Street Art & Urban Creativity | Vol. 12, No. 5, 2026 | ISSN 2183-9956
Street Art & Urban Creativity. Scientific Journal
https://doi.org/10.62161/sauc.v12.6387

AGENTES DE IA AUDITABLES DE LAZO CERRADO PARA GEMELOS DIGITALES DE PUENTES URBANOS

Validación sintética seguida de una implementación a escala municipal en Utsunomiya

TAKAYUKI SHINOHARA (SHINOHARA.TAKAYUKI@AIST.GO.JP)[1]
[1] AIST, Japón

PALABRAS CLAVE

IA agéntica
Gemelos digitales urbanos
Ciudades inteligentes
Mantenimiento de puentes
Resiliencia urbana
Planificación multiobjetivo
IA auditable

RESUMEN

El mantenimiento de los puentes urbanos es un problema de gobernanza de la ciudad inteligente que implica evidencia incierta, presupuestos restringidos y acciones críticas para la seguridad. Presentamos una arquitectura de inteligencia artificial auditable de lazo cerrado que separa las observaciones, los estados inferidos, las alternativas, las decisiones y las intervenciones. Los mecanismos se validan en SynthTown, un municipio de verdad de referencia con 50 puentes y 842 componentes, mediante la predicción de estados ocultos, la planificación multiobjetivo, el cribado por parte de los actores interesados, la recalibración por retroalimentación y salvaguardas de verificación-reparación. Una implementación con datos públicos abarca 1733 puentes de Utsunomiya e integra 1592 puntos de radar de apertura sintética interferométrico de dispersores persistentes, registros de inspección, modelos de mantenimiento tridimensionales y planes conscientes de la capacidad. La verificación eliminó las salidas inseguras en el banco de pruebas sintético

Recibido: 12 / 04 / 2026
Aceptado: 21 / 08 / 2026

1. Introducción

Las redes de puentes urbanos sostienen la movilidad diaria, el acceso de emergencia, la logística y la continuidad de los servicios públicos. Su mantenimiento no es, por tanto, solo un problema de ingeniería, sino también un problema de gobernanza de la ciudad inteligente que implica la resiliencia urbana, la asignación de recursos públicos y el uso responsable de la inteligencia artificial (IA) (Meerow et al., 2016; Pereira et al., 2018; Wirtz et al., 2019). El mantenimiento de puentes es un problema de decisión pública secuencial, más que una única tarea de predicción (Frangopol et al., 2001; Frangopol y Liu, 2007; Wu et al., 2021). Una administración observa evidencia incompleta y ruidosa, infiere el estado latente, compara planes bajo restricciones presupuestarias y de red, autoriza intervenciones y, más tarde, recibe evidencia alterada por esas intervenciones. Un sistema que predice el deterioro, pero que no puede preservar la evidencia, hacer respetar los límites de autoridad ni aprender de las acciones ejecutadas, está incompleto como sistema de apoyo a la decisión pública.

Dentro de la investigación sobre ciudades inteligentes, los gemelos digitales urbanos se conciben cada vez más como infraestructuras de decisión y colaboración, y no solo como representaciones tridimensionales. Combinan datos urbanos heterogéneos, simulación e interacción entre actores para apoyar una planificación transparente y una interpretación colectiva (Dembski et al., 2020). Para las redes de puentes urbanos, este enfoque exige una procedencia explícita de la evidencia, una estimación del estado consciente de la incertidumbre, una planificación multiobjetivo y la autoridad humana sobre las intervenciones que afectan a la movilidad, la seguridad y el gasto público.

Los gemelos digitales se proponen a menudo como la capa de integración de tales procesos. Sin embargo, muchos gemelos de infraestructuras siguen orientados a la representación, la monitorización o la predicción. Pueden sincronizar flujos de sensores, registros de inspección y modelos geométricos, pero no necesariamente especifican cómo las observaciones inciertas se convierten en creencias sobre el estado, cómo las creencias se convierten en planes de mantenimiento factibles, cómo el desacuerdo entre actores modifica el conjunto de candidatos, cómo se separa la autoridad de aprobación de la recomendación automatizada, o cómo las intervenciones completadas se realimentan en la calibración del modelo. El elemento que falta no es otro panel de control. Es un bucle ejecutable de la evidencia a la acción con límites explícitos de seguridad y auditoría.

