain-specific epidemiological constraints, from compartmental equations and contact networks to agent-based simulators, rather than replacing them. The integrated systems we survey bring five capabilities to bear on these challenges. *Text understanding* and *multimodal alignment* encode heterogeneous epidemic context, from outbreak reports and policy text to case-count graphs and genomic streams, into representations that enable joint reasoning over signals classical pipelines treat as independent channels (Du et al., 2025; Gong et al., 2025; Dudley et al., 2025b; Liu et al., 2026). *Program synthesis* with verified intermediate representations shifts the epidemiologist's task from writing simulation code to specifying scenarios in natural language, with a typed representation enforcing structural admissibility before execution (Datta et al., 2026; Samaei et al., 2025). *Behavioral simulation* embeds the LLM as a decision function inside a stepped mechanistic simulator, with persona and memory conditioning replacing hand-coded behavioral rules (Williams et al., 2023; Chopra et al., 2025; Hou et al., 2025; Wang et al., 2026; Bo et al., 2026); at population scale, this produces emergent dynamics that static decision rules cannot express, over which *policy reasoning* can ground counterfactual intervention analysis in mechanistic state rather than plausibility heuristics, where controlled experimentation is infeasible (Aoki and Ghaffarzadegan, 2026; Shi et al., 2026).

These five capabilities plug into the model lifecycle at four distinct positions, realized as the coupling patterns developed in §3 and illustrated in Figure 1.

3 Coupling Patterns in Forecasting and Simulation

Mechanism coupling takes four forms: training-time, architectural, verification-time, and runtime. We trace each through its representative systems;

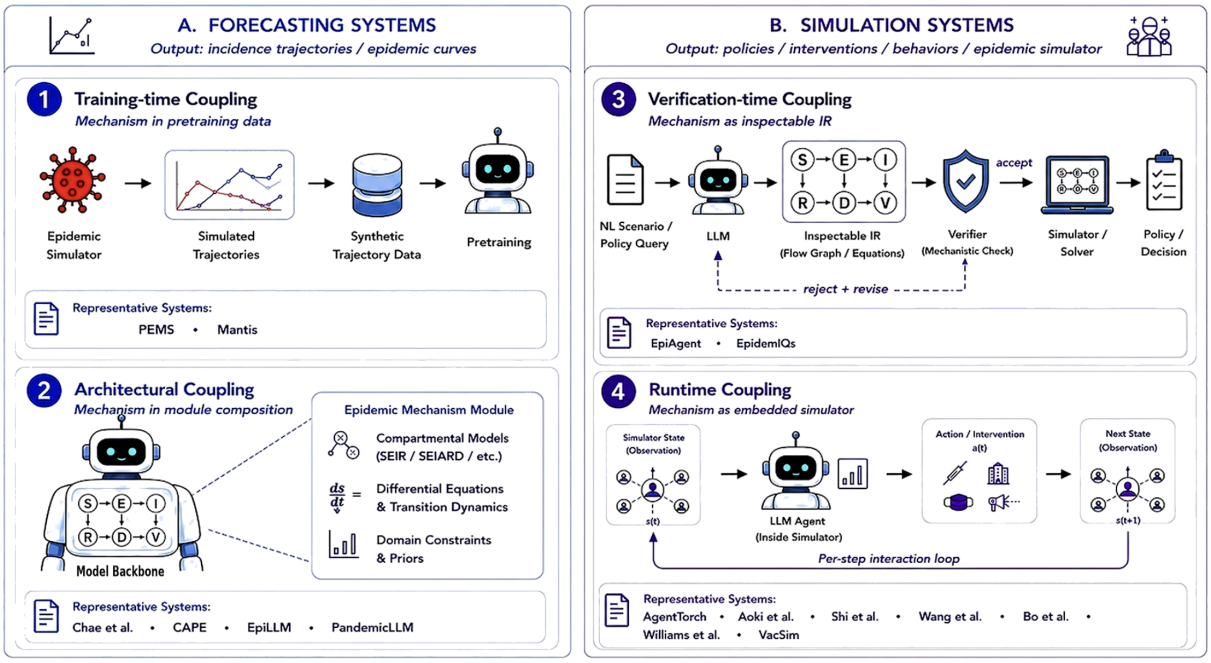

Figure 1: The four coupling patterns and their representative systems. Forecasting systems (left) bind mechanism before deployment: training-time coupling consumes simulator-generated trajectories during pretraining, while architectural coupling embeds mechanistic modules into the deployed model. Simulation systems (right) keep mechanism active during use: verification-time coupling checks LLM-generated intermediate representations against structural predicates, while runtime coupling embeds the LLM as an agent inside a stepped simulator. Each pattern is traced in detail in §3.1–§3.2.

learning paradigms are developed in §4. Figure 1 illustrates the workflow details of the two subfamilies by coupling patterns.

3.1 Forecasting

Training-time coupling Training-time coupling instantiates mechanism coupling at the level of the pretraining distribution: a parameterised collection of simulators $\mathrm{Sim}(\theta)$, indexed by configuration θ drawn from a prior $p(\Theta)$, generates plausible epidemic trajectories, and a model M is fitted to minimize an expected loss over that collection,

$$\mathcal{L}_{\mathrm{TT}} \;=\; E_{\theta\sim p(\Theta)}\, E_{\tau\sim \mathrm{Sim}(\theta)}\big[\, \ell(\hat{y}_\tau,\, y_\tau)\,\big], \qquad (1)$$

where $\hat{y}_\tau$ is the model's prediction for trajectory τ and ℓ is a proper scoring loss. The marginalization in (1) is the structural commitment: the simulator is an oracle during training and absent at deployment. Two systems instantiate this pattern. PEMS pretrains self-supervisedly on the Project Tycho multi-disease archive, with mechanism entering through multi-task signal rather than an explicit simulator collection (Kamarthi and Prakash, 2023; van Panhuis et al., 2013). Mantis pushes the marginalization onto explicit simulator parameters, pretraining exclusively on synthetic trajectories spanning compartmental, host-vector, and environmental simulators (Dudley et al., 2025b).

