

# Beyond Memory and Skills

Symbolic World Models as a Learning Primitive for AI Agents

Ninad Jagdish

Dwnrt Absrd Pte. Ltd.

Singapore

August 2026

## Abstract

The world contains more than humanity has written down and more than humanity yet understands. Large language models (LLMs) are trained on vast amounts of human-articulated knowledge and from it acquire broad latent representations of the world. Yet these representations, while powerful, are necessarily incomplete. They are limited by what has been observed, expressed, preserved, and made available. Memory and reusable skills give AI agents ways to learn beyond this inheritance: memory preserves what an agent has encountered, while skills preserve useful ways of acting. This paper examines symbolic world models as another learning primitive alongside memory and skills. Symbolic world models are persistent, explicit hypotheses about how a bounded part of the world works that an agent can inspect, interrogate, test, and revise through interaction. Such artifacts may be conceptual, including qualitative causal maps, or executable, allowing reproducible counterfactual exploration. The paper distinguishes symbolic world models from latent neural representations, memory, and skills and articulates a lifecycle for their formation, interrogation, validation, revision, and abstraction. Cross-model abstraction can turn situated learning into reusable causal archetypes, potentially reducing the observations, interactions, and representational search required in later situations. System dynamics is presented as an example of a constrained executable modeling language for organizational systems shaped by accumulation, feedback, delay, and nonlinearity. A current neurosymbolic agent implementation illustrates how this can be done in practice. Evaluation is treated as a plural problem: controlled benchmarks can measure bounded correctness, reliability, and efficiency, while situated qualitative assessment is needed to examine structural defensibility, inspectability, warranted trust, learning, and usefulness for a declared purpose.

## 1 Introduction: Beyond Memory and Skills

The world available to an AI system is not the world itself. It is a record assembled from what people and machines have observed, expressed, digitized, preserved, selected, and made available. LLMs inherit remarkable breadth from that record. Their distributed parameters and context-dependent representations encode regularities about objects, institutions, behavior, and causation, providing powerful priors for interpretation and reasoning. This paper therefore begins from the abundance of knowledge in an LLM, not its absence. Those representations may reasonably be described as *world-model-like*; this does not imply that a general-purpose LLM contains a complete or grounded simulator of reality. Linguistic form and evidence from direct interaction remain different sources of knowledge (Bender and Koller, 2020).

The incompleteness is not only organizational. Humanity has not recorded everything it has observed, does not know everything that has been recorded, and has not discovered everything there is to know. Some relevant knowledge is private or local and has never appeared in pretraining. Some is tacit: people act on practices, distinctions, expectations, and mental models they cannot fully articulate (Polanyi, 2009). Some mechanisms change after a model is trained; others remain contested or unknown. Organizations make this general epistemic condition unusually visible because documents and operational data capture only part of the knowledge distributed across people, routines, incentives, exceptions, and informal coordination (Nonaka, 1994; Ford and Sterman, 1998).

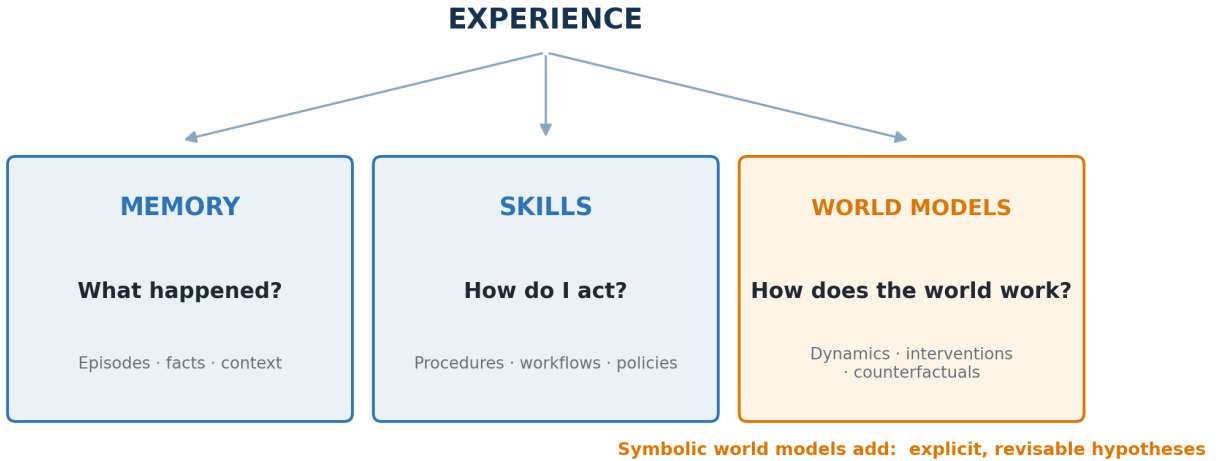
Agent systems increasingly supplement pretrained knowledge with external scaffolding. Tools provide actions; memory carries information across contexts; planners organize interaction; and reusable skills preserve successful procedures. These artifacts let an agent retain what it has encountered and reuse ways of acting (Sumers et al., 2023; Packer et al., 2023; Wang et al., 2023, 2024; Fang et al., 2026; Hu et al., 2025c).

Yet remembering and acting do not necessarily provide an explicit account of how a particular system works. Suppose an operations agent observes that a clinic’s backlog rises month after month. Memory can preserve the

episodes. A skill can encode a workflow for hiring clinicians or rescheduling appointments. The LLM may already understand accumulation, onboarding delays, and workload effects from pretraining. What is not yet explicit is which mechanisms matter in *this* clinic, how they interact, which assumptions are supported, and what follows under a contemplated intervention. The agent can possess relevant facts and procedures while lacking a situated, inspectable hypothesis of the system.

This paper examines whether a *symbolic world model* can serve as such a hypothesis. A symbolic world model may externalize knowledge already latent in the LLM, reorganize it into a testable form, combine it with local evidence, articulate previously tacit knowledge, or generate implications that were not previously visible. Whether every such implication is “new” in an absolute sense—or merely new relative to an agent, organization, or conversation—may be difficult to reliably determine. The more consequential question is whether the artifact produces useful knowledge: situated commitments that people and agents can inspect, interrogate, contest, execute where appropriate, improve with evidence, and apply to judgment or action.

