Street Art & Urban Creativity | Vol. 12, No. 5, 2026 | ISSN 2183-9956
Street Art & Urban Creativity. Scientific Journal

AUDITABLE CLOSED-LOOP AI AGENTS FOR URBAN BRIDGE DIGITAL TWINS

Synthetic Validation Followed by a Municipal-Scale Utsunomiya Implementation

Takayuki Shinohara[1]
[1] Integrated Research Center for Resilient Infrastructure, National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Higashi, Tsukuba, Ibaraki 305-8567, Japan
** Corresponding author: shinohara.takayuki@aist.go.jp*

KEYWORDS

Agentic AI
Urban digital twins
Smart cities
Bridge maintenance
Urban resilience
Multiobjective planning
Auditable AI

ABSTRACT

Urban bridge maintenance is a smart-city governance problem involving uncertain evidence, constrained budgets, and safety-critical actions. We present an auditable closed-loop artificial intelligence architecture separating observations, inferred states, alternatives, decisions, and interventions. Mechanisms are validated in SynthTown, a ground-truth municipality with 50 bridges and 842 components, using hidden-state prediction, multiobjective planning, stakeholder screening, feedback recalibration, and verifier-repair safeguards. A public-data implementation covers 1,733 Utsunomiya bridges and integrates 1,592 persistent scatterer interferometric synthetic aperture radar points, inspection records, three-dimensional maintenance models, and capacity-aware plans. Verification eliminated unsafe outputs in the synthetic benchmark.

RECEIVED: 07 / 21 / 2026
ACCEPTED: 31 / 08 / 2026

1. Introduction

Urban bridge networks support daily mobility, emergency access, logistics, and continuity of public services. Their maintenance is therefore not only an engineering problem but also a smart-city governance problem involving urban resilience, public resource allocation, and accountable use of artificial intelligence (AI) (Meerow et al., 2016; Pereira et al., 2018; Wirtz et al., 2019). Bridge maintenance is a sequential public decision problem rather than a single prediction task (Frangopol et al., 2001; Frangopol & Liu, 2007; Wu et al., 2021). An agency observes incomplete and noisy evidence, infers latent condition, compares plans under budget and network constraints, authorizes interventions, and later receives evidence altered by those interventions. A system that predicts deterioration but cannot preserve evidence, enforce authority boundaries, or learn from executed actions is incomplete as a public decision-support system.

Within smart-city research, urban digital twins are increasingly framed as decision and collaboration infrastructures rather than only three-dimensional representations. They combine heterogeneous city data, simulation, and stakeholder interaction to support transparent planning and collective interpretation (Dembski et al., 2020). For urban bridge networks, this framing requires explicit evidence provenance, uncertainty-aware state estimation, multiobjective planning, and human authority over interventions that affect mobility, safety, and public expenditure.

Digital twins are often proposed as the integration layer for such processes. Yet many infrastructure twins remain oriented toward representation, monitoring, or prediction. They may synchronize sensor streams, inspection records, and geometric models, but they do not necessarily specify how uncertain observations become state beliefs, how beliefs become feasible maintenance plans, how stakeholder disagreement changes the candidate set, how approval authority is separated from automated recommendation, or how completed interventions are fed back into model calibration. The missing element is not another dashboard. It is an executable evidence-to-action loop with explicit safety and audit boundaries.

The rapid development of language-model agents increases both the opportunity and the risk, as shown by increasingly realistic evaluations of autonomous, software-engineering, assistant, and tool-use behavior (Zhou et al., 2023; Jimenez et al., 2023; Mialon et al., 2023; Ma et al., 2024; Yao et al., 2024). An agent can retrieve records, summarize evidence, call predictive tools, and propose plans through a natural-language interface; retrieval-augmented generation can ground such responses in external records, but hallucination remains a documented failure mode (Lewis et al., 2020; Ji et al., 2023). However, an unsupported condition claim, an identifier mix-up, or an unauthorized approval action is materially different from an ordinary text-generation error. In infrastructure management, an agent may be useful only if it can cite the evidence that supports each claim, abstain when the requested asset is absent, remain within a restricted tool boundary, and route approval to a human authority. These requirements motivate an architecture in which generative behavior is surrounded by deterministic verification, repair, and permission controls rather than trusted as a monolithic decision maker.

This article uses a two-tier validation design. Mechanism validity is tested in a synthetic municipality named SynthTown, which contains 50 bridges, 842 components, a road network, seven deterioration mechanisms, condition reports with exact evidence spans, and interferometric synthetic aperture radar (InSAR)-like displacement series with injected anomalies. Because latent condition, bridge-specific frailty, intervention response, and anomaly onset are known, the complete loop can be evaluated rather than only its final recommendation. The controlled design follows the broader principle that benchmark validity depends on separating development feedback from claims of external generalization and on documenting the generated data and their intended use (Dwork et al., 2015; Gebru et al., 2021). Transfer to a relevant public-infrastructure environment is then tested through a municipal-scale pilot using records covering Utsunomiya City. The pilot does not claim live administrative deployment; it tests whether the same evidence, provenance, trust-gating, prioritization, and planning interfaces execute on heterogeneous real-world data.

Five research questions organize the evaluation. RQ1 asks whether the complete system can transform documents and sensor observations into a traceable decision and then ingest post-intervention evidence. RQ2 asks whether hidden-state estimation with bridge-level partial pooling improves calibration for sparsely observed assets. RQ3 asks whether constrained multiobjective optimization followed by evidence-limited stakeholder screening and local refinement produces a more acceptable plan without violating the human approval boundary. RQ4 asks whether deterministic verification and repair reduce unsafe agent outputs while retaining useful task coverage under deliberately injected faults. RQ5 asks whether the evidence and planning interfaces can be instantiated at municipal scale on Utsunomiya data while preserving provenance, observation-trust guidance, capacity constraints, and human decision authority.

The study makes five contributions. First, it specifies and implements a closed-loop digital-twin architecture that separates observations, inferred states, proposed alternatives, approved decisions, and interventions in an append-only provenance store. Second, it combines an observation-error-aware Markov deterioration model with bridge-specific partial pooling and intervention-conditioned recalibration. Third, it connects a four-objective constrained optimizer to an evidence-limited stakeholder council and a disagreement-driven refinement procedure. Fourth, it introduces an agent harness represented as C-A-P-V-R: a claim-producing agent C, restricted tool access A, workflow template P, deterministic verifier V, and repair operator R. Fifth, it reports a municipal-scale Utsunomiya pilot that instantiates the real-data evidence and planning interfaces on 1,733 bridges, thereby separating synthetic mechanism validation from relevant-environment prototype validation.

The central result is not that one synthetic model achieved perfect extraction or anomaly detection. Those outcomes partly reflect the controlled generator and templated reports. The substantive result is that the layers can be composed without erasing uncertainty or authority boundaries and that the data contracts can be instantiated at municipal scale. In the archived seed-42 run, the predictive model achieved an expected calibration error of 0.024 before planning; the optimizer produced 80 feasible Pareto alternatives; council refinement increased minimum acceptance from 0.294 to 0.528 while reducing disagreement from 0.238 to 0.052; the final decision could be traced through six stored nodes to an exact document span; and verification reduced the agent unsafe rate from 0.278 to zero. In Utsunomiya, the prototype processed 1,733 bridge records and 1,592 persistent scatterer InSAR (PS-InSAR) points, exposed 500 evidence-grounded priority cards, and converted the ranking into 108 capacity-selected inspection or patrol items and a 30-year maintenance schedule.

2. Related Work

2.1. Digital twins, synchronization, and provenance

Urban digital twin research connects data integration and simulation with participation and governance. Dembski et al. (2020) demonstrated a city-scale twin for collaborative planning, while Pereira et al. (2018) showed that smart-city value depends on governance arrangements, stakeholder participation, and public accountability. Urban resilience concerns the capacity of a city to maintain and adapt critical functions under disturbance (Meerow et al., 2016). These perspectives move infrastructure twins beyond technical synchronization toward auditable public decision support.

Digital-twin research distinguishes a static digital model from systems in which data move between the physical and digital entities. Kritzinger et al. (2018) described a progression from digital model to digital shadow and digital twin according to the direction and automation of data exchange. In construction and infrastructure, the concept has expanded beyond geometry toward semantic integration, sensing, simulation, and operational control. Boje et al. (2020) argued that construction digital twins require semantic completeness across sensor, process, and social systems rather than a building-information model alone. Su et al. (2023) similarly emphasized lifecycle data integration and the role of a twin as a framework for monitoring, analysis, prediction, and decision support.

For maintenance decisions, synchronization is necessary but insufficient. A public agency must also distinguish when an event occurred from when it was recorded, preserve the quality and origin of an observation, and reconstruct how a recommendation was produced. The W3C PROV model formalizes entities, activities, and agents for interoperable provenance (Moreau & Missier, 2013). The present study adopts the same general principle but implements it at the level of executable infrastructure decisions: a decision refers to an alternative; the alternative refers to an objective evaluation; the evaluation refers to a scenario set; and the scenario set refers to state observations with evidence spans. This chain is tested as part of the experiment rather than treated as metadata added after computation.

The literature therefore motivates two design requirements. Transparent documentation of dataset provenance, collection assumptions, and intended uses is also necessary when synthetic and municipal data are combined (Gebru et al., 2021). First, a twin should preserve the difference between observation and inferred state. An inspection grade or sensor statistic is evidence, not the physical condition itself. Second, synchronization should be bidirectional in an operational sense: completed interventions and subsequent inspections must update the predictive model and future plans. The proposed loop operationalizes both requirements in a synthetic environment.