Architectural coupling Architectural coupling persists mechanism as a permanent component of the deployed model's parametric form, with four systems spanning a bandwidth spectrum from scalar conditioning to full multimodal reasoning. At the narrow end, an event-interpreter LLM compresses external context into a single transmission-impact signal feeding a stochastic forecasting core (Chae et al., 2025):

$$\begin{aligned} I_t &= \mathrm{LLM}(\mathrm{context}_t), \\ y_t \,|\, I_t, x_{1:t} &\sim \mathrm{NegBin}(\mu_\theta(x_{1:t}, I_t),\, \phi)\,, \end{aligned} \qquad (2)$$

where I_t is the LLM's context-compression output, $x_{1:t}$ the historical case-count series, NegBin the negative-binomial likelihood commonly used for overdispersed count data, and ϕ a learned dispersion parameter. The language model's role is restricted to context compression: I_t modulates the forecaster's conditional mean but does not alter the distributional form. CAPE widens the interface to a K-dimensional dictionary of learned compartmental prototype vectors, representing each outbreak as a soft mixture over prototypes (Liu et al., 2026): given the outbreak's latent representation h_t at time

t,

$$\hat{h}_t = \sum_{k=1}^{K} \alpha_{t,k}\, c_k, \quad \alpha_t = \mathrm{softmax}(W h_t) \tag{3}$$

with prototype-alignment regularisers driving the c_k toward functional analogues of compartmental states. EpiLLM pushes the interface into the LLM's full embedding space: separate encoders $z_t^{\text{epi}} = f_\theta^{\text{epi}}(X_t)$ and $z_t^{\text{mob}} = f_\phi^{\text{mob}}(A_t, X_t)$ tokenize case-count graphs and mobility tensors respectively, and a gated fusion aligns both into a frozen LLM backbone (Gong et al., 2025). At the widest end, PandemicLLM reformulates forecasting as multimodal text reasoning: a fine-tuned Llama-2-13B treats spatial features, genomic variants, and policy stringency as text prompt segments alongside a recurrent bridge that encodes case-count time series into token embeddings (Du et al., 2025).

Across the forecasting subfamily, the progression from training-time to architectural coupling shifts the traceability boundary (what remains mechanistically auditable) : mechanism moves from fully absorbed in weights to partially exposed through prototype alignments or projector signals.

3.2 Simulation

Verification-time coupling Verification-time coupling instantiates mechanism coupling as symbolic synthesis under structural constraints: an LLM proposes a typed intermediate representation (IR) — a flow graph, a DSL program, a simulator configuration — which a verifier checks before execution, with rejection routing back to LLM revision within the same task instance. For a compartmental flow graph $G = (V, E)$ with compartments as nodes V and transitions as edges E, admissibility requires three properties: population mass must be conserved across all flows in G, every edge $e \in E$ must use an admissible functional form, and no compartment in V may be unreachable from the initial state. Two systems carry this pattern at different verification strengths. EpiAgent makes these three properties the core of the loop: an LLM generates a flow graph G from a natural-language scenario, the verifier checks the properties before compilation, and rejection produces a structured error trace referencing the failing property that drives revision (Datta et al., 2026). EpidemIQs adopts a less stringent verification mode in which a scientist agent coordinates task-expert agents across a five-phase workflow, with verification operating at phase-acceptance level rather than at symbolic-IR level (Samaei et al., 2025).

Runtime coupling Runtime coupling instantiates mechanism coupling as continuous co-residence: the language model is embedded inside a mechanistic simulator that interprets the LLM's outputs as agent-level state transitions or behavioral parameters at every step, with neither model's weights updated. The internal axis is the state-ownership spectrum between LLM and simulator. At the narrow end the LLM emits a scalar or categorical action consumed by a simulator that owns all state,

$$\begin{aligned} a_i(t) &\sim p_{\text{LLM}}\big(\cdot \,\big|\, \pi_i, \mathrm{context}(t)\big)\,, \\ s_i(t{+}1) &= f_{\text{sim}}(s_i(t), \{m_{ij}\}, e(t), a_i(t))\,, \end{aligned} \tag{4}$$

where π_i is agent i's persona, $\mathrm{context}(t)$ the epidemic context (local case counts and policy state), $\{m_{ij}\}$ the contact network, and $e(t)$ the environment. At the wide end the LLM owns text-valued state which it itself updates from stimuli generated by a minimal substrate,

$$(M_i(t{+}1),\, \sigma_i(t{+}1)) = \mathrm{LLM}\big(\pi_i,\ M_i(t),\ \mathrm{stim}_i(t),\ \boldsymbol{\sigma}_{\mathcal{N}_i}(t)\big), \tag{5}$$

where $\boldsymbol{\sigma}_{\mathcal{N}_i}(t)$ collects the observable states of agent i's neighbours in the contact network. The shift from Eq. (4) to Eq. (5) tracks a transfer of state ownership (which component maintains epidemic-relevant variables) from the simulator to the LLM: narrow-end systems emit a single action into a state-owning simulator; wide-end systems maintain their own memory and stance while the substrate reduces to a stimulus generator.

At the narrow end, AgentTorch queries the LLM per behavioral archetype rather than per agent (Chopra et al., 2025), while Aoki emits a single scalar restriction level into an SEIR ODE (Aoki and Ghaffarzadegan, 2026). Shi, Wang, and Bo populate intermediate positions (Shi et al., 2026; Wang et al., 2026; Bo et al., 2026). Toward the wide end, Williams runs daily per-agent reasoning (Williams et al., 2023), and VacSim reaches the wide endpoint with census-grounded personas (Hou et al., 2025).

Across the simulation subfamily, the traceability boundary shifts in the opposite direction: verification-time coupling produces fully auditable intermediate representations, while runtime coupling yields behavioral traces whose faithfulness to mechanistic constraints is difficult to verify at population scale.

The four sub-couplings admit a deeper analytical reading: where mechanism enters the lifecycle determines who owns epidemic-relevant state at inference, and state ownership in turn determines what evaluation can verify and what remains opaque, and therefore which aspects of a forecast public-health decision-makers can trace back to mechanistic factors. Table 1 records this alignment: as coupling moves later in the lifecycle, inspectability increases but the nature of opacity shifts from representational (whether weights encode mechanism) to behavioral (whether agents respect mechanism under shift). This asymmetry carries direct implications for evaluation (§5).

4 Mechanism-Shaped Training and Inference

The coupling patterns of §3 commit to positions on two axes: how much the model trains against epidemic data, and how much epidemic state it accesses at inference.