Figure 1 summarizes the functional distinction motivating the paper. Memory asks what happened; skills ask how to act; a symbolic world model asks how a bounded part of the world might work and what might follow if that part changed. These categories are not an exhaustive ontology of cognition and can overlap in implementation. They identify different persistence functions within an agent architecture.



**Figure 1:** Three functional roles for persistent agent artifacts. Memory retains experience, skills retain procedures, and symbolic world models retain explicit, revisable hypotheses.

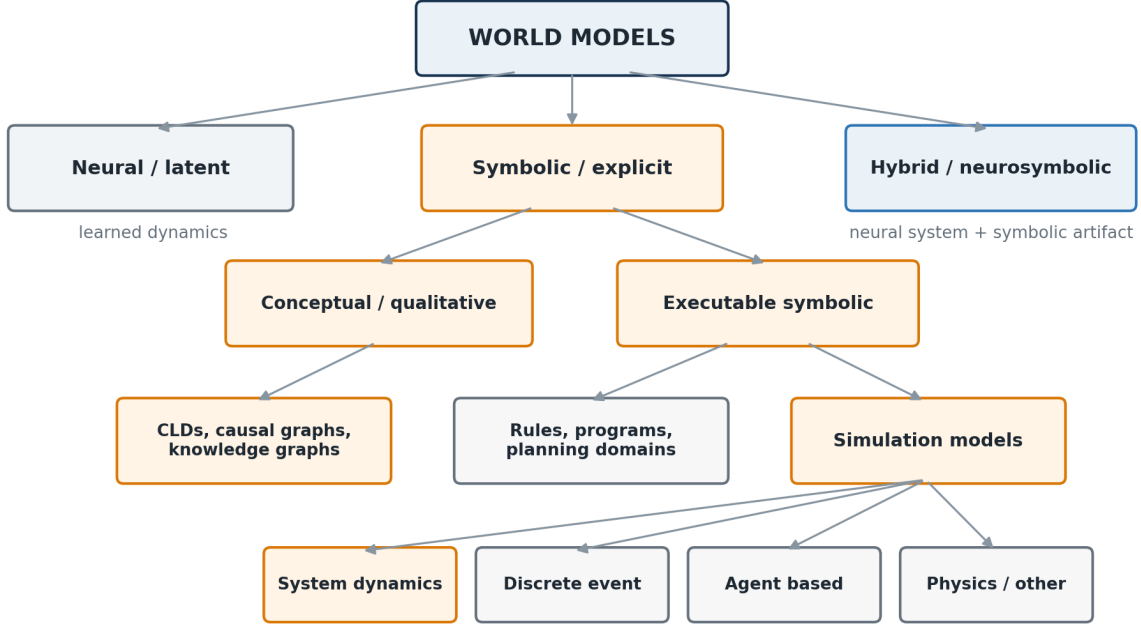
Explicit world models for agents are not new. Prior work has learned neural world models, induced rules and programs, connected agents to causal simulators, and inferred or revised symbolic environments through interaction (Ha and Schmidhuber, 2018; Schrittwieser et al., 2020; Tang et al., 2024; Agravante et al., 2023; Gkountouras et al., 2024; Khan et al., 2025; Hu et al., 2025a; Ma et al., 2026). Within this emerging landscape, this paper advances a functional account of symbolic world models as persistent learning artifacts in agent architectures. It describes the use of qualitative causal representations and executable simulation, develops a governed lifecycle for model formation, interrogation, revision, and abstraction, and proposes an evaluation framework that combines benchmarkable technical conformance with situated modeling usefulness. System dynamics (SD) provides a deliberately constrained and practically consequential setting through which to develop this broader argument.

## 2 Latent Knowledge and Explicit World-Model Artifacts

The term *model* carries several meanings in AI and simulation. An LLM is a neural model. A learned latent dynamics representation may be called a world model. A causal-loop diagram is a conceptual model. A stock–flow simulation is an executable model. Precision requires keeping these levels separate.

Figure 2 locates the paper’s subject within this larger family. Neural world models encode regularities in distributed learned representations. Symbolic world models express commitments in variables, relationships, rules, equations, or programs. A neurosymbolic system combines neural interpretation with one or more symbolic artifacts. Within the symbolic branch, conceptual models need not have complete execution semantics; executable models do.

Simulation models are executable models that generate behavior or state trajectories through time. SD simulation models are one subclass among several.



**Figure 2:** A working taxonomy. The paper focuses on external symbolic world models, including conceptual causal maps and executable simulation models; SD is one executable simulation language.

**Definition.** A *symbolic world model* is a persistent, inspectable artifact that explicitly represents hypotheses about the state, structure, relationships, dynamics, or constraints of a bounded part of the world and can be interrogated and revised. A *conceptual symbolic world model* represents these commitments qualitatively without complete execution semantics. An *executable symbolic world model* adds formal semantics that produce reproducible consequences under actions, interventions, or alternative assumptions. A *simulation model* is an executable world model whose operation generates behavior or state trajectories through time.

The word *bounded* matters. A useful model does not attempt to reproduce reality in full. It selects variables, mechanisms, time horizons, and levels of aggregation in relation to a purpose. Two models of the same organization may therefore differ legitimately because one investigates workforce capacity while another investigates customer adoption. Even for the same brief, competent modelers may choose different boundaries or structures, each illuminating a different mechanism. Model plurality is not evidence that quality is arbitrary; it means quality must be judged relative to purpose, evidence, and use rather than correspondence with a single canonical diagram.

Table 1 contrasts the artifact functions. The distinction being made here is functional. What changes is what the artifact is meant to preserve and how it can be interrogated.

**Table 1:** Functional distinctions among agent resources.

| Resource                     | Primary persistence              | Characteristic question                              | Characteristic affordance                                   |
|------------------------------|----------------------------------|------------------------------------------------------|-------------------------------------------------------------|
| Latent neural representation | Learned statistical regularities | What patterns can I infer from training and context? | Broad generalization without an explicit domain artifact    |
| Memory                       | Episodes, facts, context         | What happened or was learned here?                   | Retrieval across interactions                               |
| Skill                        | Procedure, workflow, policy      | How can I act?                                       | Reusable execution of a method                              |
| Symbolic world model         | Explicit structural hypothesis   | What could follow if this part of the world changed? | Inspection, counterfactual reasoning, testing, and revision |

The relationship between latent and symbolic knowledge is many-to-many. A symbolic artifact may make an existing prior explicit, expose contradictions among several priors, combine prior knowledge with private documents, interviews, observations, or data, or derive implications from familiar mechanisms. Each can be useful even when it is impossible to establish whether the resulting knowledge is novel in an absolute sense. Externalization can make tacit or distributed understanding discussable; formalization can make its consequences examinable; interaction can refine it; and comparison across models can reveal reusable patterns. The learning value lies in this progression from situated articulation to testing, revision, and transfer.