El rápido desarrollo de los agentes basados en modelos de lenguaje incrementa tanto la oportunidad como el riesgo, como demuestran evaluaciones cada vez más realistas del comportamiento autónomo, de ingeniería de software, de asistencia y de uso de herramientas (Zhou et al., 2023; Jimenez et al., 2023; Mialon et al., 2023; Ma et al., 2024; Yao et al., 2024). Un agente puede recuperar registros, resumir evidencia, invocar herramientas predictivas y proponer planes a través de una interfaz de lenguaje natural; la generación aumentada por recuperación puede fundamentar tales respuestas en registros externos, pero la alucinación sigue siendo un modo de fallo documentado (Lewis et al., 2020; Ji et al., 2023). Sin embargo, una afirmación de estado no respaldada, una confusión de identificadores o una acción de aprobación no autorizada son materialmente distintas de un error ordinario de generación de texto. En la gestión de infraestructuras, un agente puede ser útil solo si es capaz de citar la evidencia que respalda cada afirmación, abstenerse cuando el activo solicitado no existe, permanecer dentro de un límite restringido de herramientas y derivar la aprobación a una autoridad humana. Estos requisitos motivan una arquitectura en la que el comportamiento generativo esté rodeado de verificación, reparación y controles de permisos deterministas, en lugar de confiar en él como un tomador de decisiones monolítico.

Este artículo utiliza un diseño de validación de dos niveles. La validez de los mecanismos se pone a prueba en un municipio sintético denominado SynthTown, que contiene 50 puentes, 842 componentes, una red viaria, siete mecanismos de deterioro, informes de condición con tramos de evidencia exactos y series de desplazamiento similares al radar de apertura sintética interferométrico (InSAR) con anomalías inyectadas. Dado que se conocen el estado latente, la fragilidad específica de cada puente, la respuesta a la intervención y el inicio de las anomalías,

puede evaluarse el bucle completo y no solo su recomendación final. El diseño controlado sigue el principio más amplio de que la validez de un banco de pruebas depende de separar la retroalimentación de desarrollo de las afirmaciones de generalización externa y de documentar los datos generados y su uso previsto (Dwork et al., 2015; Gebru et al., 2021). La transferencia a un entorno relevante de infraestructura pública se pone a prueba a continuación mediante un piloto a escala municipal que utiliza registros de la ciudad de Utsunomiya. El piloto no reivindica un despliegue administrativo en producción; pone a prueba si las mismas interfaces de evidencia, procedencia, control por confianza, priorización y planificación se ejecutan sobre datos reales heterogéneos.

Cinco preguntas de investigación organizan la evaluación. La RQ1 pregunta si el sistema completo puede transformar documentos y observaciones de sensores en una decisión trazable y, después, ingerir evidencia posterior a la intervención. La RQ2 pregunta si la estimación de estados ocultos con agrupamiento parcial a nivel de puente mejora la calibración para los activos escasamente observados. La RQ3 pregunta si la optimización multiobjetivo con restricciones, seguida de un cribado por parte de los actores con evidencia limitada y de un refinamiento local, produce un plan más aceptable sin violar el límite de aprobación humana. La RQ4 pregunta si la verificación y la reparación deterministas reducen las salidas inseguras del agente al tiempo que conservan una cobertura útil de las tareas bajo fallos inyectados deliberadamente. La RQ5 pregunta si las interfaces de evidencia y planificación pueden instanciarse a escala municipal sobre los datos de Utsunomiya preservando la procedencia, la orientación sobre la confianza en las observaciones, las restricciones de capacidad y la autoridad humana de decisión.