4.1 Pretraining

The training-time systems of §3.1 differ in how they structure the pretraining objective. Kamarthi and Prakash (2023) encode epidemic structure as auxiliary self-supervised tasks: peak-masking forces reconstruction of surges from surrounding context, and a four-way season classification embeds seasonal regularity, with a two-stage transfer protocol and reversible instance normalisation handling cross-disease scale differences. CAPE shifts mechanism into a learnable prior: K compartmental prototype embeddings interact with patch representations through cross-attention, regularised by a differentiable spectral-radius proxy

$$\frac{\sigma_{\min}(F)}{\sigma_{\max}(V)} \leq \rho(FV^{-1}) \leq \frac{\sigma_{\max}(F)}{\sigma_{\min}(V)}, \tag{6}$$

where F and V are the next-generation matrices of new infections and transitions, and $\rho(FV^{-1})$ is the basic-reproduction number R_0; the bound replaces the unstable matrix inverse with stable singular-value ratios, exposing R_0 as a target for gradient-based regularization (Liu et al., 2026). Mantis takes the heaviest position, pretraining entirely on synthetic trajectories; ablations confirm that both the encoder choice and simulation-only pretraining are critical (Dudley et al., 2025b). A contrastive baseline bypasses mechanism coupling entirely: Kalahasti et al. (2025) evaluate off-the-shelf foundation models including TabPFN-TS, Chronos, and TimesFM on respiratory-disease surveillance without epidemic-specific pretraining to establish a zero-shot performance baseline.

4.2 Supervised Fine-tuning

The architectural-coupling systems of §3.1 split along a training-effort axis. Gong et al. (2025) keep the LLM backbone frozen and train only an external projector that maps graph-neural-network and mobility-tensor encodings into the LLM embedding space, with direction-aware soft edge prompts and per-timestep gating fusing the two branches. PandemicLLM takes the opposite position, fully fine-tuning LLaMA-2 over ordinal hospitalisation-trend classification (Du et al., 2025); a GRU bridge encodes case-count time series into trainable special tokens, and ablation demonstrates that this small recurrent encoder remains critical for serialising continuous epidemic signal into the LLM context.

4.3 Prompt-driven Synthesis

The verification-time and narrow-end simulation systems are prompt-only at the language-model level, with mechanism entering through inspectable intermediate outputs rather than parameter updates. Output bandwidth ranges from a single scalar (Chae et al., 2025) to typed flow graphs under verification predicates (Datta et al., 2026) and multi-phase workflows with tool calling (Samaei et al., 2025).

4.4 Agent Conditioning

The runtime systems of §3.2 share a common technical surface: persona conditioning and retrieved memory, with no supervised fine-tuning or reinforcement learning. Chopra et al. (2025) address the population-scale cost through archetype-level prompting: GPT-3.5 is queried once per demographic cluster α, and each agent's Bernoulli compliance probability — whether to follow or deviate from a public-health intervention — is approximated by Monte Carlo averaging,

$$\hat{p}_\alpha = \frac{1}{M}\sum_{m=1}^{M} \mathrm{LLM}(\mathrm{prompt}_\alpha, \mathrm{context}_t)\,, \tag{7}$$

where prompt_α encodes the demographic persona of archetype α, $\mathrm{context}_t$ is the epidemic context at time t, and each of M independent calls returns a binary comply/deviate decision. Structural simulator parameters such as transmission and recovery rates are calibrated end-to-end by

Coupling pattern	When active	State owner	What evaluation can verify	What remains opaque	Representative systems
Training-time	Before deployment	Neural model	Forecast accuracy on held-out real outbreaks	Whether learned representations encode mechanistic structure	Dudley et al., 2025b; Kamarthi and Prakash, 2023
Architectural	Forward pass	Neural model with embedded prior	Accuracy + prototype / projector alignment signals	Whether prototypes encode true compartmental semantics	Liu et al., 2026; Gong et al., 2025; Du et al., 2025; Chae et al., 2025
Verification-time	After generation, before execution	Verifier + typed IR	Structural admissibility (mass conservation, well-typedness)	Whether structural correctness implies causal correctness	Datta et al., 2026; Samaei et al., 2025
Runtime	Each simulation step	Simulator + LLM agents	Agent action consistency; trajectory fit to real data	Whether agent decisions respect constraints under distribution shift	Chopra et al., 2025; Williams et al., 2023; Hou et al., 2025; Shi et al., 2026; Bo et al., 2026; Wang et al., 2026; Aoki and Ghaffarzadegan, 2026

Table 1: Coupling position determines what evaluation can verify. Each row shows what a coupling pattern makes verifiable and what it leaves beyond evaluation's reach. As coupling moves later in the lifecycle (top to bottom), inspectability increases but the nature of opacity shifts from representational to behavioral. This asymmetry motivates the coupling-aware evaluation dimensions proposed in §6.

gradient flow around the frozen LLM through the differentiable substrate. The remaining systems share this prompt-only surface: Williams et al. (2023) demonstrate that SEIR-style dynamics emerge from prompted reasoning alone, while Aoki and Ghaffarzadegan (2026) show that a short paragraph of mechanistic knowledge substantially improves policy stability.

An observation links the techniques across §4.1–§4.4: as training effort decreases, mechanism migrates leftward along the model lifecycle, from synthetic training data, to a learnable prior with a differentiable theory bound, to verification predicates over intermediate representations, to a brief mechanistic paragraph in the prompt. This inverts the pattern of imposing stronger constraints through heavier training; mechanism instead surfaces into inspectable interfaces (prototype banks, flow graphs, structured action vocabularies) that §5 must contend with. Appendix B details how grounding method constrains their form.

5 Evaluation and Benchmarks

5.1 Shared Infrastructure

The epidemic-intelligence community has converged on three operational forecasting hubs with shared protocols: CDC FluSight, running since 2013–14 (McGowan et al., 2019); the US COVID-19 Forecast Hub, active since 2020 (Cramer et al., 2022); and the ECDC RespiCast hub, launched in 2023 (European Centre for Disease Prevention and Control, 2023). All three score weekly quantile forecasts against the Weighted Interval Score family of proper scoring rules introduced by Bracher et al. (2021), complemented by mean absolute error and prediction-interval coverage. The Hubverse standard (Consortium of Infectious Disease Modeling Hubs et al., 2025) provides the submission schema and multi-team ensembling toolkit across these hubs. For scenario projections under counterfactual policy assumptions, the COVID-19 Scenario Modeling Hub has been evaluated by Howerton et al. (2023), which is the community's canonical reference for scenario-model evaluation under mixed factual and counterfactual conditions. Non-pharmaceutical intervention causal estimation lacks a comparable shared benchmark.