### 3 Related Work and Intellectual Neighbors

Research on agent memory now spans episodic, semantic, procedural, and working forms, together with consolidation and retrieval dynamics (Park et al., 2023; Packer et al., 2023; Du, 2026; Hu et al., 2025c). Skills and procedural memories preserve reusable ways of acting and can themselves evolve through experience (Wang et al., 2023, 2024; Fang et al., 2026; Gao et al., 2026). The present argument is not that these artifacts are insufficient in general, but that a persistent hypothesis about world structure performs a distinguishable function.

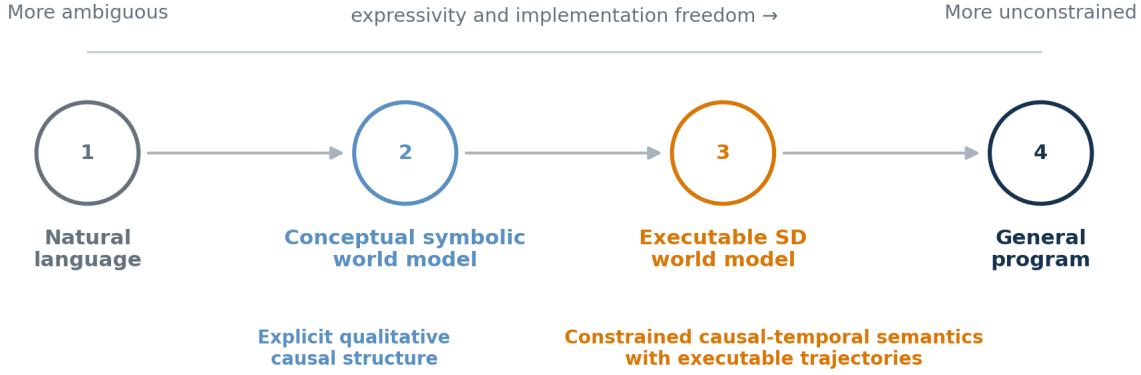
World models have a long history in model-based learning and control. Neural approaches learn compact latent dynamics for prediction and planning (Ha and Schmidhuber, 2018; Schrittwieser et al., 2020; Hafner et al., 2025). Programmatic and symbolic approaches make world rules explicit and executable (Tang et al., 2024; Hu et al., 2025a,b; Khan et al., 2025). ProPlay, for example, maintains a procedural world model that is revised with feedback (Ma et al., 2026). These works establish a neighboring research landscape within which persistent causal and simulation artifacts deserve more focused study.

Causal representation learning seeks variables and mechanisms that support intervention, transfer, and reasoning beyond observed correlations (Schölkopf et al., 2021). LLMs show causal competence but remain unreliable across formal causal tasks (Kiciman et al., 2023). Language Agents Meet Causality connects an LLM agent to a queryable causal world model, illustrating directly how situated causal structure can complement knowledge inherited from pretraining (Gkountouras et al., 2024). Causal-loop diagrams, causal graphs, and directed acyclic graphs can function as conceptual symbolic world models; with equations or probabilistic mechanisms, some become executable. SD differs from Pearlian structural causal models in its typical representation of feedback through coupled differential or difference equations rather than acyclic graphs (Pearl, 2009).

SD contributes a mature tradition of mental-model elicitation, endogenous explanation, learning through simulation, and model validation (Forrester, 1961; Sterman, 2000; Ford and Sterman, 1998; Barlas, 1996). Recent work has begun using LLMs to generate causal-loop diagrams and dynamical models, translate narrative descriptions into specialized modeling formalisms, and assist with model explanation and repair (Frydenlund et al., 2024; Liu and Keith, 2025; Luo et al., 2025; Giabbanelli et al., 2025). The BEAMS initiative develops open tests for several of these modeling operations (Metcalf and Schoenberg, 2026). This work is valuable for assessing technical conformance; Section 9 returns to the different problem of evaluating open-ended modeling usefulness.

### 4 From Representation to Experimentation

An LLM can reason *about* a system in language. A conceptual symbolic world model makes the hypothesized structure persistent and inspectable. An executable symbolic world model permits reasoning *with* the formalized hypothesis. The progression is illustrated in Figure 3. It is not a mandatory pipeline: a qualitative causal map can remain the appropriate endpoint when evidence does not justify quantification.



**Figure 3:** A representation spectrum. Moving right increases expressivity and implementation freedom. SD occupies a constrained middle position between qualitative causal structure and arbitrary generated code.

Suppose an agent hypothesizes that increasing clinic capacity should reduce delay. A causal map can expose the relevant accumulations, delays, feedback loops, and disputed links. An executable model can additionally ask: if capacity rises by 20 percent at month 12, under declared onboarding and productivity assumptions, what backlog trajectory follows over the next 36 months? Execution makes hidden commitments operational, reveals dimensional or logical inconsistencies, and produces repeatable consequences for comparison.

Execution does not make a model true. It can reveal error, but it can also launder unsupported assumptions into precise-looking trajectories. The benefit is therefore not numerical authority. It is disciplined experimentation on an explicit theory. Simulation generates synthetic experience from a fallible model; observation and intervention provide evidence from the world.

Natural language is expressive and accessible but ambiguous. Arbitrary code is also highly expressive, but it creates an enormous search space and permits idiosyncratic representations that are difficult to compare, validate, or revise. A constrained modeling language supplies an inductive bias. By limiting admissible constructs, it can make generation more tractable, artifacts more comparable, and validation more systematic. Constraint is valuable precisely because arbitrary representation is easy.

## 5 Why System Dynamics?

SD is not proposed as a universal world-model language. It is useful for a particular class of questions: how aggregate behavior arises over time from accumulation, feedback, delay, nonlinearity, and decision rules. Its core commitments map naturally onto a temporal causal hypothesis.