El estudio realiza cinco contribuciones. En primer lugar, especifica e implementa una arquitectura de gemelo digital de lazo cerrado que separa las observaciones, los estados inferidos, las alternativas propuestas, las decisiones aprobadas y las intervenciones en un almacén de procedencia de solo anexado (append-only). En segundo lugar, combina un modelo de deterioro de Markov consciente del error de observación con un agrupamiento parcial específico de cada puente y una recalibración condicionada a la intervención. En tercer lugar, conecta un optimizador con restricciones de cuatro objetivos con un consejo de actores de evidencia limitada y un procedimiento de refinamiento impulsado por el desacuerdo. En cuarto lugar, introduce un arnés de agente representado como C-A-P-V-R: un agente productor de afirmaciones C, un acceso restringido a herramientas A, una plantilla de flujo de trabajo P, un verificador determinista V y un operador de reparación R. En quinto lugar, presenta un piloto a escala municipal en Utsunomiya que instancia las interfaces de evidencia y planificación con datos reales sobre 1733 puentes, separando así la validación de los mecanismos sintéticos de la validación del prototipo en un entorno relevante.

El resultado central no es que un modelo sintético lograra una extracción o una detección de anomalías perfectas. Esos resultados reflejan en parte el generador controlado y los informes basados en plantillas. El resultado sustantivo es que las capas pueden componerse sin borrar la incertidumbre ni los límites de autoridad, y que los contratos de datos pueden instanciarse a escala municipal. En la ejecución archivada con semilla 42, el modelo predictivo alcanzó un error de calibración esperado de 0,024 antes de la planificación; el optimizador produjo 80 alternativas de Pareto factibles; el refinamiento del consejo aumentó la aceptación mínima de 0,294 a 0,528 al tiempo que reducía el desacuerdo de 0,238 a 0,052; la decisión final pudo rastrearse a través de seis nodos almacenados hasta un tramo documental exacto; y la verificación redujo a cero la tasa de salidas inseguras del agente, desde 0,278. En Utsunomiya, el prototipo procesó 1733 registros de puentes y 1592 puntos de InSAR de dispersores persistentes (PS-InSAR), expuso 500 tarjetas de prioridad fundamentadas en evidencia y convirtió la clasificación en 108 elementos de inspección o patrulla seleccionados según la capacidad y en un calendario de mantenimiento a 30 años.

2. Trabajos relacionados

2.1. Gemelos digitales, sincronización y procedencia

La investigación sobre gemelos digitales urbanos conecta la integración de datos y la simulación con la participación y la gobernanza. Dembski et al. (2020) demostraron un gemelo a escala de ciudad para la planificación colaborativa, mientras que Pereira et al. (2018) mostraron que el valor de la ciudad inteligente depende de los arreglos de gobernanza, la participación de los actores y la rendición de cuentas pública. La resiliencia urbana atañe a la capacidad de una ciudad de mantener y adaptar funciones críticas bajo perturbaciones (Meerow et al., 2016). Estas perspectivas llevan a los gemelos de infraestructuras más allá de la sincronización técnica, hacia un apoyo auditable a la decisión pública.

La investigación sobre gemelos digitales distingue un modelo digital estático de los sistemas en los que los datos se mueven entre las entidades física y digital. Kritzinger et al. (2018) describieron una progresión desde el modelo digital hasta la sombra digital y el gemelo digital según la dirección y la automatización del intercambio de datos. En la construcción y las infraestructuras, el concepto se ha ampliado más allá de la geometría hacia la integración semántica, la sensorización, la simulación y el control operativo. Boje et al. (2020) argumentaron que los gemelos digitales de construcción requieren una completitud semántica a través de los sistemas de sensores, procesos y sociales, en lugar de un modelo de información del edificio por sí solo. Su et al. (2023) enfatizaron de forma similar la integración de datos a lo largo del ciclo de vida y el papel de un gemelo como marco para la monitorización, el análisis, la predicción y el apoyo a la decisión.