5.2 Evaluation Practice

Table 2 in appendix records the evaluation methodology and output type across the surveyed systems grouped by sub-coupling pattern. The forecasting subfamily shows uneven protocol adoption: Mantis backtests against the COVID-19 Forecast Hub archive with WIS scoring across 78 models and 16 diseases (Dudley et al., 2025b); Kalahasti et al. (2025) evaluate on multi-disease surveillance across France and Italy with MAE, MAPE, and prediction-interval coverage; Chae et al. (2025) report MAE with 90% prediction-interval coverage against foundation-model baselines (Ansari et al., 2024; Das et al., 2024); while CAPE and EpiLLM report point metrics only (Liu et al., 2026; Gong et al., 2025). No forecasting system in the corpus operationally submits to a hub round. The simulation subfamily shows no shared protocol: nine systems yield nine distinct evaluation strategies, with

scoring rules ranging from flow-graph admissibility predicates (Datta et al., 2026) to Kolmogorov–Smirnov tests against survey baselines (Hou et al., 2025; Bo et al., 2026), and no common dataset standard or submission infrastructure. Their output types further fragment evaluation: the forecasting subfamily produces incidence trajectories or ordinal trend classes (Du et al., 2025), while the simulation subfamily spans executable models (Datta et al., 2026), scientific manuscripts (Samaei et al., 2025), behavior traces (Chopra et al., 2025; Williams et al., 2023), and policy trajectories (Shi et al., 2026; Aoki and Ghaffarzadegan, 2026) — artifact categories that no single scoring rule can accommodate. Three simulation systems evaluate on self-generated populations without real-world calibration anchors (Williams et al., 2023; Wang et al., 2026; Aoki and Ghaffarzadegan, 2026), which limit external validity claims to qualitative plausibility.

6 Directions and Challenges

The transition from model-centric to system-centric research (Chen et al., 2025; Lee et al., 2025) requires evaluation paradigms that test coupling quality, not just model skill, and positions foundation models not only as predictive tools but as instruments for mechanistic discovery. The gaps surfaced by our evaluation audit point toward three directions and three challenges.

6.1 Directions and Vision

Coupling-aware evaluation Current evaluation does not distinguish coupling positions. The taxonomy shows a structural reason: forecast uncertainty must decompose into model and substrate contributions, and this decomposition differs by coupling position. A coupling-aware benchmark should test each system at its coupling boundary, from generalization to held-out transmission modes in training-time systems to agent consistency under simulator perturbation in runtime systems. Boundary-level failure attribution remains difficult (Zhang et al., 2025), and current behavior-driven evaluation frameworks (Bhagat et al., 2025) do not yet index by coupling position. What fraction of a training-time-coupled model's accuracy is attributable to mechanistic structure rather than statistical pattern matching, and does that fraction degrade under the distribution shifts that novel pathogens inevitably introduce? Coupling-specific benchmarks are the missing evaluation infrastructure.

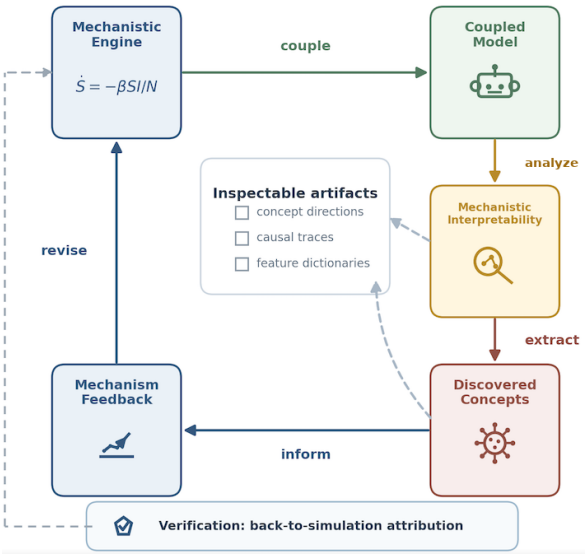

Figure 2: Bidirectional interpretability paradigm for responsible epidemic intelligence. Each stage and the inspectable artifacts it yields are described in §6.1.

Interpretability as a bidirectional bridge The coupling taxonomy produces structurally distinct inspectable artifacts at each lifecycle position (Table 1), yet no system in the corpus applies interpretability methods to any of them (Table 2). Recent advances make this feasible: attribution graphs trace computational structure through domain reasoning (Lindsey et al., 2025; Ameisen et al., 2025), sparse autoencoders recover meaningful features from foundation models (Simon and Zou, 2025; Marks et al., 2025; Mishra, 2026), and back-to-simulation attribution links predictions to simulated regimes (Dudley et al., 2025a). General-purpose time-series foundation models already perform competitively on epidemic surveillance data without explicit mechanistic coupling (Kalahasti et al., 2025), raising the question of what epidemiological structure these models implicitly encode.

We envision a bidirectional paradigm for responsible epidemic intelligence (Figure 2). Mechanistic interpretability techniques analyze the coupled model's internal representations, yielding inspectable artifacts: concept directions identifying activation subspaces for epidemiological factors such as transmission regime or seasonal forcing, causal traces revealing which components drive specific forecasts, and feature dictionaries decomposing computations into interpretable units. These artifacts are decoded into discovered concepts: do specific features activate selectively for known epi-

demiological quantities such as effective reproduction number or generation interval, or do they encode statistical regularities with no mechanistic counterpart? Crucially, some discovered concepts may represent knowledge beyond current epidemiological understanding, just as concepts extracted from AlphaZero's representations proved to encode novel chess strategies that grandmasters verified and learned from (Schut et al., 2025). In epidemic intelligence, domain experts would similarly validate whether a discovered concept reflects a real epidemiological factor before it informs mechanism refinement: if a learned feature maps to behavioral response but no compartment captures this, the validated discovery motivates adding one. The refined mechanism is coupled back and verified through back-to-simulation attribution, testing whether updated predictions align with the simulated regimes from training. Such a cycle would transform interpretability from a post-hoc audit into a generative tool for responsible system design (Sharkey et al., 2025), ensuring that the factors driving a forecast remain traceable to mechanistic reasoning that domain experts can challenge.