**Table 2:** SD constructs as commitments in an executable symbolic world model.

| Construct                       | World-model commitment                                                             |
|---------------------------------|------------------------------------------------------------------------------------|
| Stock                           | A state variable that accumulates and cannot change instantaneously without a flow |
| Flow                            | A rate governing change in a stock                                                 |
| Auxiliary                       | A relationship, decision rule, or intermediate quantity                            |
| Feedback loop                   | An endogenous causal pathway through which consequences return to influence causes |
| Delay                           | A separation between action, information, adjustment, and consequence              |
| Equation and unit               | Executable semantics and dimensional commitments                                   |
| Parameter and initial condition | Explicit assumptions about magnitude and starting state                            |

Organizations are especially useful as an epistemic testing ground. They are partially observable and tacit-knowledge-heavy. Their records omit informal routines, expectations, political constraints, workarounds, and local

causal beliefs. Their outcomes often emerge from backlogs, capacity, hiring, learning, churn, inventories, adoption, and delayed responses. SD does not remove ambiguity from these systems; it supplies a disciplined language in which selected ambiguities become discussable.

The constraint also has costs. Aggregate stocks can obscure heterogeneity. Continuous flows may misrepresent discrete events. Feedback models may omit strategic actors, spatial structure, or rare transitions. A responsible agent must therefore decide whether to stop at a causal map, proceed to SD, or select another formalism such as discrete-event simulation, agent-based modeling, a probabilistic causal model, a physics model, or a general program. The deeper capability is not always building an SD model. It is choosing the simplest representational commitment warranted by the question and evidence.

## 6 A Current System Instantiation

A current neurosymbolic agent implementation illustrates one way this progression can be integrated into a usable workflow. It accepts natural-language descriptions, conversations, documents, and structured organizational information; proposes a qualitative causal structure; preserves assumptions and relationships for inspection; and can formalize appropriate structures as executable SD models. A separate simulation runtime executes equations and runs scenarios, while the LLM interprets outputs and supports further interrogation. The qualitative artifact can remain useful without being forced into execution.

The current implementation<sup>1</sup> provides a concrete instantiation of the architecture: natural-language elicitation, qualitative causal representation, stock-flow formalization, simulation, and human interrogation coexist in one working system.

Consider the clinic example. Conversations and documents identify arrivals, treatment capacity, clinician availability, workload, onboarding, and demand responses. The system first exposes these as a causal hypothesis that domain participants can challenge. If execution is warranted, selected quantities become stocks, flows, auxiliaries, units, equations, and initial conditions. Scenarios then examine possible trajectories under changes in capacity, hiring, productivity, or demand. The important artifact is not the chart alone; it is the bundle of scope, assumptions, structure, evidence, simulation settings, results, uncertainty, and revision history.

This example also clarifies the epistemic claim. The LLM may already know that arrivals can outpace completions and that onboarding takes time. Formalization need not create those concepts. The situated model identifies whether they matter here, connects them to local evidence, and exposes their joint consequences. If later observations contradict those consequences, the retained artifact provides a target for diagnosis and revision.

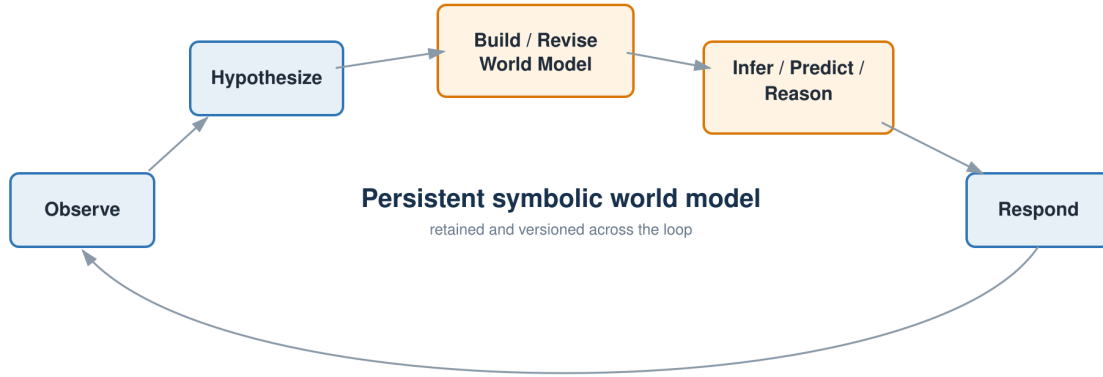
## 7 When Model Construction Becomes Learning

A symbolic world model can serve as a learning artifact in different ways. *Model formation* brings latent priors, observations, documents, data, interviews, and earlier artifacts into an explicit, situated hypothesis. The act of formation can itself be valuable: it may articulate knowledge that was previously tacit, distributed, contradictory, or difficult to use. *Model interrogation* makes that hypothesis productive by tracing implications, comparing explanations, revealing missing evidence, and, where execution semantics exist, running counterfactuals and policy experiments. *Model revision* incorporates experience by changing boundaries, relationships, assumptions, parameters, equations, qualitative expectations, or the weight assigned to competing hypotheses. *Model abstraction* compares learning across cases and compresses recurring structures into reusable knowledge.

Figure 4 shows the situated cycle in compact form. The agent observes, forms a hypothesis, builds or revises a symbolic world model, uses the model to infer, predict, or reason, responds, and observes again. Persistence is not a separate stage: it is a property of the retained artifact, which carries identity, provenance, versions, and declared commitments across the loop. A response may be an explanation, a recommendation, a question, a request for evidence, a scenario comparison, or—when the agent is authorized—an action in the world. At a minimum, model-centered learning includes evidence-responsive revision that improves prospective explanation, judgment, or action. It can also include the articulation of previously unusable knowledge and the distillation of transferable structures across models. Cross-model abstraction operates across retained models and revision histories produced by multiple such cycles.

---

<sup>1</sup>The implementation described here is commercially developed as SupaSim by Dwnrt Absrd Pte. Ltd. A public demonstration of SupaSim is available in Jagdish (2026).



**Figure 4:** A compact symbolic world-model learning loop. The agent observes, hypothesizes, builds or revises a persistent model, uses it to infer, predict, or reason, responds, and observes again.