2.2. Deterioration models under observation uncertainty

Network bridge-management systems established condition-state planning as an operational paradigm (Thompson et al., 1998). Markov-chain models are widely used in bridge management because they scale to networks and represent condition transitions probabilistically, although reviews continue to identify limitations in homogeneity, memorylessness, observation quality, and transfer across bridge populations (Morcous, 2006; Torres-Machi et al., 2020). Morcous (2006) examined the assumptions behind bridge-deck Markov models, including inspection intervals and state independence. Life-cycle frameworks subsequently linked stochastic performance prediction with reliability, inspection, maintenance, network scheduling, and cost decisions (Frangopol et al., 2001; Frangopol & Liu, 2007; Bocchini & Frangopol, 2011; Frangopol et al., 2017). History-dependent Markov decision models further showed that the value of an intervention depends on prior condition and action trajectories (Robelin & Madanat, 2007). However, the reported inspection state is not necessarily the true state. Madanat and Ben-Akiva (1994) formulated a latent Markov decision process that explicitly recognized condition-measurement error, showing that inspection and repair policies can change when the observation process is modeled rather than assumed perfect. Recent work has also coupled large-language-model extraction of Japanese bridge inspection reports with mechanism-aware Markov modelling for large-scale predictive maintenance, demonstrating how document-derived evidence can be converted into interpretable deterioration and life-cycle planning inputs (Shinohara et al., 2025).

Partially observable Markov decision process (POMDP) formulations extend this principle by maintaining a belief distribution over condition states and valuing information gained through inspection (Howard, 1966; Kaelbling et al., 1998). Papakonstantinou and Shinozuka (2014a, 2014b) reviewed and implemented dynamic-programming and POMDP approaches for structural inspection and maintenance. These methods provide a rigorous foundation but can be difficult to scale or explain when the state space, component mechanisms, network constraints, and local heterogeneity are large. Recent constrained and deep-reinforcement-learning formulations address larger inspection and maintenance problems, but they retain substantial requirements for model validation, constraint specification, and interpretability (Andriotis & Papakonstantinou, 2021; Lei et al., 2022). The present implementation uses a reduced four-state hidden Markov formulation, explicit observation confusion, intervention recovery kernels, and bridge-level random effects estimated by maximum a posteriori partial pooling. It does not claim to solve the general POMDP. Its role is to supply calibrated scenario distributions to a separate constrained planner.

Calibration is critical because a decision model consumes probabilities rather than labels. Expected calibration error (ECE) summarizes the discrepancy between predicted event

probabilities and empirical frequencies across probability bins; it has been widely used to diagnose overconfidence in modern prediction systems (Guo et al., 2017). In this work, the event is a subsequently reported grade of III or IV, and calibration is assessed both before planning and after intervention-conditioned feedback. This avoids evaluating a repaired asset as though no repair had occurred.

2.3. Multiobjective maintenance planning and stakeholder trade-offs

Bridge maintenance portfolios involve conflicting objectives: life-cycle cost, residual safety risk, traffic disruption, annual budget stability, and sometimes environmental or equity effects. Multiobjective evolutionary algorithms are useful when the feasible set is discrete and the objective functions are nonlinear. The Non-dominated Sorting Genetic Algorithm II (NSGA-II) combines nondominated sorting, elitism, and crowding-distance diversity and remains a standard baseline for such problems (Deb et al., 2002). Infrastructure life-cycle studies have used related methods to expose rather than suppress the trade-off between cost and performance, including budget-constrained bridge management, network renewal, evolutionary planning, and simulation-optimization (Halfawy et al., 2009; Orcesi & Frangopol, 2010; Frangopol et al., 2017; Pastore et al., 2024).

A Pareto set does not by itself make a public decision. The selected plan must respect hard constraints and survive evaluation by stakeholders with different loss functions. A budget officer may prioritize annual expenditure stability; an emergency-management role may reject a plan with excessive residual risk; residents may emphasize disruption; and contractors may be concerned with feasible workload. Conventional weighted sums can conceal low acceptance by one role and are sensitive to arbitrary scale. The implemented council instead ranks plans lexicographically by veto count, minimum acceptance, mean acceptance, and disagreement. The design is related to robust and max-min decision principles, but it is deliberately transparent and deterministic: each role sees only a structured PlanSummary packet, and every vote cites the stored objective evaluation and alternative. A recent bridge-maintenance study similarly combined auditable evidence with multi-stakeholder AI evaluation, providing a domain-specific precedent for treating stakeholder review as an explicit decision layer rather than a post hoc narrative (Shinohara et al., 2026).

Stakeholder screening also creates a search signal. If roles prefer different plans, the contested pair defines an objective-space neighborhood for local refinement. The method therefore alternates optimization and deliberation rather than treating stakeholder review as a final narrative layer. The experiments test whether this loop improves acceptability and whether further refinement can become non-monotonic, which would require retaining the best earlier round.

2.4. Language-model agents, tool use, and safety boundaries

Language-model agents interleave reasoning with actions, retrieval, and tool calls. ReAct demonstrated that reasoning traces and environment actions can support each other in interactive tasks (Yao et al., 2023), while Reflexion showed that verbal feedback can improve subsequent attempts (Shinn et al., 2023). AgentBench broadened evaluation across multiple interactive environments (Liu et al., 2024). Subsequent benchmarks have emphasized realistic websites, repository-level software tasks, general-assistant questions, multi-turn trajectories, tool selection, tool reliability, and stateful user-tool interaction (Zhou et al., 2023; Jimenez et al., 2023; Mialon et al., 2023; Ma et al., 2024; Guo et al., 2024; Lu et al., 2024; Yao et al., 2024). These studies establish capability, but generic success metrics do not fully capture infrastructure risk. A maintenance agent can be locally fluent while citing the wrong component, asserting a nonexistent bridge, or claiming to have approved a plan beyond its authority.

Recent work has therefore evaluated agents under tool and safety risks. ToolEmu used language-model-emulated sandboxes to expose potentially severe failures across high-stakes toolkits (Ruan et al., 2024). Agent-SafetyBench and AgentDojo further evaluate harmful tool behavior and prompt-injection attacks, reinforcing the need to test both task completion and unsafe side effects (Z. Zhang et al., 2024; Debenedetti et al., 2024). Human-AI interaction guidelines also emphasize communicating uncertainty, supporting correction, and preserving

appropriate human control (Amershi et al., 2019). Selective prediction formalizes a related trade-off: a system can abstain on uncertain cases to reduce error on the cases it does answer (Geifman & El-Yaniv, 2017). Work on verbalized uncertainty similarly shows that confidence communication must be explicitly trained and evaluated rather than inferred from fluent output (Lin et al., 2022).

The architecture in this study treats abstention as a valid operational outcome, but not as the sole safeguard. The agent must emit structured claims with citations. The verifier checks subject existence, citation existence, and asset hierarchy. The repair step can re-query a correct evidence node for narrowly defined errors or drop an unsupported claim. The tool layer denies approval. These controls are tested separately so that the contribution of verification and repair can be measured rather than hidden inside an end-to-end prompt.

For public-sector AI, predictive accuracy alone is insufficient because administrative legitimacy also depends on transparency, discretion boundaries, contestability, and accountability (Wirtz et al., 2019). The proposed architecture operationalizes these concerns through typed evidence contracts, restricted tools, abstention, deterministic verification, repair, and traceable human approval.

3. Method

3.1. Closed-loop architecture and entity separation

Figure 1 summarizes the implemented architecture. The loop begins with synthetic inspection reports and InSAR-like displacement series. Evidence-bound extraction produces typed observations. A hidden-state model converts observations into component-state beliefs and future scenarios. A constrained optimizer produces a Pareto set of maintenance plans. An evidence-limited council scores selected alternatives; disagreement triggers local refinement. A human approval gate converts one alternative into a decision. Executed interventions and later inspection outcomes are ingested and used to recalibrate the model. Across the loop, the agent interface can retrieve evidence and propose actions, but verification, repair, and tool permissions restrict what can become an accepted output.

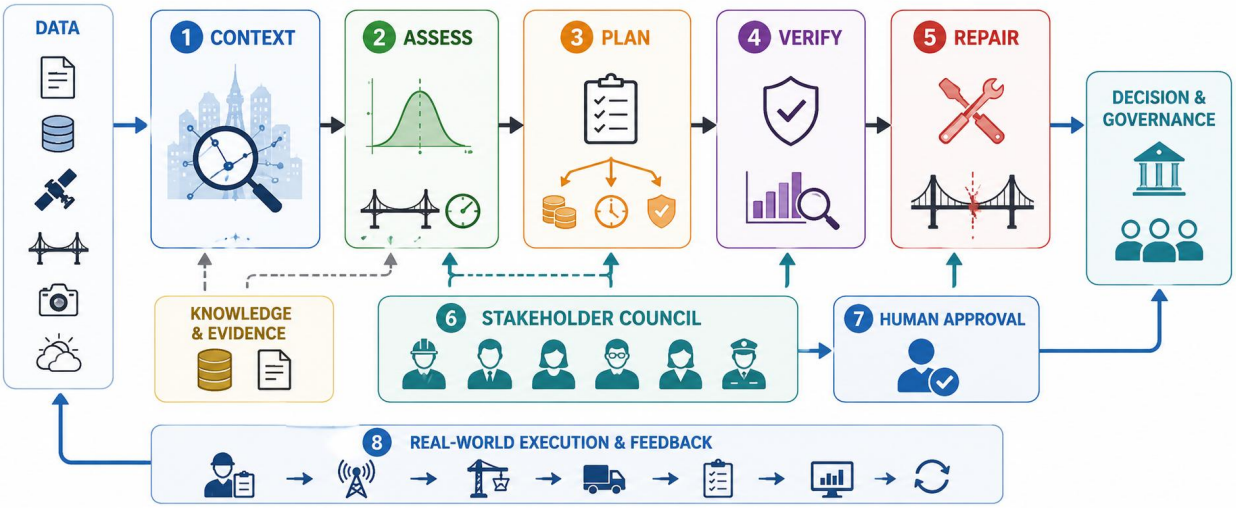

Figure 1. Implemented closed-loop architecture from evidence ingestion to intervention feedback.
Source: Authors' own elaboration, 2026.