Design principles and cross-domain transfer The taxonomy implies a design-space trade-off: training-time coupling maximizes generalizability but renders mechanism opaque; verification-time coupling maximizes inspectability but constrains expressivity; runtime coupling maximizes behavioral flexibility but limits scale and introduces validation challenges only beginning to be formalized (Larooij and Törnberg, 2025). The choice depends on the deployment priority. These patterns are not specific to epidemic intelligence: training-time coupling appears wherever models pretrain on simulator output (Meta FAIR, 2025); verification-time coupling wherever formal checkers gate model output (Hubert et al., 2026); runtime coupling wherever agents operate inside structured environments (Piao et al., 2026). A systems-theoretic perspective is emerging (Miehling et al., 2025), but no existing taxonomy foregrounds lifecycle position as the organizing axis. Whether hybrid strategies can soften this trade-off remains open. Epidemic intelligence offers a distinctive test bed: compartmental models provide a mechanistic verification signal stronger than LLM-as-judge evaluation (Lau et al., 2026; Ghareeb et al., 2026; Yamada et al., 2025).

6.2 Open Challenges

Three challenges surface from limitations the surveyed systems themselves admit.

Mechanism-faithful verification Structural admissibility does not guarantee that a verified intermediate representation encodes the correct causal mechanism: Datta et al. (2026) acknowledge that parameter fitting can compensate for structural mis-specification. No method currently exists to verify mechanistic faithfulness of either typed IRs or long-horizon agent decision traces, a gap paralleling semantic correctness evaluation in structured code generation. Current code-generation metrics verify execution correctness but not whether the generated program encodes the intended causal structure; mechanism-faithful verification requires metrics distinguishing admissible from causally correct representations.

Cross-disease and cross-context generalization Validation across the corpus is predominantly single-disease and single-region (Du et al., 2025; Liu et al., 2026), with documented cross-context accuracy degradation (Bo et al., 2026). Systematic cross-disease transfer evaluation does not exist in the corpus, connecting to domain adaptation for grounded generation systems. Unlike standard domain adaptation, where only the text distribution shifts, cross-disease transfer also changes the mechanistic grounding: the generation component must adapt to new transmission dynamics across pathogens.

Reliability and scalability at deployment scale Multiple systems report reliability concerns: Samaei et al. (2025) find hallucinated references in generated reports; Hou et al. (2025) document inconsistent demographic-profile responses; Shi et al. (2026) observe policy variability in weaker backbones. Scalability compounds this: Chopra et al. (2025) show that per-agent LLM queries do not scale to population-sized simulations, and their archetype approximation trades heterogeneity for tractability. Figure 2 offers a partial path: if mechanistic interpretability can identify which agent behaviors are grounded in mechanistic reasoning and which arise from language-model priors, practitioners could selectively constrain the latter while preserving the former. Realizing this at population scale remains open.

Limitations

Scope and taxonomy rationale The corpus comprises mechanism-coupled epidemic-intelligence systems published between late 2022 and May 2026 (post-GPT era) in which a mechanistic component is coupled to a foundation model, excluding classical hybrid models without a foundation-model component (Ye et al., 2025), pure time-series foundation models without explicit mechanistic coupling (Lau et al., 2026), and the broader AI-for-epidemiology landscape (Kraemer et al., 2025). Public-health modeling traditionally distinguishes forecasting (predicting future trajectories) from scenario modeling (evaluating counterfactual interventions). Our taxonomy groups systems by lifecycle position rather than by the traditional forecasting-versus-scenario-modeling distinction. This choice means that systems serving different public-health functions (prediction versus counterfactual policy analysis) may share a row, potentially obscuring functional differences that matter to practitioners. Some system placements will inevitably be contested; we view this ambiguity as characteristic of an developing field rather than a flaw in the taxonomy.

Taxonomy and analytical claims The four coupling patterns are derived from a principled axis (lifecycle position) rather than ad hoc clustering, but they may not be exhaustive: reward-signal coupling has not yet been observed in epidemic intelligence but could emerge, and alternative axes such as data modality or deployment maturity would yield different groupings. Systems are assigned to the coupling position governing their primary inference pathway; borderline cases are discussed in §3. The argument that coupling position determines what evaluation can verify is a structural observation grounded in the cross-reading of Table 1 and Appendix Table 2, not an empirically tested hypothesis; controlled experiments comparing failure modes across coupling positions under matched distribution shift would strengthen or refute this claim.

Conceptual Proposal The bidirectional interpretability paradigm proposed in §6.1 is a conceptual proposal grounded in recent advances in mechanistic interpretability; no empirical validation of this cycle exists for epidemic-intelligence systems.

References

Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan, and B. Aditya Prakash. 2019. EpiDeep: Exploiting embeddings for epidemic forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 577–586.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Tom Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. Circuit tracing: Revealing computational graphs in language models.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. 2024. Chronos: Learning the language of time series. *Preprint*, arXiv:2403.07815.

Goshi Aoki and Navid Ghaffarzadegan. 2026. AI agents as policymakers in simulated epidemics. *Preprint*, arXiv:2601.04245.

Pranav Bhagat, K. N. Ajay Shastry, Pranoy Panda, and Chaitanya Devaguptapu. 2025. Evaluating compound AI systems through behaviors, not benchmarks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

Lujia Bo, Mingxuan Chen, Youduo Chen, Xiaofan Gui, Jiang Bian, Chunyan Wang, and Yi Liu. 2026. From risk perception to behavior: Large language models-based simulation of pandemic prevention behaviors. *Preprint*, arXiv:2601.03552.

Johannes Bracher, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich. 2021. Evaluating epidemic forecasts in an interval format. *PLOS Computational Biology*, 17(2):e1008618.

John S. Brownstein, Clark C. Freifeld, and Lawrence C. Madoff. 2009. Digital disease detection — harnessing the Web for public health surveillance. *New England Journal of Medicine*, 360(21):2153–2157.

John S. Brownstein, Benjamin Rader, Christina M. Astley, and Huaiyu Tian. 2023. Advances in artificial intelligence for infectious-disease surveillance. *New England Journal of Medicine*, 388(17):1597–1607.

Joongwon Chae, Runming Wang, Chen Xiong, Gong Yunhan, Lian Zhang, Ji Jiansong, Dongmei Yu, and Peiwu Qin. 2025. Beyond curve fitting: Neuro-symbolic agents for context-aware epidemic forecasting. *Preprint*, arXiv:2511.23276.