Every symbolic world model embodies expectations about the world. Its structure expresses expectations about what exists inside the model boundary, which factors are connected, where delays occur, and which mechanisms matter. An executable model also produces behavioral expectations: patterns or trajectories that should follow if those structural commitments and assumptions are adequate. Learning can therefore begin before any simulated time series is compared with observed data. A participant may identify an omitted relationship, a new operating constraint may appear, or a causal direction may prove implausible. Behavioral evidence can also reveal that a plausible structure produces the wrong pattern. Both kinds of discrepancy are legitimate prompts for revision.

A governed revision process is deliberately easier to audit than an unrecorded prompt rewrite. When evidence and the model do not fit, the agent first asks what kind of mismatch has occurred: bad or incomplete observations, an incorrect assumption, the wrong magnitude, a missing mechanism, an inappropriate boundary, or a genuine change in the world. It then proposes explicit candidate changes and states why each might help. Where numbers are involved, parameters may be estimated through declared procedures; where structure is at issue, relationships, boundaries, or qualitative expectations may change. Candidate versions are then checked against the purpose and available evidence. If the evidence does not distinguish among them, the system retains alternatives rather than manufacturing certainty. The record shows what changed, why it changed, and why a version was retained, down-weighted, rejected, or retired.

Cross-model distillation constitutes another significant form of learning. One way to understand this process is as controlled semantic defocusing: the agent looks past domain-specific variable names and details while preserving the functional roles, causal relationships, delays, constraints, feedback structures, and behavioral expectations embodied in the model. Like squinting at a diagram, this can make similarities visible that were obscured by differences in domain vocabulary.

The degree of defocusing matters. Too little leaves the representation tied to a particular clinic, factory, service team, or maintenance operation. Too much reduces it to method-generic elements such as stocks, flows, and auxiliaries, producing similarities that are formally correct but not especially informative. The useful level often lies between them. Patients awaiting treatment, unresolved support tickets, and machines awaiting repair may all become a backlog; clinicians, support staff, and maintenance crews may become service capacity; treatment completions, resolved tickets, and completed repairs may become throughput. Domain-specific concepts are replaced by functional roles such as demand, capacity, backlog, resource acquisition, adjustment delay, attrition, congestion, or quality pressure without erasing the causal mechanism.

These defocused representations can support both comparison and retrieval. Models can be indexed by their functional and causal structure rather than only by their domain vocabulary. When a new session, observation, or document suggests an emerging structure, the agent can defocus it in the same way and retrieve relevant archetypes or prior models from other domains. A capacity–backlog–quality pattern first articulated in a clinic might therefore help interpret a service team or maintenance operation.

Cross-model abstraction also represents an attempt to improve future efficiency. By retrieving and adapting relevant archetypes, an agent may reduce the observations, interactions, and representational search required to form a useful hypothesis in a new situation. What was learned through one or more earlier modeling episodes can

thereby lower the cost of understanding a later case. This benefit is not automatic: an inappropriate archetype may accelerate convergence toward the wrong explanation. The relevant objective is therefore not speed alone, but the efficient formation of sufficiently accurate, defensible, and appropriately bounded hypotheses.

Retrieved archetypes are candidate hypotheses rather than predetermined answers. The agent must map them back to the new situation, examine whether their purposes, boundaries, mechanisms, and behavioral expectations remain meaningful, and retain alternatives when the evidence is insufficient. Each abstraction should remain connected to the situated models from which it arose and preserve the provenance, mappings, boundary conditions, counterexamples, and conditions of use associated with those models. Useful abstraction therefore requires blurring away the domain without blurring away the mechanism. In this way, model learning can operate both within a case—through articulation, interrogation, and revision—and across cases through abstraction and retrieval.

Across a fleet of agents, governed abstraction could allow useful archetypes to accrue across deployments. The shared artifact would not be a source model or a memory of a particular organization, but a candidate general theory expressed through functional roles and recurring mechanisms. It should contain neither confidential organizational detail nor personally identifiable information. Because semantic defocusing does not itself guarantee anonymity, promotion to a shared archetype library requires privacy checks, provenance controls, and explicit conditions of use.

## 8 Validation

A simulation model is not verified as true in the way a theorem is proved. Validation is better understood as a cumulative process of building trust and clarifying whether, how, and for what purpose a model should be used (Barlas and Carpenter, 1990; Barlas, 1996). The zone of validity of a model can expand, contract, or change as evidence accumulates.

For an SD model, the zone of validity can be assessed using boundary and structural assessment, dimensional consistency, extreme-condition tests, sensitivity analysis, behavior-pattern reproduction, policy robustness, parameter provenance, and review by people who know the system. Structure matters because a model that reproduces history for the wrong reasons may fail under intervention. Point prediction and conventional backtesting are therefore neither the only nor always the primary criteria; the appropriate tests depend on what the model is intended to explain or support (Sterman, 2000; Naugle et al., 2025).

Trust is not equivalent to user confidence. A polished interface or precise trajectory can induce confidence without warrant. Conversely, a model may be valuable because it lowers confidence by revealing uncertainty, disputed structure, or policy fragility. A governed artifact should expose assumptions, sources, failed tests, sensitivity, competing structures, and the conditions under which a recommendation changes. The aim is calibrated reliance: neither automatic acceptance nor reflexive rejection.

Explicitness creates a distinctive collaboration surface. Domain participants can challenge a boundary, reverse a causal link, dispute an equation, request evidence, or compare alternative structures. An LLM can explain its answer, but its internal learned representation is not directly inspectable and editable as a shared domain-level causal artifact. Symbolic externalization does not eliminate judgment; it makes more of that judgment addressable.

The modeling process itself can be consequential. Eliciting assumptions may reveal conflicting mental models; experimenting with policies may alter what participants believe is possible; and constructing a small model may build shared intuition even if the artifact is later replaced (Forrester, 1985; Ford and Sterman, 1998). Model validation must therefore attend not only to numerical fit but also to the transparency, credibility, and learning generated through use.

## 9 Evaluating Symbolic World Models

Given this account of validation, evaluation must match the epistemic structure of the claim. Controlled benchmarks are well suited to bounded questions of correctness, reliability, transfer, and efficiency. Situated qualitative assessment addresses whether open-ended models are defensible, inspectable, trusted appropriately, and useful in practice. The framework therefore distinguishes three levels: closed-world technical conformance, open-ended model construction, and situated organizational usefulness.