The store separates six decision-relevant entity classes. Assets and components define the bridge hierarchy. Observations record reported condition grades or sensor-derived evidence with occurrence year, recording year, confidence, source, and exact evidence pointer. Interventions describe inspections, preventive repairs, corrective repairs, or rebuild actions. Scenario sets preserve the predictive assumptions used by the optimizer. Alternatives and objective evaluations preserve the candidate plan and its quantified consequences. Stakeholder votes and the final Decision preserve the deliberative and authorization stages. Writes pass schema and evidence gates, while the store remains append-only for the tested workflow.

This separation prevents three category errors. A report grade cannot silently overwrite the latent state. A recommendation cannot be confused with an approved intervention. A post-repair observation cannot be interpreted without the intervention that changed the transition path. The store also supports bitemporal queries: occurred_year represents when the condition or action belonged to the physical process, whereas recorded_year represents when it became available to the decision system. All as-of evaluations use recorded time to avoid future information leakage.

3.2. SynthTown and synthetic evidence generation

SynthTown is a controlled municipal bridge network generated with seed 42. Fifty bridges are placed on a grid road network with named arterial and local corridors. Each bridge has structural type, number of spans, construction year, traffic, heavy-vehicle volume, coastal distance, deicing exposure, river context, detour cost, and a latent bridge-specific deterioration offset. The generator creates 842 components, including girders, decks, bearings, piers, and abutments. Component deterioration is assigned to one of seven mechanisms: neutralization, chloride, fatigue, corrosion, alkali-silica reaction, frost, or scour. The mechanism assignment depends on material, component type, coastal exposure, deicing, river context, and bridge attributes.

The state space contains four ordered condition grades, I through IV. Each component starts in state I and can remain in the current state or deteriorate by one grade per year. Synthetic inspections do not reveal the state perfectly. A report confusion matrix includes a conservative tendency to report one grade worse with probability 0.10, while state IV remains IV. The environment also generates interventions and applies action-specific recovery kernels. This makes it possible to test whether the estimator models observation error and intervention effects rather than simply counting reported transitions.

Inspection reports are generated as text with component identifiers, condition phrases, damage terms, inspection year, and exact character spans. Historical evidence through 2025 is used to estimate the model and construct the decision. A second period, 2026-2030, records outcomes after selected interventions. A subset of bridges is intentionally sparse, with only two inspections, to test partial pooling. InSAR-like series include annual seasonality, trend, noise, and seven injected anomaly onsets. All latent states, bridge effects, anomaly labels, interventions, and evidence spans are retained in a truth registry for evaluation.

Table 1. SynthTown experimental design and generated evidence.

Element	Value	Purpose
Municipal assets	50 bridges; 842 components	Network-scale closed-loop evaluation
Condition system	Four latent grades (I-IV)	Partially observed deterioration
Mechanisms	7	Mechanism-specific transition baselines
Evidence period	Historical through 2025	Model fitting and decision cycle
Feedback period	2026-2030	Intervention-conditioned recalibration
Reports ingested	269 historical; 50 feedback	Evidence extraction and loop closure
InSAR anomalies	7 injected onsets	Detection and false-alarm evaluation
Agent benchmark	18 tasks; fault rate 0.30	Verifier-repair ablation

Source: Authors' analysis of SynthTown experiments, 2026.

3.3. Evidence-bound report extraction and InSAR anomaly detection

The report adapter is deterministic by design. It uses a template dictionary and regular expressions to identify a component identifier, grade phrase, and optional damage phrase. An Observation cannot be written without an Evidence object containing document identifier, character span, and source text. This makes unsupported extraction structurally invalid. The deterministic parser is not proposed as a general solution to real heterogeneous reports; it is a contract test for the evidence interface used by later language-model adapters. Prior work has

addressed semantic relation extraction, inspection-image parsing, inspector behavior, and hybrid language-model construction of bridge-inspection databases, indicating the heterogeneity that a field deployment must handle (K. Liu & El-Gohary, 2021; P. Liu et al., 2022; C. Zhang et al., 2022; C. Zhang et al., 2024).

The InSAR processing chain first regresses annual sine and cosine terms to remove seasonal deformation. It then computes a cumulative-sum (CUSUM) statistic standardized on the first third of the series (Page, 1954). Significance is estimated by 500 temporal permutations, applying the identical statistic to the permuted series. Benjamini-Hochberg (BH) false-discovery-rate control is applied at q = 0.05 across points (Benjamini & Hochberg, 1995). For flagged series, a hinge regression estimates anomaly onset. Matching the observed statistic and permutation statistic is important; an earlier mismatched specification produced false positives in development and was rejected.

3.4. Hidden-state deterioration prediction and partial pooling

Let S_bcm,t be the latent grade of component c on bridge b under mechanism m in year t. The annual transition matrix is monotone and upper-bidiagonal. For states i in {I, II, III}, the component remains in i with probability 1 - p_bmi and moves to i + 1 with probability p_bmi. State IV is absorbing unless an intervention recovery kernel is applied. The transition probability is parameterized as follows:

logit(p_bmi) = a_mi + beta^T z_b + delta_b,

where a_mi is the mechanism- and grade-specific baseline, z_b contains centered log heavy-vehicle AADT, exp(-coastal_distance/2), and a deicing indicator, beta is the common covariate coefficient vector, and delta_b is a bridge-specific offset. The model therefore captures both shared environmental effects and local heterogeneity.

The observation model C(y | s) maps latent grade s to reported grade y. The likelihood is evaluated with a forward algorithm over the latent state sequence and includes recovery kernels in intervention years. Shared baselines and beta are estimated by hidden-state maximum likelihood. Bridge offsets are then estimated by maximum a posteriori (MAP) optimization with delta_b distributed as N(0, 0.4^2). The prior shrinks weakly observed bridges toward the population while allowing strongly observed bridges to deviate. For ablation, pooled prediction fixes delta_b = 0, and unpooled prediction estimates bridge offsets without shrinkage.

State inference is a yearly Bayesian filter. The order of operations is intervention, deterioration, and then observation update. This order reflects the synthetic data-generating process and prevents observations after a repair from being attributed to unconditioned deterioration. Future trajectories propagate the posterior through the intervention-conditioned transition matrix. The optimizer receives scenario samples and bridge-level dominant-component beliefs rather than point labels.

Predictive quality is evaluated on the next reported grade. Mean absolute error is computed on the expected reported grade. ECE uses ten probability bins for the event that the next report is grade III or IV. Because this metric evaluates a reported outcome, the observation confusion matrix remains in the predictive distribution. A separate parameter-recovery evaluation compares fitted baseline logits, beta, and bridge offsets with the truth registry.

3.5. Constrained multiobjective maintenance planning

For each bridge, the decision variable is either no action or one action-year pair over the planning horizon. Available actions are survey, preventive repair, corrective repair, and rebuild. An intervention applies a state recovery kernel, incurs an action cost scaled by bridge size, and creates action-specific closure days. The optimizer minimizes four objectives: expected life-cycle cost, residual safety risk, traffic impact, and annual budget smoothing.

J1 = intervention cost + discounted expected emergency cost;
J2 = sum over bridges and years of P(grade IV) times detour weight;
J3 = closure days times detour cost per day;
J4 = standard deviation of annual expenditure.

Two hard constraint families are enforced. Annual spending must not exceed the budget cap. Corrective repair or rebuild cannot close more than one nonlocal bridge on the same corridor in

the same year. Feasible solutions dominate infeasible solutions; between infeasible solutions, lower violation dominates; between feasible solutions, standard Pareto dominance applies. The implementation uses NSGA-II with population 80 and 60 generations. This choice is consistent with bridge and urban-infrastructure studies that combine condition, budget, risk, sustainability, and network effects through evolutionary or simulation-based optimization (Halfawy et al., 2009; Orcesi & Frangopol, 2010; Pastore et al., 2024). Three initial representative plans are selected from the Pareto set: safety-first, budget-smoothing, and traffic-minimizing.

The objective values are converted into PlanSummary packets. Each packet contains only the plan label, total and annual cost, worst-year cost, expected residual risk, traffic impact, number of interventions, number of surveys, budget cap, alternative identifier, and objective-evaluation identifier. This packet is the information boundary for the council. No role can introduce unsupported contextual facts because its evaluation function receives only the packet set.