Jiayi Chen, Junyi Ye, and Guiling Wang. 2025. From standalone LLMs to integrated intelligence: A survey of compound AI systems. *Preprint*, arXiv:2506.04565.

Ayush Chopra, Shashank Kumar, Nurullah Giray-Kuru, Ramesh Raskar, and Arnau Quera-Bofarull. 2025. On the limits of agency in agent-based models. In *Proceedings of AAMAS*.

Consortium of Infectious Disease Modeling Hubs, Melissa Kerr, Rebecca Borchering, Alvaro Castro Rivadeneira, Lucie Contamin, Sebastian Funk, Harry Hochheiser, Emily Howerton, Anna Krystalli, Li Shandross, and Nicholas G. Reich. 2025. Coordinating collaborative infectious disease modeling projects with the Hubverse. *medRxiv*.

Estee Y. Cramer, Yuxin Huang, Yijin Wang, Evan L. Ray, Matthew Cornell, Johannes Bracher, Andrea Brennen, Alvaro J. Castro Rivadeneira, Aaron Gerding, Katie House, Dasuni Jayawardena, Abdul H. Kanji, Ayush Khandelwal, Khoa Le, Jarad Niemi, Apurv Shah, Nutcha Wattanachit, Martha W. Zorn, Nicholas G. Reich, and 1 others. 2022. The united states COVID-19 forecast hub dataset. *Scientific Data*, 9:462.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 10148–10167.

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

Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534.

Hongru Du, Jianan Zhao, Yang 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*, 5:467–480.

Carson Dudley, Reiden Magdaleno, Christopher Harding, and Marisa Eisenberg. 2025a. Simulation as supervision: Mechanistic pretraining for scientific discovery. *Preprint*, arXiv:2507.08977.

Carson Dudley, Reiden Magdaleno, Christopher Harding, Ananya Sharma, Emily Martin, and Marisa Eisenberg. 2025b. Mantis: A foundation model for mechanistic disease forecasting. *Preprint*, arXiv:2508.12260.

Stephen Eubank, Hasan Guclu, V. S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, Zoltán Toroczkai, and Nan Wang. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184.

European Centre for Disease Prevention and Control. 2023. RespiCast: The new european respiratory diseases forecasting hub.

Neil M. Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, and 1 others. 2020. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Report 9, Imperial College London.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Jon M. Laurent, Muhammed T. Razzak, Andrew D. White, Michaela M. Hinks, Samuel G. Rodriques, and 1 others. 2026. A multi-agent system for automating scientific discovery. *Nature*.

Chenghua Gong, Rui Sun, Yuhao Zheng, Juyuan Zhang, Tianjun Gu, Liming Pan, and Linyuan Lü. 2025. EpiLLM: Unlocking the potential of large language models in epidemic forecasting. *Preprint*, arXiv:2505.12738.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 16115–16152. PMLR.

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. In *Proceedings of the Conference on Language Modeling (COLM)*.

Emily Howerton, Lucie Contamin, Luke C. Mullany, Michelle Qin, Nicholas G. Reich, Samantha Bents, Rebecca K. Borchering, and 1 others. 2023. Evaluation of the US COVID-19 Scenario Modeling Hub for informing pandemic response under uncertainty. *Nature Communications*, 14:7260.

Thomas Hubert, Rishi Mehta, Laurent Sartran, Miklós Z. Horváth, Goran Žužić, Eric Wieser, Aja Huang, Julian Schrittwieser, Yannick Schroecker, and 1 others. 2026. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651(8106):607–613. Published online 12 November 2025.

Shuyu Jiang, Hao Ren, Yichang Gao, Yi-Cheng Zhang, Li Qi, Dayong Xiao, Jie Fan, Rui Tang, and Wei Wang. 2026. Large language models for spreading dynamics in complex systems.

Suprabhath Kalahasti, Benjamin Faucher, Boxuan Wang, Claudio Ascione, Ricardo Carbajal, Maxime Enault, Christophe Vincent Cassis, Titouan Launay, Caroline Guerrisi, Pierre-Yves Boëlle, Federico

Baldo, and Eugenio Valdano. 2025. Foundation time series models for forecasting and policy evaluation in infectious disease epidemics. medRxiv 2025.02.24.25322795.

Harshavardhan Kamarthi, Lingkai Kong, Alexander Rodríguez, Chao Zhang, and B. Aditya Prakash. 2021. When in doubt: Neural non-parametric uncertainty quantification for epidemic forecasting. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 19796–19807.

Harshavardhan Kamarthi and B. Aditya Prakash. 2023. PEMS: Pre-trained epidemic time-series models. *Preprint*, arXiv:2311.07841.

William Ogilvy Kermack and Anderson Gray McKendrick. 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115:700–721.

Moritz U. G. Kraemer, Joseph L.-H. Tsui, Serina Y. Chang, Spyros Lytras, Mark P. Khurana, Samantha Vanderslott, Sumali Bajaj, Samuel V. Scarpino, Oliver G. Pybus, Simon Cauchemez, Samir Bhatt, and 1 others. 2025. Artificial intelligence for modelling infectious disease epidemics. *Nature*, 638(8051):623–635.

Maik Larooij and Petter Törnberg. 2025. Validation is the central challenge for generative social simulation: a critical review of LLMs in agent-based modeling. *Artificial Intelligence Review*.

Max S. Y. Lau, C. Jessica E. Metcalf, Zewen Liu, Bryan T. Grenfell, and Wei Jin. 2026. Toward AI foundation models for epidemics: Promise, challenges, and paths forward. *Proceedings of the National Academy of Sciences*, 123(13):e2526192123.

Yu-Ang Lee, Guan-Ting Yi, Mei-Yi Liu, Jui-Chao Lu, Guan-Bo Yang, and Yun-Nung Chen. 2025. Compound AI systems optimization: A survey of methods, challenges, and future directions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 28748–28763, Suzhou, China. Association for Computational Linguistics.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, and 1 others. 2025. On the biology of a large language model. Anthropic Transformer Circuits Thread, March 27.

Zewen Liu, Juntong Ni, Bohan Wang, Max S. Y. Lau, and Wei Jin. 2026. Pre-training epidemic time series forecasters with compartmental prototypes. In *Proceedings of ICLR*.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *ICLR 2025 (Oral)*.