### 9.1 What controlled benchmarks can establish

When a task has a specified answer, quantitative testing is appropriate. An evaluator can ask whether a system translated a stated causal relationship correctly, preserved a supplied structure, produced syntactically valid

equations, maintained units, found a deliberately inserted error, or explained the behavior of a fixed model. Synthetic environments can also test narrow capabilities such as recognizing accumulation or reasoning over delay.

BEAMS provides an important example of standardized evaluation. Its tests cover causal translation, model iteration, causal reasoning, conformance, model explanation, suggested modeling steps, and error correction (Metcalf and Schoenberg, 2026). Synthetic “alternate universes” allow some tests to have deterministic ground truth. Such tasks can reveal brittleness, compare systems, and support regression testing. With repeated trials and matched resource accounting, they can also measure how frequently a system succeeds and what interactions and computational resources it requires.

Model construction need not, however, be the endpoint of a benchmark. A controlled environment can evaluate whether an agent uses its representation to answer difficult prediction, explanation, intervention, or counterfactual questions. Success includes both correct answers and appropriately calibrated uncertainty when the evidence is insufficient. Results can then be reported in terms of both success and efficiency: how reliably the agent reaches an adequate answer and how much evidence, interaction, time, and computation it requires.

These results establish bounded but important capabilities. A task with a specified target cannot by itself determine which model will be most useful for an ill-structured organizational problem. That distinction marks the scope of standardized evaluation, not a shortcoming of it.

## 9.2 Why open-ended model construction is different

Give ten competent modelers the same organizational brief and they may produce ten different models. They may select different boundaries, aggregation levels, feedback mechanisms, time horizons, or policy levers. Some differences are superficial; some structures are behaviorally equivalent under available evidence; others illuminate different legitimate purposes. Several models may be useful without being interchangeable.

A single-reference benchmark provides valuable evidence of conformance to a specified target, but it is not a complete measure of modeling quality. Even equivalence-aware graph scoring remains incomplete when alternative models answer different questions. For open-ended modeling, a further question must therefore be asked: *Is the model defensible and useful for this purpose, given this evidence, context, and uncertainty?* This does not imply that anything goes. Models can be challenged on structural plausibility, evidence, boundary adequacy, behavior, robustness, transparency, and consequences. It implies that the target is purpose-relative and may be plural.

The broader benefits are also difficult to isolate experimentally. Organizational learning can unfold over long periods. Later outcomes cannot be attributed cleanly to one modeling interaction. The model changes the conversation and potentially the system itself. User expertise, institutional context, interface design, facilitation, and the political stakes all affect what the artifact accomplishes. Attempting to control these away may remove the phenomenon under study.

## 9.3 Situated qualitative assessment

Situated qualitative assessment complements controlled benchmarks by investigating whether a model surfaces tacit assumptions, makes disagreement precise, supports counterfactual exploration, earns warranted trust, or improves a decision process. These claims cannot always be reduced to a predetermined correct answer. Rigor instead comes from a prespecified framework, transparent case selection, multiple evidence sources, attention to negative cases, explicit rival explanations, and within- and cross-case analysis (Eisenhardt, 1989; Gioia et al., 2013).

Organizational pilots provide especially valuable research settings because the questions, consequences, stakeholders, constraints, and tacit knowledge are real. Participants have reasons to challenge an inappropriate boundary, reject an implausible mechanism, or notice that an apparently useful model does not support the decision they actually face. Academic studies can assess bounded comprehension, usability, or technical operations, but they do not recreate the same stakes or situated expertise. Table 3 outlines a prospective pilot evaluation sequence. It becomes longitudinal when follow-up is repeated over time and connected to artifact versions, later observations, and consequential use. The aim is analytical rather than statistical generalization: to understand under which conditions, for what purposes, and through what mechanisms symbolic world models prove useful.

**Table 3:** A prospective pilot evaluation sequence.

| Phase           | Questions                                                                                                                   | Evidence                                                                      |
|-----------------|-----------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| Before modeling | What is the decision, initial understanding, uncertainty, disagreement, and available evidence?                             | Interviews, problem statement, initial assumptions, candidate actions         |
| During modeling | What knowledge is elicited? Which boundaries, links, and equations are challenged? What alternatives emerge?                | Observation, transcripts, artifact versions, scenario and revision logs       |
| After modeling  | Did the artifact improve clarity, inspectability, counterfactual exploration, trust calibration, or decision defensibility? | Neutral Likert-like items, interviews, decision record, unresolved questions  |
| Follow-up       | Was the model used, revised, rejected, or retired? What evidence changed its domain of warranted trust and appropriate use? | Version history, later observations, follow-up interviews, critical incidents |

Likert-like responses can structure participant perceptions but should not be converted automatically into a single usefulness score, particularly in small pilots. Responses depend on participant baselines, and ordinal categories need not have equal intervals. Participant-level interpretation, accompanying reasons, observed behavior, and artifact evidence are more informative. Across cases, recurring mechanisms and contrasting failures can gradually establish a domain of applicability.

Commercial pilots create additional obligations. Participants should know when activity is being used as research; organizational information requires consent, access control, and redaction; publication should not expose identifiable material without permission; and the developer’s commercial interest must be disclosed. Independent review, negative-case reporting, and preservation of rejected models can reduce—but not eliminate—researcher and vendor bias.

## 9.4 Matching evidence to claims

Table 4 makes the plural evaluation logic explicit. Quantitative and qualitative evidence are complementary, but they answer different questions. Evidence that supports one claim should not be treated as proof of a different claim: syntactic correctness does not establish organizational usefulness, and favorable user judgment does not establish numerical validity.

**Table 4:** Claims and the evidence appropriate to them.

| Claim                                                   | Appropriate evidence                                                                                                                            |
|---------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|
| The system can perform a specified modeling operation   | Controlled task, unit test, reliability and cost measures                                                                                       |
| Symbolic world models improve bounded agent reasoning   | Controlled tasks; answer correctness and calibration; success across repeated trials; and required evidence, interaction, time, and computation |
| A particular executable artifact is internally coherent | Compilation, units, conservation, extreme-condition and numerical tests                                                                         |
| A particular model is defensible for a declared purpose | Structural review, evidence provenance, behavior-pattern tests, sensitivity and policy robustness                                               |
| People can inspect and challenge the artifact           | Observed use, interviews, revision records, and comprehension evidence                                                                          |
| The artifact improves a decision process                | Situated evidence collected before, during, and after the case, with explicit rival explanations                                                |
| The model learns through interaction                    | Longitudinal records of predictions, observations, discrepancies, and versions                                                                  |
| The approach transfers across contexts                  | Replicated cases, cross-case comparison, and identified boundary conditions                                                                     |

Quantitative benchmarks are therefore indispensable when the evaluator can specify the task and judge its outcome. They can establish whether symbolic world models help agents reason more correctly, reliably, or efficiently under controlled conditions. Open-ended modeling introduces additional questions because the appropriate representation—and sometimes the appropriate question—is not known in advance. Those questions require situated evidence of

how models are constructed, challenged, trusted, and used.