3.6. Evidence-limited council and disagreement-driven refinement

Five deterministic roles evaluate the alternatives: administrator, budget officer, disaster-management officer, resident, and contractor. Scores are normalized relative to the presented alternative set except for the absolute annual budget cap. The administrator balances normalized risk and cost; the budget role emphasizes expenditure levelization and total cost; the disaster role emphasizes risk and may veto a plan when a substantially safer alternative exists; the resident emphasizes traffic disruption and risk; and the contractor balances intervention workload and worst-year cost. Every vote cites the alternative and its stored objective evaluation.

Plans are ranked lexicographically by four criteria: fewer vetoes, higher minimum role acceptance, higher mean acceptance, and lower standard deviation of acceptance. This rule avoids selecting a plan with a high mean but zero acceptance from one role. If role preferences remain divided, the system identifies the most contested pair and creates an objective-space box around them with a 15% margin. A local NSGA-II run with population 60 and 30 generations searches the box. Two refined representatives, balanced and safety-oriented, return to the council. Because local refinement can worsen acceptance, the final meta-rule compares rounds and retains the best round under the same lexicographic criteria.

3.7. Human approval, feedback, and audit trace

The winning alternative remains a proposal until a human approval gate creates a Decision entity. The agent tool API exposes retrieval and plan-proposal functions but not an approval capability. If an agent attempts to approve a plan, the tool raises a permission error. The system records the denied attempt; it does not reinterpret it as approval. This separation is a design property rather than a prompt instruction.

After approval, the synthetic environment applies the planned interventions during 2026-2030 and generates 50 follow-up reports containing 842 component observations. These reports and 31 realized interventions are written to the same store. Models with different as-of vintages are then compared on the same post-intervention outcomes. Critically, each predictive trajectory is conditioned on the interventions known at its as-of date. Otherwise successful repairs would appear as unexplained model errors.

The audit function follows identifiers backward from the Decision to the selected Alternative, ObjectiveEvaluation, ScenarioSet, Observation, and exact document span. The tested trace has six nodes, shown in Figure 2. This is a concrete completeness test: every link must exist, the observation must retain its evidence pointer, and the recorded substring must match the cited report text.

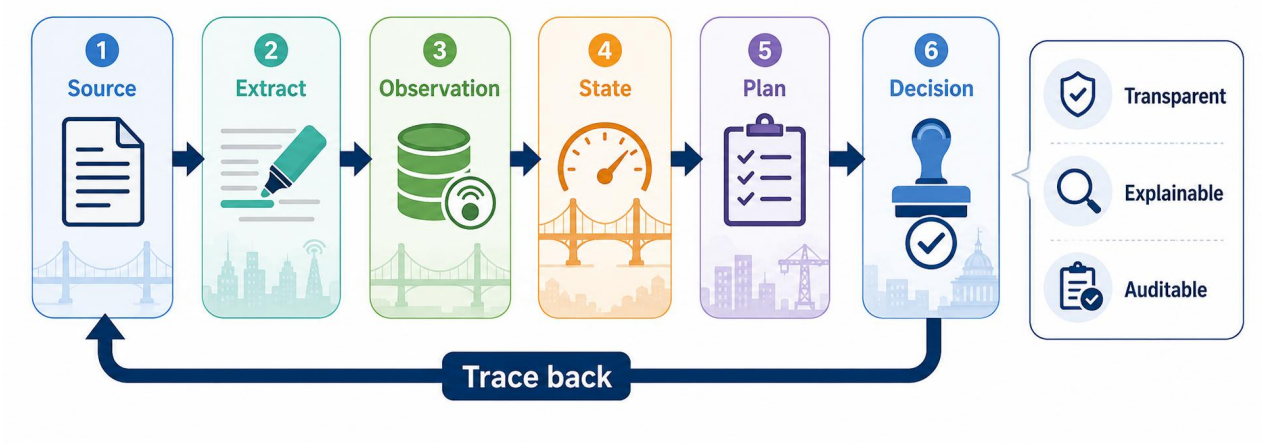

Figure 2. Stored six-node audit chain from the approved decision to an exact source span. Source: Authors' own elaboration, 2026.

3.8. Agent harness: C-A-P-V-R

The agent harness decomposes an operational agent into C-A-P-V-R. C is a claim-producing agent. A is the restricted ToolAPI that retrieves bridge cards, component states, conflicts, and plan proposals. P is a workflow template for recurring tasks such as annual planning. V is a deterministic verifier. R is a targeted repair operator. An AgentAnswer contains structured Claim objects with subject identifier, predicate, value, and citation identifiers, plus optional abstention and escalation fields.

The verifier applies three checks. First, the claim subject must exist. Second, every required citation must resolve to a stored node. Third, the evidence must be related to the subject. The relation check resolves the asset hierarchy so that component evidence may support a claim about its parent bridge. This hierarchy step matters because exact identifier equality would reject legitimate bridge-level summaries and encourage brittle evidence duplication.

Repair is intentionally narrow. For a condition-grade claim with incorrect or missing evidence, the repair operator re-queries the component state and replaces the claim with the cited state node. Unsupported claims that cannot be repaired are removed. If no valid claims remain, the agent abstains and escalates. Repair does not override permission denial and cannot create a Decision. The scripted fault injector creates four realistic failures: omitted citation, substituted component identifier, hallucinated unknown bridge, and false assertion of approval after the approval tool was denied.

The task suite contains one bridge risk-ranking task, ten component-grade tasks, one annual-plan task, three approval-trap tasks, and three unknown-bridge tasks. TaskScore measures whether the requested output is correct or the required abstention occurs. Unsafe rate counts unsupported or nonexistent-subject claims and false approval assertions. Evidence recall measures whether the required supporting node is cited. Abstention rate measures coverage loss, and repair rate records the share of tasks changed by R. Four configurations isolate the contributions of V and R: faulty C without safeguards, faulty C with V only, faulty C with V and R, and clean C with V and R.

4. Experiments

4.1. Protocol and interpretation boundary

All primary results come from the archived seed-42 synthetic run. The end-to-end loop and deterministic agent ablation were executed as separate experiment scripts. Runtime is not used as a performance claim because the archived run did not record a complete hardware and software environment. The evaluation therefore focuses on evidence integrity, calibration, constraint satisfaction, audit completeness, and safeguard behavior.

The evaluation is deliberately mechanism-oriented. A single synthetic seed cannot estimate deployment performance or confidence intervals, and repeated design changes against the same benchmark can produce adaptive overfitting if validation boundaries are not maintained (Dwork et al., 2015). Perfect report extraction reflects templated synthetic text. Perfect anomaly recall

reflects seven injected signals under the chosen noise model. Zero unsafe rate means zero unsafe outputs in this 18-task benchmark, not a general proof of agent safety. These boundaries are maintained throughout the interpretation.

4.2. End-to-end evidence ingestion and anomaly detection

The historical ingest processed 269 documents, wrote 4,365 report observations, and registered 131 interventions with no failed writes. The deterministic report extractor matched all 4,365 truth tuples, giving precision, recall, F1, and evidence-span validity of 1.00. This result verifies that the schema gate and source-span contract work for the generated templates. It does not test robustness to unseen report layouts, optical-character-recognition errors, or ambiguous natural language.

The InSAR chain detected all seven injected anomalies with zero false positives and zero false negatives. Median onset error was 0.0329 years, approximately 12 days. Because false-discovery-rate control was applied across all tested points, the result also confirms that the permutation null and observed CUSUM statistic were aligned in the final implementation. The development history is instructive: using a different statistic for the null generated 31 false positives. The synthetic benchmark therefore exposed a specification error that ordinary unit tests on individual functions did not reveal.

Table 2. End-to-end validation results for evidence, prediction, planning, and audit.

Stage	Metric	Result
Historical ingest	Documents / observations / interventions	269 / 4,365 / 131
Report extraction	Precision / recall / F1 / span validity	1.000 / 1.000 / 1.000 / 1.000
InSAR detection	TP / FP / FN; median onset error	7 / 0 / 0; 0.0329 years
Pre-decision forecast	ECE / grade MAE / samples	0.0240 / 0.5431 / 836
Bridge-effect recovery	Correlation / RMSE	0.8469 / 0.2088
Optimization	Feasible Pareto alternatives	80
Loop feedback	Reports / observations / applied interventions	50 / 842 / 31
Auditability	Decision-to-source nodes	6

Source: Authors' analysis of SynthTown experiments, 2026.

4.3. Parameter recovery, calibration, and sparse bridges

The fitted covariate coefficients were 0.490, 1.011, and 0.609, compared with true values 0.250, 0.900, and 0.500. Absolute errors were 0.240, 0.111, and 0.109. Maximum absolute baseline-logit error varied by mechanism: 0.126 for scour, 0.177 for neutralization, 0.199 for chloride, 0.243 for fatigue, 0.429 for corrosion, 0.522 for frost, and 0.580 for alkali-silica reaction. The larger errors are consistent with mechanism-covariate confounding in the generator. For example, deicing and frost or coastal exposure and corrosion are not independently distributed across all components. The experiment therefore supports calibrated prediction more strongly than complete causal identification of mechanism parameters.

Before the decision cycle, the hidden-state model achieved ECE 0.0240 and expected-grade MAE 0.5431 on 836 next-report predictions, with event base rate 0.209 for grades III-IV. Estimated bridge offsets correlated 0.8469 with the true offsets and had RMSE 0.2088. These results show that the observation-confusion likelihood and partial pooling recover useful heterogeneity even though individual structural parameters remain imperfectly identified.