Craig J. McGowan, Matthew Biggerstaff, Michael Johansson, Kara M. Apfeldorf, Michal Ben-Nun, Logan Brooks, Matteo Convertino, Madhav Erraguntla, David C. Farrow, John Freeze, Saurabh Ghosh, Sangwon Hyun, Sasikiran Kandula, Joceline Lega, Yang Liu, Nicholas Michaud, Hiroyuki Morita, Jarad Niemi, Naren Ramakrishnan, and 9 others. 2019. Collaborative efforts to forecast seasonal influenza in the united states, 2015–2016. *Scientific Reports*, 9:683.

Meta FAIR. 2025. Code world model. Released September 2025.

Erik Miehling, Karthikeyan Natesan Ramamurthy, Kush R. Varshney, Matthew Riemer, Djallel Bouneffouf, John T. Richards, Amit Dhurandhar, Elizabeth M. Daly, Michael Hind, Prasanna Sattigeri, Dennis Wei, Ambrish Rawat, Jasmina Gajcin, and Werner Geyer. 2025. Agentic AI needs a systems theory. *Preprint*, arXiv:2503.00237.

Anurag Mishra. 2026. Dissecting Chronos: Sparse autoencoders reveal causal feature hierarchies in time series foundation models. *Preprint*, arXiv:2603.10071.

Oliver W. Morgan, Ximena Aguilera, Andrea Ammon, John Amuasi, Ibrahima Socé Fall, Tom Frieden, David Heymann, Chikwe Ihekweazu, Eun-Kyeong Jeong, Gabriel M. Leung, Barbara Mahon, John Nkengasong, Farah Naz Qamar, Anne Schuchat, Lothar H. Wieler, and Scott F. Dowell. 2021. Disease surveillance for the COVID-19 era: Time for bold changes. *The Lancet*, 397(10292):2317–2319.

Quynh C. Nguyen, Isha Yardi, Francia Ximena Marin Gutierrez, Heran Mane, and Xiaohe Yue. 2022. Leveraging 13 million responses to the U.S. COVID-19 Trends and Impact Survey to examine vaccine hesitancy, vaccination, and mask wearing, January 2021–February 2022. *BMC Public Health*, 22:1911.

Viet Bach Nguyen, Truong Son Hy, Long Tran-Thanh, and Nhung Nghiem. 2023. Predicting COVID-19 pandemic by spatio-temporal graph neural networks: A New Zealand's study. *Preprint*, arXiv:2305.07731.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations (ICLR)*.

George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. 2021. Transfer graph neural networks for pandemic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4838–4845.

Christophe Paquet, Denis Coulombier, Reinhard Kaiser, and Massimo Ciotti. 2006. Epidemic intelligence: A new framework for strengthening disease surveillance in Europe. *Euro Surveillance*, 11(12):212–214.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2026. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

Mohammad Hossein Samaei, Faryad Darabi Sahneh, Lee W. Cohnstaedt, and Caterina Scoglio. 2025. EpidemIQs: Prompt-to-paper LLM agents for epidemic modeling and analysis. *Preprint*, arXiv:2510.00024.

Lisa Schut, Nenad Tomašev, Thomas McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. 2025. Bridging the human–ai knowledge gap through concept discovery and transfer in alphazero. *Proceedings of the National Academy of Sciences*, 122(13):e2406675122.

Lee Sharkey, Bilal Chughtai, Joshua Batson, and 1 others. 2025. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*.

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. *Preprint*, arXiv:2601.09264.

Elana Simon and James Zou. 2025. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22:2107–2117.

Willem G. van Panhuis, John Grefenstette, Su Yon Jung, Nian Shong Chok, Anne Cross, Heather Eng, Bruce Y. Lee, Vladimir Zadorozhny, Shawn Brown, Derek Cummings, and Donald S. Burke. 2013. Contagious diseases in the United States from 1888 to the present. *New England Journal of Medicine*, 369(22):2152–2158.

Shanrui Wang, Huiyong Liu, S. Zhang, and Q. Yang. 2026. An LLM-driven multi-agent simulation framework for coupled epidemic–economic dynamics. *Information*, 17(3):259.

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

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 53140–53164. PMLR.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *Preprint*, arXiv:2504.08066.

Yang Ye, Abhishek Pandey, Carolyn Bawden, Dewan Md Sumsuzzman, Rimpi Rajput, Affan Shoukat, Burton H. Singer, Seyed M. Moghadas, and Alison P. Galvani. 2025. Integrating artificial intelligence with mechanistic epidemiological modeling: a scoping review of opportunities and challenges. *Nature Communications*, 16:581.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11121–11128.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. 2025. AgenTracer: Who is inducing failure in the LLM agentic systems? *Preprint*, arXiv:2509.03312.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115.

A Complete Evaluation Details

Table 2: Evaluation methodology and output type across mechanism-coupled epidemic-intelligence systems, grouped by sub-coupling pattern (§3). No system in the corpus reports a mechanistic-interpretability probe of model internals, a uniform gap motivating the methodology-transfer opportunities developed in §6.1. †PandemicLLM outputs an ordinal hospitalisation-trend class, not a numerical trajectory. ‡Kalahasti et al. (2025) evaluate frozen off-the-shelf time-series foundation models without mechanism coupling; this system is not part of the mechanism-coupled corpus but is included as a contrastive zero-shot baseline.