## 10 Discussion

The philosophical premise of this paper is simple but consequential: recorded knowledge is not identical to the world. A foundation model inherits patterns from traces of prior observation and articulation. Those traces support impressive generalization, yet omit what was inaccessible, tacit, changing, or undiscovered. Retrieval can expose additional records, but it cannot retrieve what was never articulated, and a collection of facts need not form a coherent causal-temporal account.

Symbolic world models offer a bridge between inherited knowledge and continued experience. They do not replace latent neural representations. In many cases they externalize or reorganize what the LLM already knows. Their distinctive value is that they are situated, explicit, inspectable, contestable, persistent, revisable, and shareable; when executable, they also support reproducible counterfactual exploration. These properties turn assumptions about a particular part of the world into artifacts that can participate in collective reasoning and continued learning.

Memory persists observations. Skills persist ways of acting. Symbolic world models can persist hypotheses about structure. Conceptual artifacts make those hypotheses communicable and revisable. Executable artifacts add a generative affordance: an agent can ask *what if* before acting in a costly system. This is synthetic experience generated by a fallible theory, not experience supplied by the world. That distinction is central to responsible use.

The argument also suggests an adaptive representational architecture. An agent should be able to stop at a qualitative map when that is sufficient, choose SD for aggregate feedback systems, use discrete-event simulation for queues and processes, select agent-based models for heterogeneous interaction, or construct other formal artifacts for physical, algorithmic, or probabilistic domains. Choosing the right formalism is one component of a more consequential longer-term question: can agents become more useful in the world by selecting, constructing, validating, and revising explicit models suited to the situations they encounter? And more broadly, can learning accumulated and distilled across vast sets of such generated models be a source of novel articulated knowledge.

Repeated modeling episodes may therefore create a functionally indexed library of experience. Controlled semantic defocusing can make recurring mechanisms visible across domains, while retrieval can bring those mechanisms back into reasoning when a new situation is encountered. This connects situated model construction to experience compression and reusable knowledge. The central risk is premature analogy: a familiar archetype may illuminate a new setting or force it into the wrong frame. Archetypes must therefore remain conditional, contestable, and connected to the provenance, boundaries, counterexamples, and alternatives that govern their appropriate use.

Finally, the paper advances a plural methodological position. Closed-world benchmarks provide standardized evidence of bounded correctness, reliability, transfer, and efficiency. Situated assessment examines complementary questions about how models expose assumptions, support inquiry, earn or lose trust, and contribute to consequential decisions. Together, these methods allow technical capability and practical usefulness to be studied without treating either as a substitute for the other.

## 11 Limitations and Risks

The following risks bound the claims of the paper and identify where governance and further research are most important.

**Epistemic and causal limits.** No neural or symbolic representation exhausts the part of the world it addresses. Multiple structures can reproduce the same history, and explicit construction cannot identify a mechanism without discriminating evidence. A generated diagram or equation set remains a hypothesis; intervention effects require appropriate causal and domain evidence. Cross-model abstraction can also mistake superficial structural resemblance for a transferable causal mechanism, especially when contextual boundaries and counterexamples have been blurred away.

**Representational limits.** Qualitative causal maps and SD privilege causal structure, while SD additionally privileges aggregate stocks, flows, and feedback. These choices can obscure heterogeneity, discrete events, strategic interaction, spatial structure, and rare transitions. Constraint reduces search but can exclude the right explanation.

**False precision and trust.** LLMs can invent variables, reverse polarities, misuse units, omit constraints, and produce executable equations that encode an incoherent theory. Execution may increase unwarranted confidence.

Validation builds warranted trust for a purpose; it cannot prove that a model is complete or permanently adequate.

**Revision and computational limits.** Plausible structures grow combinatorially with variables, boundaries, and parameters. Premature pruning creates false certainty, while retaining and simulating every alternative is infeasible. Model maintenance also adds interaction, review, latency, and compute cost.

**Evaluation limits.** Situated qualitative assessment can reveal mechanisms and usefulness that benchmarks miss, but it introduces its own risks: participant self-report, facilitator influence, selection bias, demand characteristics, incomparable contexts, and weak longitudinal attribution. Pilot evidence should not be generalized statistically beyond the cases studied. Cross-case replication and transparent negative cases are necessary for broader claims.

**Commercial interests in field evaluation.** Pilots are valuable because they involve real stakes and stakeholders, but commercial relationships can influence case selection, participant responses, facilitation, interpretation, and the reporting of successes or failures. Transparent disclosure, preservation of negative cases, and independent review where practical help make those incentives visible; they do not eliminate them.

**Human responsibility and governance.** Elicitation, boundary selection, validation, interpretation, and accountability remain consequential human roles. A responsible outcome may be to reject a model, retain several alternatives, seek more evidence, restrict its use, or decline to model the problem at all. Explicit artifacts can support responsibility, but they do not transfer it from people to the system.

## 12 Conclusion

LLMs inherit broad world-model-like knowledge from recorded traces of reality, but those traces are neither reality in full nor the limit of what can be learned. Agent systems increasingly supplement this inheritance through memory and reusable skills. This paper has examined symbolic world models as another type of persistent artifact: explicit, bounded, and revisable hypotheses about how parts of the world may work.

Conceptual models make causal commitments inspectable. Executable models additionally make selected consequences reproducible. System dynamics offers one deliberately constrained language for this progression in aggregate organizational systems. The larger proposal is representation-agnostic: agents may benefit from constructing, selecting, interrogating, validating, and revising explicit models suited to the problems they encounter.

Evaluation likewise requires complementary methods. Controlled tests can establish whether symbolic world models help agents reason correctly, reliably, and efficiently on bounded tasks. Particular models earn trust through purpose-appropriate validation, while organizational usefulness must be studied through observed use, participant judgment, artifact histories, negative cases, and longitudinal evidence. Together, these forms of inquiry can reveal not simply whether symbolic world models work, but where, why, for whom, and under what conditions they become useful learning artifacts.