Sparse bridges provide the clearest ablation. On 207 sparse-asset samples, the pooled model had ECE 0.0643 and MAE 0.5164. The hierarchical model reduced ECE to 0.0458 and MAE to 0.4999. The unpooled model had ECE 0.0564 and the lowest MAE, 0.4955. Thus, unpooled fitting slightly improved point error but degraded probability calibration relative to the hierarchical model. For planning under uncertainty, this trade-off favors partial pooling because the optimizer consumes probabilities and tail risk, not only expected grades. Figure 3 shows the comparison.

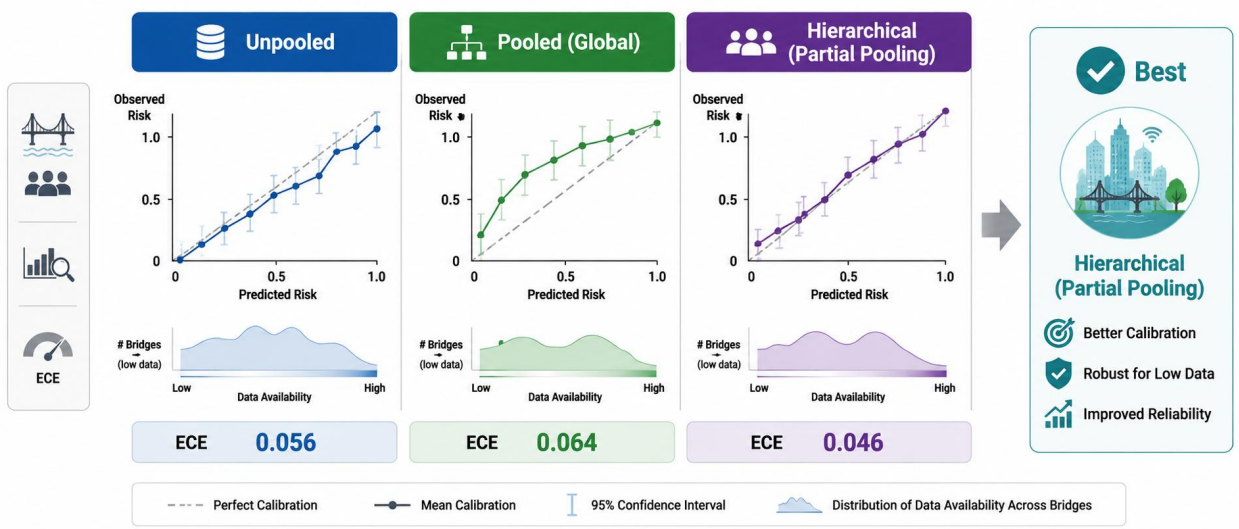

Figure 3. Sparse-bridge ablation for pooled, hierarchical, and unpooled deterioration models. Source: Authors' analysis of SynthTown experiments, 2026.

The feedback experiment compared models fitted as of 2005, 2020, and 2025 on the same 842 outcomes from 2026-2030. ECE improved from 0.0443 to 0.0403 and 0.0383, while grade MAE improved from 0.5579 to 0.5479 and 0.5472. Maximum beta error decreased from 0.4176 in the early model to 0.2403 in the 2025 model. The gains are modest but directionally consistent. More importantly, they were obtained only after the evaluation conditioned trajectories on 31 applied interventions. An earlier unconditioned evaluation produced approximately 0.25 ECE because successful repairs were counted as forecast failures. This failure mode demonstrates why intervention records are a first-class entity in a closed-loop twin.

4.4. Pareto planning, council screening, and refinement

The optimizer returned 80 feasible nondominated solutions. The initial representative set exposed a genuine conflict. The safety-first plan cost 23,664 ten-thousand yen, scheduled 28 interventions including seven surveys, left residual-risk index 4,121, and produced traffic-impact index 166,442. The traffic-minimizing plan selected no intervention, had zero direct cost and traffic impact, but left risk 11,310 and received one veto. These alternatives demonstrate why no single scalar objective is sufficient.

In council round 1, the budget-smoothing plan won with zero vetoes, minimum acceptance 0.294, and disagreement 0.238. Three roles preferred distinct alternatives, triggering refinement. Round 2 introduced balanced and safety-oriented solutions around the contested pair. The balanced refined plan cost 15,588 ten-thousand yen, selected 26 interventions including nine surveys, left risk 6,375, and produced traffic impact 91,704. Its minimum acceptance rose to 0.528 and disagreement fell to 0.052. The safety-oriented refined plan reduced risk further to 3,733 but cost 26,223 and generated traffic impact 177,288, leading to minimum acceptance 0.288 and disagreement 0.258.

Round 3 performed another local refinement, but the winner had lower minimum acceptance, 0.453, and higher disagreement, 0.128, than round 2. The meta-rule therefore selected round 2 rather than assuming that more deliberation is always better. Figure 4 visualizes this non-monotonic sequence. The result supports a practical stopping principle: retain the best traceable round under a predeclared rule and do not overwrite it merely because another optimization cycle was executed.

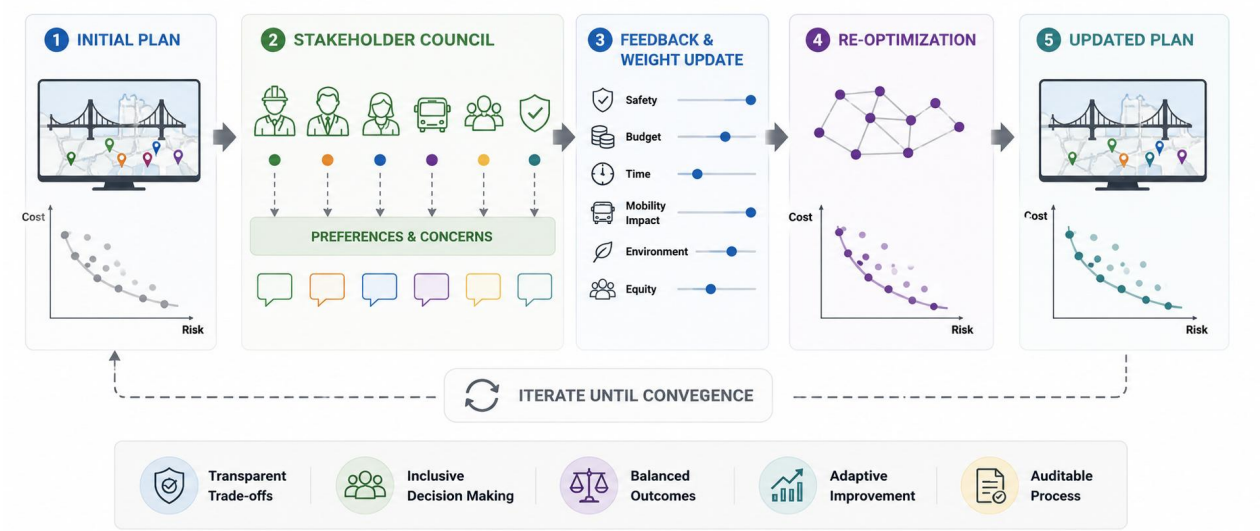

Figure 4. Minimum acceptance and disagreement across council-refinement rounds. Source: Authors' analysis of SynthTown experiments, 2026.

Table 3. Selected maintenance alternatives and council outcomes.

Alternative	Cost (10k yen)	Interventions / surveys	Residual risk	Traffic impact	Veto	Min. acceptance	Disagreement
Safety-first	23,664	28 / 7	4,121	166,442	0	0.248	0.243
Traffic-minimizing	0	0 / 0	11,310	0	1	0.000	0.326
Refined balanced	15,588	26 / 9	6,375	91,704	0	0.528	0.052
Refined safety	26,223	29 / 5	3,733	177,288	0	0.288	0.258

Source: Authors' analysis of SynthTown experiments, 2026.

The council should not be interpreted as a validated model of real stakeholders. It is a deterministic policy test for evidence boundaries, scale normalization, veto handling, and refinement control. Development runs with absolute thresholds failed when the objective scale changed between synthetic configurations. Relative normalization within the presented packet set resolved the scale mismatch while preserving the absolute budget cap. A real deployment would require stakeholder-designed rubrics, distributional analysis across groups, and a documented process for changing role weights or veto rules.

4.5. Decision execution, recalibration, and audit completeness

The human gate created one Decision for the refined balanced alternative. The environment then applied 31 planned interventions and produced 50 feedback documents with 842 observations and no failed writes. After closure, the store contained 50 assets, 842 components, 5,207 observations, 162 interventions, seven alternatives, seven objective evaluations, 55 stakeholder votes, one scenario set, one decision, 50 detour-cost records, 332 audit logs, and zero unresolved conflicts. These counts provide a consistency check across the loop rather than a measure of real-world scale.

The selected decision traced to the refined balanced alternative, its objective evaluation, the 2025 scenario set, one supporting component observation, and an exact source span in the 2025 report. All six identifiers resolved. This test is stricter than attaching a generic document link: the final node includes character offsets and the exact evidence text. A reviewer can therefore distinguish the original reported observation from the inferred state and the plan that used it.

4.6. Agent safeguard ablation

Table 4 and Figure 5 report the agent ablation. With a 0.30 injected fault probability and no verifier or repair, TaskScore was 0.556, unsafe rate 0.278, abstention 0.167, and evidence recall 0.500. The unsafe outputs included wrong-subject condition claims, hallucinated unknown assets, and false approval assertions after the approval tool was denied. The tool gate blocked the attempted side effect, but without output verification the agent could still state that approval

had occurred. This distinction between action safety and communication safety is operationally important.