System	Dataset	Metric	Baseline	Output
Training-time coupling (forecasting)				
PEMS (Kamarthi and Prakash, 2023)	Project Tycho (van Panhuis et al., 2013)	RMSE + cross-entropy (cross-disease)	Informer (Zhou et al., 2021), PatchTST (Nie et al., 2023), DLinear (Zeng et al., 2023), EpiDeep (Adhikari et al., 2019), EpiFNP (Kamarthi et al., 2021) + SSL ablations	Point forecast
Mantis (Dudley et al., 2025b)	16 diseases incl. COVID Hub archives (Cramer et al., 2022)	MAE + WIS + 50%/95% PI coverage	78 baselines across 16 diseases (incl. CDC COVID-19 Forecast Hub) (Cramer et al., 2022)	Prob. forecast
Architectural coupling (forecasting)				
Chae (Chae et al., 2025)	HFMD surveillance + context signals	MAE + 90% PI coverage	Chronos (Ansari et al., 2024), TimesFM (Das et al., 2024)	Prob. forecast + rationale
CAPE (Liu et al., 2026)	17 diseases from Project Tycho (curated via EpiLearn)	MSE / MAE / RMSE	PatchTST (Nie et al., 2023), DLinear (Zeng et al., 2023), EpiDeep (Adhikari et al., 2019), MOMENT (Goswami et al., 2024), Chronos (Ansari et al., 2024), Moirai (Woo et al., 2024), ARIMA	Point forecast
EpiLLM (Gong et al., 2025)	Spatio-temporal COVID-19	MAE / RMSE	MPNN (Panagopoulos et al., 2021), ATMGNN (Nguyen et al., 2023), LSTM, ARIMA	Point forecast
PandemicLLM† (Du et al., 2025)	19-month US 50-state COVID-19 (Dong et al., 2020)	Accuracy + Brier + RPS	GRU, LSTM, bi-LSTM, ARIMA, PrevTrend	Ordinal trend class
Verification-time coupling (simulation)				
EpiAgent (Datta et al., 2026)	MIDAS scenario case studies	Flow-graph admissibility (V&V)	Unguided LLM generation (no IR)	Executable model
EpidemIQs (Samaei et al., 2025)	Prompt-to-paper workflow	79% task success + LLM-as-judge	Single-agent LLM (same backbone)	Scientific manuscript
Runtime coupling (simulation)				
AgentTorch (Chopra et al., 2025)	NYC 8.4M agents + mobility data (Dong et al., 2020)	Wave + unemployment trajectory fit	Heuristic vs. LLM agent architectures	Behavior trace
Williams (Williams et al., 2023)	Self-generated agent population	Qualitative behavioral plausibility	Three prompt-information conditions	Behavior trace
VacSim (Hou et al., 2025)	100 agents + Delphi CTIS survey (Nguyen et al., 2022)	Alignment vs. CTIS + LLM-as-judge	Six LLM backbones (Llama, Qwen, GPT-4o, Claude)	Attitude trajectory
Shi (Shi et al., 2026)	US 50-state COVID-19 2020 (Dong et al., 2020)	Counterfactual gain (63.7% / 40.1%)	Observed US 2020 NPI outcomes	Policy trajectory
Bo (Bo et al., 2026)	Beijing residents survey (R1 / R2)	Behavior calibration vs. survey (zero-shot / few-shot / cross-context)	Survey ground truth	Behavior trajectory

Continued on next page

System	Dataset	Metric	Baseline	Output
Wang (Wang et al., 2026)	Self-generated coupled epi–econ. simulation	Cross-LLM trajectory consistency (infections, mortality, unemployment, wealth)	Different LLM backends + policy regimes	Epi-econ. trajectory
Aoki (Aoki and Ghaffarzadegan, 2026)	Self-generated SEIR sandbox	Decision quality + stability across memory	Single vs. ensemble agents + Knowledge prompt	Policy decision
Zero-shot baseline (no mechanism coupling)[‡]				
Kalahasti (Kalahasti et al., 2025)	7 diseases across France / Italy	MAE + MAPE + PI coverage	SARIMA + mechanistic SEIR baselines	Prob. forecast

B Grounding and Validation Analysis

Table 3: Grounding strategy and validation anchor across the 16 surveyed systems. *Grounding* refers to how mechanistic domain knowledge (compartmental dynamics, demographic priors, epidemiological constraints) is injected into or encoded within the model. The first five columns characterize the grounding pathway and the system's functional interface; the sixth records the strongest external evidence each system validates against. Systems sharing the same grounding method are split into sub-rows when their validation anchors differ substantially.

Grounding method	Mechanism type	Model / agent role	Backbone	Weights	Validation anchor	Systems
Distributional pretraining on simulator output or multi-disease archive	Compartmental ODE	Forecast from learned representations	Custom transformer	Trained	Real-world forecasting benchmark (CDC Forecast Hub; national surveillance archives)	PEMS (Kamarthi and Prakash, 2023), Mantis (Dudley et al., 2025b)
Prototype alignment with compartmental semantics	Compartmental prototypes	Forecast through prototype-aligned encoding	Custom transformer + prototype dictionary	Trained	Real-world forecasting benchmark (multi-disease held-out evaluation)	CAPE (Liu et al., 2026)
Signal encoder (projector or bridge tokenizing epidemiological signal)	Epidemiological encoder	Forecast over encoded signal	Frozen LLM + learned projector	Projector / fine-tuned	Real-world forecasting benchmark (state-level hospitalisation; spatio-temporal case counts)	EpiLLM (Gong et al., 2025), PandemicLLM (Du et al., 2025)
Prompt-based knowledge injection	Agent-based SEIR / network dynamics	Citizen / policymaker emitting behavioral parameters	Frozen commercial LLM	Frozen	Real survey or administrative data (Delphi vaccine survey; Beijing resident survey; US Bureau of Labor Statistics)	VacSim (Hou et al., 2025), Bo (Bo et al., 2026), AgentTorch (Chopra et al., 2025)
Prompt-based knowledge injection	Compartmental ODE	Policymaker emitting policy parameters	Frozen commercial LLM	Frozen	Counterfactual policy evaluation	Shi (Shi et al., 2026), Aoki (Aoki and Ghaffarzadegan, 2026)
Prompt-based knowledge injection	Social network / agent-based SEIR + multi-layer network	Citizen emitting behavioral parameters	Frozen commercial LLM	Frozen	Self-generated scenarios (synthetic population; sandbox simulation)	Williams (Williams et al., 2023), Wang (Wang et al., 2026)

Grounding method	Mechanism type	Model / agent role	Backbone	Weights	Validation anchor	Systems
LLM-as-context-interpreter + neuro-symbolic statistical calibration	Statistical count model	Compress context into model input	Frozen commercial LLM	Frozen	Real surveillance with probabilistic calibration (HFMD case counts; Poisson/NB scoring)	Chae (Chae et al., 2025)
RAG + flow-graph verification (EpiAgent); Multi-agent workflow + LLM-as-judge (EpidemIQs)	Compartmental ODE + social network	Model builder generating simulator	Frozen commercial LLM + RAG	Frozen	Self-generated scenarios (LLM-as-judge; workflow-completion rate)	EpiAgent (Datta et al., 2026), EpidemIQs (Samaei et al., 2025)