## Disclosure

The author is the developer of SupaSim and has a commercial interest in Dwnrt Absrd Pte. Ltd.

## References

- Don Joven Agravante, Daiki Kimura, Michiaki Tatsubori, Asim Munawar, and Alexander Gray. Learning neuro-symbolic world models with conversational proprioception. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 648–656, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.57. URL <https://aclanthology.org/2023.acl-short.57/>.
- Yaman Barlas. Formal aspects of model validity and validation in system dynamics. *System Dynamics Review*, 12(3):183–210, 1996. doi: 10.1002/(SICI)1099-1727(199623)12:3<183::AID-SDR103>3.0.CO;2-4.
- Yaman Barlas and Stanley Carpenter. Philosophical roots of model validation: Two paradigms. *System Dynamics Review*, 6(2):148–166, 1990. doi: 10.1002/sdr.4260060203.

- Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463.
- Pengfei Du. Memory for autonomous LLM agents: Mechanisms, evaluation, and emerging frontiers. *arXiv preprint arXiv:2603.07670*, 2026.
- Kathleen M. Eisenhardt. Building theories from case study research. *Academy of Management Review*, 14(4): 532–550, 1989. doi: 10.5465/amr.1989.4308385.
- Runnan Fang et al. MemP: Exploring agent procedural memory. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. arXiv:2508.06433.
- David N. Ford and John D. Sterman. Expert knowledge elicitation to improve formal and mental models. *System Dynamics Review*, 14(4):309–340, 1998. doi: 10.1002/(SICI)1099-1727(199824)14:4<309::AID-SDR154>3.0.CO;2-5.
- Jay W. Forrester. *Industrial Dynamics*. MIT Press, Cambridge, MA, 1961.
- Jay W. Forrester. “The” Model versus a Modeling “Process”. *System Dynamics Review*, 1(1):133–134, 1985. doi: 10.1002/sdr.4260010112.
- Erika Frydenlund, Joseph Martínez, Jose J. Padilla, Katherine Palacio, and David Shuttleworth. Modeler in a box: How can large language models aid in the simulation modeling process? *SIMULATION*, 100(7):727–749, 2024. doi: 10.1177/00375497241239360.
- Huan-ang Gao et al. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *Transactions on Machine Learning Research*, 2026.
- Philippe J. Giabbanelli, John Beverley, Istvan David, and Andreas Tolk. From over-reliance to smart integration: Using large-language models as translators between specialized modeling and simulation tools. *arXiv preprint arXiv:2506.11141*, 2025.
- Dennis A. Gioia, Kevin G. Corley, and Aimee L. Hamilton. Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods*, 16(1):15–31, 2013. doi: 10.1177/1094428112452151.
- John Gkountouras, Matthias Lindemann, Phillip Lippe, Efstratios Gavves, and Ivan Titov. Language agents meet causality: Bridging LLMs and causal world models. *arXiv preprint arXiv:2410.19923*, 2024.
- David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.
- Danijar Hafner et al. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.
- Mengkang Hu, Bowei Xia, Yuran Wu, Ailing Yu, Yude Zou, Qiguang Chen, Shijian Wang, Jiarui Jin, Kexin Li, Wenxiang Jiao, Yuan Lu, and Ping Luo. Agent2World: Learning to generate symbolic world models via adaptive multi-agent feedback. *arXiv preprint arXiv:2512.22336*, 2025a.
- Mengkang Hu et al. Text2World: Benchmarking large language models for symbolic world model generation. *arXiv preprint arXiv:2502.13092*, 2025b.
- Yuyang Hu et al. Memory in the age of AI agents. *arXiv preprint arXiv:2512.13564*, 2025c.
- Ninad Jagdish. SupaSim V2. YouTube video, 2026. URL <https://www.youtube.com/watch?v=CD1C5hYBLuI>. Accessed 2026-08-22.
- Zaid Khan, Archiki Prasad, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. One life to learn: Inferring symbolic world models for stochastic environments from unguided exploration. *arXiv preprint arXiv:2510.12088*, 2025.
- Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

- Ning-Yuan Georgia Liu and David R. Keith. Leveraging large language models for automated causal loop diagram generation: Enhancing system dynamics modeling through curated prompting techniques. *arXiv preprint arXiv:2503.21798*, 2025.
- Xiao Luo et al. How do large language models perform in dynamical system modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 866–880, 2025. doi: 10.18653/v1/2025.findings-naacl.50. URL <https://aclanthology.org/2025.findings-naacl.50/>.
- Yijun Ma, Zehong Wang, Yiyang Li, Ziming Li, Xiaoguang Guo, Weixiang Sun, Chuxu Zhang, and Yanfang Ye. ProPlay: Procedural world models for self-evolving LLM agents. *arXiv preprint arXiv:2606.12780*, 2026.
- Sara Metcalf and William Schoenberg. BEAMS: Benchmarking and evaluating AI for modeling and simulation. *arXiv preprint arXiv:2605.28994*, 2026. doi: 10.48550/arXiv.2605.28994.
- Asmeret B. Naugle, Indu Manickam, Scott Steinmetz, Paul Schutte, Matt Sweitzer, and Alex Washburne. Trusted simulation: Considering model quality in the context of user trust. *System Dynamics Review*, 41(4):e70011, 2025. doi: 10.1002/sdr.70011.
- Ikujiro Nonaka. A dynamic theory of organizational knowledge creation. *Organization Science*, 5(1):14–37, 1994. doi: 10.1287/orsc.5.1.14.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763.
- Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.
- Michael Polanyi. *The Tacit Dimension*. University of Chicago Press, Chicago, 2009. Originally published 1966.
- Julian Schrittwieser et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588: 604–609, 2020. doi: 10.1038/s41586-020-03051-4.
- Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.
- John D. Sterman. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Irwin/McGraw-Hill, Boston, MA, 2000.
- Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.
- Hao Tang, Darren Key, and Kevin Ellis. WorldCoder, a model-based LLM agent: Building world models by writing code and interacting with the environment. In *Advances in Neural Information Processing Systems*, volume 37, pages 70148–70212, 2024. doi: 10.52202/079017-2243. arXiv:2402.12275.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.
- Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024.