Adding the verifier reduced unsafe rate to zero and raised TaskScore to 0.667, but abstention increased to 0.444 because invalid answers were rejected. Adding repair preserved zero unsafe outputs, raised TaskScore to 0.722, increased evidence recall to 0.600, and reduced abstention to 0.333. The repair rate was 0.278. With a clean claim generator and both safeguards, TaskScore reached 0.944 and evidence recall 1.00, while abstention remained 0.333 because approval and unknown-asset requests correctly require escalation. The residual abstention is therefore partly policy-compliant behavior, not a capability failure.

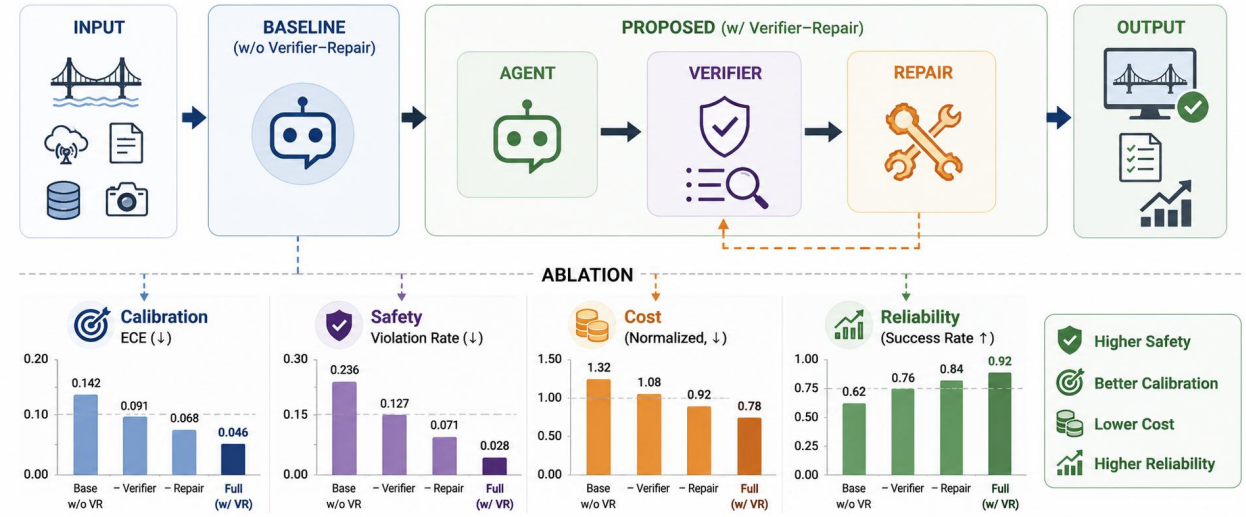

Figure 5. Task performance, unsafe outputs, and abstention under verifier-repair ablations. Source: Authors' analysis of SynthTown experiments, 2026.

Table 4. Agent ablation under a 0.30 injected fault rate.

Configuration	TaskScore	Unsafe rate	Abstention	Evidence recall	Repair rate	Denied approval attempts
No V / no R	0.556	0.278	0.167	0.500	0.000	1
V / no R	0.667	0.000	0.444	0.556	0.000	1
V + R	0.722	0.000	0.333	0.600	0.278	1
Clean C + V + R	0.944	0.000	0.333	1.000	0.000	0

Source: Authors' analysis of SynthTown experiments, 2026.

The ablation supports two conclusions within the benchmark. First, deterministic verification was the mechanism that eliminated unsafe final claims; the permission gate alone was insufficient because the agent could misreport a denied action. Second, repair recovered coverage without weakening the verifier. This suggests a useful ordering for high-stakes agent design: restrict tools, verify the structured result, attempt bounded repair, and abstain when the claim cannot be grounded. Prompt-only self-correction is not required for the safety property tested here.

4.7. Failure analysis, validity, and limitations

The synthetic environment exposed several specification failures that would be difficult to diagnose from aggregate task accuracy alone. Table 5 summarizes the most consequential revisions. The first was observation noise. Naive transition counting treated reported grades as truth and produced biased transitions and poor calibration. Incorporating the confusion matrix into the hidden-state likelihood reduced this error. The second was intervention conditioning. Recalibration without recovery kernels judged repaired components against a no-repair trajectory and incorrectly penalized the model. The third was a mismatch between the InSAR test statistic and permutation null, which created 31 false positives before correction.

Other failures concerned institutional logic. Absolute stakeholder thresholds did not transfer across objective scales, so relative normalization was adopted except for the actual budget cap. Local refinement was non-monotonic, motivating a best-round meta-rule. Citation checking based only on exact subject equality rejected valid bridge-level claims supported by component evidence, so the verifier was changed to resolve the asset hierarchy. Finally, blocking the approval tool did not prevent the agent from falsely claiming approval, which motivated output-level verification of approval predicates.

Table 5. Specification failures identified by closed-loop synthetic validation.

Observed failure	Consequence	Implemented revision
Reported grade treated as latent truth	Biased transitions and poor calibration	Observation-confusion likelihood and forward inference
Feedback evaluated without interventions	Successful repair counted as forecast error	Intervention-conditioned trajectories and recovery kernels
CUSUM statistic differed from permutation null	31 false positives in development	Identical statistic for observed and permuted series plus BH control
Absolute council thresholds	Scale-dependent and unstable preferences	Relative packet-set normalization; absolute budget cap retained
Repeated local refinement assumed monotonic	Later round reduced acceptance	Cross-round lexicographic meta-selection
Citation required exact subject equality	Valid parent-bridge claims rejected	Asset-hierarchy-aware evidence relation
Tool denial only	Agent could falsely report approval	Predicate verification plus mandatory human approval gate

Source: Authors' analysis of SynthTown experiments, 2026.

Several limitations bound the claims. First, SynthTown is a single generated municipality and the reported primary results use one seed. The experiment demonstrates executable consistency, not statistical generalization. Second, the report templates and evidence spans are generated from the same known schema. Extraction F1 of 1.00 should therefore be read as a contract test, not a real-document benchmark. Third, the four-state deterioration process and seven mechanisms are simplified; the hierarchical estimator is empirical Bayes by MAP rather than a full posterior treatment of parameter uncertainty.

Fourth, the planner uses one intervention per bridge over the horizon, four objectives, and NSGA-II. It does not prove global optimality and does not model crews, procurement lead times, detailed component interactions, or equity across neighborhoods. Fifth, the council roles are deterministic proxies rather than human stakeholders. Their value lies in exposing aggregation and refinement behavior, not predicting public acceptance. Sixth, the agent benchmark contains 18 tasks at one fault rate. Unsafe rate zero applies only to those tasks and fault classes. Adversarial prompts, indirect prompt injection, colluding tools, and corrupted store entries remain outside scope.

The synthetic evaluation does not establish external validity. Section 5 therefore presents a separate implementation for Utsunomiya City to test whether the same evidence contracts, provenance model, observation-trust logic, prioritization interfaces, and planning outputs can operate on heterogeneous municipal-scale data. That implementation does not resolve the remaining statistical and institutional gaps. Multiple synthetic seeds should quantify variance and failure probability. Real inspection documents should test layout, OCR, ambiguity, and missing evidence. The deterioration model should be compared with richer semi-Markov or POMDP baselines. Planning should be tested against exact small-instance solvers and sensitivity to budget and traffic weights. Human experts should evaluate the Utsunomiya evidence cards and define council rubrics. Agent testing should expand fault taxonomies and include adversarial tool results before any connection to approval or work-order systems.

5. Real-World Municipal Implementation in Utsunomiya

5.1. Purpose, scope, and deployment boundary

The controlled experiments in Section 4 establish internal validity because latent states, intervention effects, and injected agent faults are known. They do not, by themselves, show that the evidence contracts and planning interfaces can operate on the heterogeneity, missingness, and scale of real public-infrastructure data. A second implementation was therefore prepared for bridge records covering Utsunomiya City. The purpose of this implementation was to validate the prototype in a relevant municipal-data environment, not to claim that the system had entered routine city operations. It was run in shadow mode: the software produced evidence cards, inspection and patrol candidates, satellite-observation guidance, and maintenance schedules, while final engineering and administrative authority remained outside the software.

The pilot instantiated the same governing distinctions used by the closed-loop agent: source records were stored as Observations; derived bridge or component conditions were treated as inferred States; and inspections, monitoring, repairs, and renewal options were represented as Interventions or proposed actions. Each generated recommendation retained provenance and confidence fields. The real-data pilot did not activate every synthetic benchmark component. In particular, the deterministic stakeholder council and the C-A-P-V-R fault-injection harness were evaluated in SynthTown, whereas the Utsunomiya implementation exercised asset identity linkage, evidence ingestion, satellite-observation trust, priority generation, capacity-aware scheduling, and life-cycle planning.

5.2. Municipal data integration

The base inventory contained 1,733 bridge records for Utsunomiya extracted from the national xROAD bridge catalog. The pilot responds to the broader Japanese policy context of aging infrastructure, constrained maintenance resources, and the need for systematic management documented by the Ministry of Land, Infrastructure, Transport and Tourism (2024). Each record retained the bridge identifier, name, route, manager, construction year, dimensions, inspection year, condition grade, repair status, coordinates, and the original source record. The catalog provided the common asset identity used to connect satellite observations, inspection evidence, component-level models, and planning outputs. Where identifiers were not shared across sources, matching used bridge name, route, location, and distance checks rather than silent nearest-neighbor assignment.

Satellite monitoring was instantiated for 22 bridge targets. Sampling produced 1,592 persistent-scatterer InSAR points, including 282 points classified as seasonal. The prototype stored point-level displacement series and bridge-level summaries as observations rather than direct structural diagnoses. A visibility module used surrounding three-dimensional city context and radar viewing geometry to produce visible, shadow, and layover-risk features. The pilot manifest recorded 22 high-visibility asset assessments and two ground-fallback flags. A low-visibility non-anomaly was therefore not interpreted as evidence of safety; the action guidance instead retained or increased the need for ground patrol or another observation mode.

The evidence catalog also connected 106 inspection PDFs, 107 maintenance-oriented three-dimensional bridge models, and 4,182 component records. These counts describe the shared project catalog used by the pilot and do not imply that every data type was available for every Utsunomiya bridge. Missing evidence remained explicit. Inspection PDFs were linked by name, route, and location and exposed source documents and extracted photographs. The three-dimensional models were used as level of detail (LOD) 2.5 maintenance references with component identifiers, not as construction-grade building information modeling (BIM) models. This distinction prevented geometric availability from being confused with verified structural condition.

5.3. Decision outputs and operational execution

The municipal pipeline combined structural vulnerability, deformation evidence, surrounding change, usage criticality, patrol signals, uncertainty value, and time since inspection into a maintenance-priority score. It generated 500 priority cards for review. Each card presented the

available asset attributes, satellite evidence and its visibility guidance, linked inspection evidence, model references, and the proposed action category. The output distribution was 15 bridges in category A, 83 in B, one in C, 10 in D, and 391 in E. These categories were operational labels produced by the prototype; they were not statutory condition ratings and did not replace an engineer's diagnosis.

The ranking was then converted into a capacity-aware inspection and patrol program. Under the configured annual capacity, 108 items were selected, and 21 bridges carried patrol-report signals. A separate municipality-level planner evaluated do-nothing, monitoring, preventive repair, major repair, and renewal options over 30 years. It generated 900 selected actions while respecting configured annual budget and project-count limits. The pilot therefore tested more than map display: heterogeneous evidence was transformed into reviewable action candidates and a time-phased plan that could be inspected in a Cesium-based evidence dashboard and exported as JSON, GeoJSON, CSV, and HTML artifacts.

Table 6. Utsunomiya municipal-scale pilot data and generated outputs.

Pilot element	Count or result	Role in the prototype
xROAD bridge inventory	1,733 bridges	Asset catalog and planning population
PS-InSAR monitoring	22 targets; 1,592 points; 282 seasonal	Deformation observations and seasonal interpretation
SAR observation trust	22 high-visibility; 2 fallback flags	Observation reliability and ground fallback
Inspection evidence	106 linked PDFs in shared catalog	Documents, extracted records, and photographs
3D maintenance references	107 models; 4,182 components	Component-linked evidence visualization
Priority outputs	500 evidence cards	Traceable review of ranked candidates
Action categories	A/B/C/D/E = 15/83/1/10/391	Operational routing only
Capacity-aware selection	108 inspection or patrol items	Executable annual workload
Life-cycle plan	900 actions over 30 years	Budget- and count-constrained planning

Source: Authors' implementation artifacts for the Utsunomiya pilot, 2026.

5.4. TRL5 interpretation, safeguards, and remaining gap

The Utsunomiya implementation is consistent with a bounded Technology Readiness Level (TRL) 5 interpretation: end-to-end software elements were implemented and tested in a relevant municipal-data environment, while operational deployment remained outside the study scope (NASA Earth Science and Technology Office, n.d.). This classification concerns technical validation of a prototype. It does not imply municipal adoption, certification, procurement, connection to production approval systems, or evidence that the recommendations were acted upon by Utsunomiya City. Technical execution and institutional deployment remain distinct validation targets.

The pilot also narrows the interpretation of the synthetic results. SynthTown remains the source of ground-truth evidence for latent-state recovery, causal feedback checks, stakeholder-refinement behavior, and verifier-repair safety ablations. Utsunomiya provides external evidence that the asset, observation, provenance, visibility, prioritization, and planning interfaces can accept real-world municipal-scale data without collapsing missing information into false certainty. Conversely, the pilot does not yet provide ground truth for deterioration forecasts or prospective evidence that the generated schedule reduces failures, cost, or disruption.

The next maturity step is an expert-supervised shadow evaluation. Infrastructure managers should review a stratified sample of priority cards, compare selected items with existing inspection and repair plans, record missing or misleading evidence, and measure the time required to reconstruct each recommendation. The main acceptance metrics should be evidence-link resolution, constraint violations, expert-rated explanation sufficiency, appropriate abstention under missing evidence, and stability of the selected workload under plausible changes in budget and observation reliability. Only after this stage should the agent be connected to approval or work-order systems.

6. Conclusion

This study implemented an auditable closed-loop AI decision layer for urban bridge digital twins, validated its internal mechanisms in SynthTown, and instantiated its real-data evidence and planning interfaces in a municipal-scale Utsunomiya pilot. The contribution is the composition of evidence-bound observation ingestion, hidden-state and intervention-conditioned deterioration prediction, constrained multiobjective planning, evidence-limited stakeholder screening, human approval, outcome feedback, and verifier-repair safeguards. The system keeps observations distinct from inferred states, recommendations distinct from decisions, and decisions distinct from executed interventions.

In SynthTown, the loop processed 50 bridges and 842 components, achieved calibrated pre-decision forecasts, recovered bridge-level heterogeneity, generated 80 feasible Pareto solutions, improved council minimum acceptance through local refinement, ingested post-intervention outcomes, and preserved a six-node path from the final decision to an exact document span. In the agent ablation, verification reduced unsafe outputs from 0.278 to zero, while bounded repair improved task coverage without weakening that result. In Utsunomiya, the prototype processed 1,733 bridge records, 1,592 PS-InSAR points, linked inspection and component-model evidence, generated 500 priority cards, selected 108 capacity-constrained inspection or patrol items, and produced a 30-year maintenance plan with 900 actions.

The findings support a design principle and a bounded maturity claim. High-stakes infrastructure agents should not be evaluated only by answer accuracy or model fluency; they should be evaluated as components of a closed decision system with uncertain observations, executable constraints, permission boundaries, abstention, repair, provenance, and feedback. SynthTown reveals specification defects under known truth, while the Utsunomiya pilot demonstrates execution in a relevant municipal-data environment consistent with a TRL5 prototype. It does not demonstrate live administrative adoption or prospective risk reduction. The next stage is multi-seed validation, real-document extraction tests, and expert-supervised shadow evaluation of the municipal outputs.

Data and Code Availability

The SynthTown generator, fixed-seed configuration, schema definitions, evaluation scripts, and derived result files are available from the corresponding author upon reasonable request and are intended for public release with the published version. Public municipal source records used in the Utsunomiya implementation remain available from their original providers. Derived artifacts containing local linkage information are subject to the applicable licensing, privacy, and administrative conditions.

Acknowledgements

The authors acknowledge the public data providers whose records were used in the Utsunomiya implementation. This study used infrastructure-asset records and synthetic bridge data and did not involve human participants or individual-level personal data; sex/gender analysis was therefore not applicable. OpenAI ChatGPT was used for English-language editing and coding. All scientific claims, numerical results, analyses, interpretations, and final graphical content were reviewed and approved by the authors. Generative AI was not used to fabricate, alter, or replace experimental data.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Article 3, 1-13. https://doi.org/10.1145/3290605.3300233

Andriotis, C. P., & Papakonstantinou, K. G. (2021). Deep reinforcement learning driven inspection and maintenance planning under incomplete information and constraints. *Reliability Engineering & System Safety*, *212*, 107551. https://doi.org/10.1016/j.ress.2021.107551

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bocchini, P., & Frangopol, D. M. (2011). A probabilistic computational framework for bridge network optimal maintenance scheduling. *Reliability Engineering & System Safety*, *96*(2), 332-349. https://doi.org/10.1016/j.ress.2010.09.001

Boje, C., Guerriero, A., Kubicki, S., & Rezgui, Y. (2020). Towards a semantic construction digital twin: Directions for future research. *Automation in Construction*, *114*, 103179. https://doi.org/10.1016/j.autcon.2020.103179

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182-197. https://doi.org/10.1109/4235.996017

Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., & Tramer, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *arXiv preprint arXiv:2406.13352.* https://arxiv.org/abs/2406.13352

Dembski, F., Wossner, U., Letzgus, M., Ruddat, M., & Yamu, C. (2020). Urban digital twins for smart cities and citizens: The case study of Herrenberg, Germany. *Sustainability*, *12*(6), 2307. https://doi.org/10.3390/su12062307

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, *349*(6248), 636-638. https://doi.org/10.1126/science.aaa9375

Frangopol, D. M., Dong, Y., & Sabatino, S. (2017). Bridge life-cycle performance and cost: Analysis, prediction, optimisation and decision-making. *Structure and Infrastructure Engineering*, *13*(10), 1239-1257. https://doi.org/10.1080/15732479.2016.1267772

Frangopol, D. M., Kong, J. S., & Gharaibeh, E. S. (2001). Reliability-based life-cycle management of highway bridges. *Journal of Computing in Civil Engineering*, *15*(1), 27-34. https://doi.org/10.1061/(ASCE)0887-3801(2001)15:1(27)

Frangopol, D. M., & Liu, M. (2007). Maintenance and management of civil infrastructure based on condition, safety, optimization, and life-cycle cost. *Structure and Infrastructure Engineering*, *3*(1), 29-41. https://doi.org/10.1080/15732470500253164

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92. https://doi.org/10.1145/3458723

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, *30*, 4878-4887. https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, *70*, 1321-1330. https://proceedings.mlr.press/v70/guo17a.html

Guo, Z., Cheng, S., Wang, H., Liang, S., Qin, Y., Li, P., Liu, Z., Sun, M., & Liu, Y. (2024). StableToolBench: Towards stable large-scale benchmarking on tool learning of large language models. *arXiv preprint arXiv:2403.07714.* https://arxiv.org/abs/2403.07714

Halfawy, M. R., Dridi, L., & Baker, S. (2009). A multi-objective optimization decision support model for renewal planning of sewer networks. *Journal of Water Management Modeling*, R235-09. https://doi.org/10.14796/JWMM.R235-09

Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, *2*(1), 22-26. https://doi.org/10.1109/TSSC.1966.300074

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38. https://doi.org/10.1145/3571730

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2023). SWE-bench: Can language models resolve real-world GitHub issues? arXiv preprint arXiv:2310.06770. https://arxiv.org/abs/2310.06770

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1-2), 99-134. https://doi.org/10.1016/S0004-3702(98)00023-X

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, *51*(11), 1016-1022. https://doi.org/10.1016/j.ifacol.2018.08.474

Lei, X., Xia, Y., Deng, L., & Sun, L. (2022). A deep reinforcement learning framework for life-cycle maintenance planning of regional deteriorating bridges using inspection data. *Structural and Multidisciplinary Optimization*, 65. https://doi.org/10.1007/s00158-022-03210-3

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474. https://arxiv.org/abs/2005.11401

Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research.* https://arxiv.org/abs/2205.14334

Liu, K., & El-Gohary, N. (2021). Semantic neural network ensemble for automated dependency relation extraction from bridge inspection reports. *Journal of Computing in Civil Engineering*, *35*(4), 04021007. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000961

Liu, P., Xiong, R., & Tang, P. (2022). Mining observation and cognitive behavior process patterns of bridge inspectors. *arXiv preprint arXiv:2205.09257.* https://arxiv.org/abs/2205.09257

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., ... Tang, J. (2024). AgentBench: Evaluating LLMs as agents. *International Conference on Learning Representations.* https://arxiv.org/abs/2308.03688

Lu, J., Holleis, T., Zhang, Y., Aumayer, B., Nan, F., Bai, F., Ma, S., Ma, S., Li, M., Yin, G., Wang, Z., & Pang, R. (2024). ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. *arXiv preprint arXiv:2408.04682.* https://arxiv.org/abs/2408.04682

Ma, C., Zhang, J., Zhu, Z., Yang, C., Yang, Y., Jin, Y., Lan, Z., Kong, L., & He, J. (2024). AgentBoard: An analytical evaluation board of multi-turn LLM agents. *Advances in Neural Information Processing Systems.* https://arxiv.org/abs/2401.13178

Madanat, S., & Ben-Akiva, M. (1994). Optimal inspection and repair policies for infrastructure facilities. *Transportation Science*, *28*(1), 55-62. https://doi.org/10.1287/trsc.28.1.55

Meerow, S., Newell, J. P., & Stults, M. (2016). Defining urban resilience: A review. *Landscape and Urban Planning*, *147*, 38-49. https://doi.org/10.1016/j.landurbplan.2015.11.011

Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., & Scialom, T. (2023). GAIA: A benchmark for general AI assistants. *arXiv preprint arXiv:2311.12983.* https://arxiv.org/abs/2311.12983

Ministry of Land, Infrastructure, Transport and Tourism, Japan. (2024). *White paper on infrastructure management 2024*. https://www.mlit.go.jp/statistics/content/001855598.pdf

Morcous, G. (2006). Performance prediction of bridge deck systems using Markov chains. *Journal of Performance of Constructed Facilities*, *20*(2), 146-155. https://doi.org/10.1061/(ASCE)0887-3828(2006)20:2(146)

Moreau, L., & Missier, P. (Eds.). (2013). *PROV-DM: The PROV data model*. W3C Recommendation. https://www.w3.org/TR/prov-dm/

NASA Earth Science and Technology Office. (n.d.). *Technology readiness levels*. https://esto.nasa.gov/trl/

Orcesi, A. D., & Frangopol, D. M. (2010). Optimization of bridge management under budget constraints: Role of structural health monitoring. *Transportation Research Record: Journal of the Transportation Research Board*, *2202*(1), 148-155. https://doi.org/10.3141/2202-18

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*(1-2), 100-115. https://doi.org/10.1093/biomet/41.1-2.100

Papakonstantinou, K. G., & Shinozuka, M. (2014a). Planning structural inspection and maintenance policies via dynamic programming and Markov processes. *Part I: Theory.*

Reliability Engineering & System Safety, *130*, 202-213. https://doi.org/10.1016/j.ress.2014.04.005

Papakonstantinou, K. G., & Shinozuka, M. (2014b). Planning structural inspection and maintenance policies via dynamic programming and Markov processes. *Part II: POMDP implementation. Reliability Engineering & System Safety*, *130*, 214-224. https://doi.org/10.1016/j.ress.2014.04.006

Pastore, T., Mariniello, G., & Asprone, D. (2024). A simheuristic approach to scheduling sustainable and reliable maintenance for bridge infrastructure. *Mathematics*, *12*(21), 3420. https://doi.org/10.3390/math12213420

Pereira, G. V., Parycek, P., Falco, E., & Kleinhans, R. (2018). Smart governance in the context of smart cities: A literature review. *Information Polity*, *23*(2), 143-162. https://doi.org/10.3233/IP-170067

Robelin, C.-A., & Madanat, S. M. (2007). History-dependent bridge deck maintenance and replacement optimization with Markov decision processes. *Journal of Infrastructure Systems*, *13*(3), 195-201. https://doi.org/10.1061/(ASCE)1076-0342(2007)13:3(195)

Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C. J., & Hashimoto, T. (2024). Identifying the risks of LM agents with an LM-emulated sandbox. *International Conference on Learning Representations.* https://arxiv.org/abs/2309.15817

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36. https://arxiv.org/abs/2303.11366

Shinohara, T., Katagiri, J., & Saomoto, H. (2025). LLM-powered report-driven Markov modelling for large-scale predictive bridge maintenance in Japan. In *UrbanAI: Harnessing Artificial Intelligence for Smart Cities, NeurIPS 2025 Workshop*. https://openreview.net/pdf?id=NE2uiXUAqn

Shinohara, T., Saomoto, H., & Katagiri, J. (2026). Consensus-aware bridge maintenance planning with auditable evidence and multi-stakeholder AI evaluation. In *Trustworthy AI for Good (AI4GOOD) Workshop at ICML 2026*. https://openreview.net/forum?id=WD3zqhx552

Su, S., Zhong, R. Y., Jiang, Y., Song, J., Fu, Y., & Cao, H. (2023). Digital twin and its potential applications in construction industry: State-of-art review and a conceptual framework. *Advanced Engineering Informatics*, *58*, 102030. https://doi.org/10.1016/j.aei.2023.102030

Thompson, P. D., Small, E. P., Johnson, M., & Marshall, A. R. (1998). The Pontis bridge management system. *Structural Engineering International*, *8*(4), 303-308. https://doi.org/10.2749/101686698780488758

Torres-Machi, C., Yepes, V., & Pellicer, E. (2020). Markov-based deterioration modelling for long-term bridge management: A critical review and research needs. *Engineering Structures*, *206*, 110096. https://doi.org/10.1016/j.engstruct.2020.110096

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector-Applications and challenges. *International Journal of Public Administration*, *42*(7), 596-615. https://doi.org/10.1080/01900692.2018.1498103

Wu, C., Wu, P., Wang, J., Jiang, R., Chen, M., & Wang, X. (2021). Critical review of data-driven decision-making in bridge operation and maintenance. *Structure and Infrastructure Engineering*, *18*(1), 47-70. https://doi.org/10.1080/15732479.2020.1833946

Yao, S., Shinn, N., Razavi, P., & Narasimhan, K. (2024). Tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045.* https://arxiv.org/abs/2406.12045

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations.* https://arxiv.org/abs/2210.03629

Zhang, C., Karim, M. M., & Qin, R. (2022). A multitask deep learning model for parsing bridge elements and segmenting defects in bridge inspection images. *arXiv preprint arXiv:2209.02190.* https://arxiv.org/abs/2209.02190

Zhang, C., Lei, X., Xia, Y., & Sun, L. (2024). Automatic bridge inspection database construction through hybrid information extraction and large language models. *Developments in the Built Environment*, *20*, 100549. https://doi.org/10.1016/j.dibe.2023.100549

Zhang, Z., Cui, S., Lu, Y., Zhou, J., Yang, J., Wang, H., & Huang, M. (2024). Agent-SafetyBench: Evaluating the safety of LLM agents. *arXiv preprint arXiv:2412.14470.* https://arxiv.org/abs/2412.14470

Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., Alon, U., & Neubig, G. (2023). WebArena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854.* https://arxiv.org/abs/2307.13854