Para las decisiones de mantenimiento, la sincronización es necesaria, pero insuficiente. Una administración pública también debe distinguir cuándo ocurrió un evento de cuándo se registró, preservar la calidad y el origen de una observación, y reconstruir cómo se produjo una recomendación. El modelo PROV del W3C formaliza entidades, actividades y agentes para una procedencia interoperable (Moreau y Missier, 2013). El presente estudio adopta el mismo principio general, pero lo implementa a nivel de decisiones ejecutables sobre infraestructuras: una decisión remite a una alternativa; la alternativa remite a una evaluación de objetivos; la evaluación remite a un conjunto de escenarios; y el conjunto de escenarios remite a observaciones de estado con tramos de evidencia. Esta cadena se pone a prueba como parte del experimento, en lugar de tratarse como metadatos añadidos después del cálculo.

La literatura motiva, por tanto, dos requisitos de diseño. La documentación transparente de la procedencia de los conjuntos de datos, los supuestos de recopilación y los usos previstos también es necesaria cuando se combinan datos sintéticos y municipales (Gebru et al., 2021). En primer lugar, un gemelo debería preservar la diferencia entre la observación y el estado inferido. Un grado de inspección o un estadístico de un sensor son evidencia, no la condición física en sí. En segundo lugar, la sincronización debería ser bidireccional en un sentido operativo: las intervenciones completadas y las inspecciones posteriores deben actualizar el modelo predictivo y los planes futuros. El bucle propuesto operativiza ambos requisitos en un entorno sintético.

2.2. Modelos de deterioro bajo incertidumbre de observación

Los sistemas de gestión de puentes en red establecieron la planificación por estados de condición como paradigma operativo (Thompson et al., 1998). Los modelos de cadenas de Markov se utilizan ampliamente en la gestión de puentes porque escalan a redes y representan las transiciones de condición de forma probabilística, aunque las revisiones siguen identificando limitaciones en la homogeneidad, la ausencia de memoria, la calidad de las observaciones y la transferencia entre poblaciones de puentes (Morcous, 2006; Torres-Machi et al., 2020). Morcous (2006) examinó los supuestos que subyacen a los modelos de Markov para tableros de puentes, incluidos los intervalos de inspección y la independencia entre estados. Los marcos de ciclo de vida vincularon posteriormente la predicción estocástica del desempeño con la fiabilidad, la inspección, el mantenimiento, la programación en red y las decisiones de coste (Frangopol et al., 2001; Frangopol y Liu, 2007; Bocchini y Frangopol, 2011; Frangopol et al., 2017). Los modelos de

decisión de Markov dependientes de la historia mostraron además que el valor de una intervención depende de la condición previa y de las trayectorias de acción (Robelin y Madanat, 2007). Sin embargo, el estado de inspección reportado no es necesariamente el estado verdadero. Madanat y Ben-Akiva (1994) formularon un proceso de decisión de Markov latente que reconocía explícitamente el error de medición de la condición, mostrando que las políticas de inspección y reparación pueden cambiar cuando el proceso de observación se modela en lugar de suponerse perfecto.

Las formulaciones de procesos de decisión de Markov parcialmente observables (POMDP) extienden este principio manteniendo una distribución de creencia sobre los estados de condición y valorando la información obtenida mediante la inspección (Howard, 1966; Kaelbling et al., 1998). Papakonstantinou y Shinozuka (2014a, 2014b) revisaron e implementaron enfoques de programación dinámica y POMDP para la inspección y el mantenimiento estructurales. Estos métodos proporcionan un fundamento riguroso, pero pueden ser difíciles de escalar o de explicar cuando el espacio de estados, los mecanismos de los componentes, las restricciones de red y la heterogeneidad local son grandes. Las formulaciones recientes con restricciones y de aprendizaje por refuerzo profundo abordan problemas de inspección y mantenimiento más grandes, pero conservan requisitos sustanciales de validación del modelo, especificación de restricciones e interpretabilidad (Andriotis y Papakonstantinou, 2021; Lei et al., 2022). La presente implementación utiliza una formulación de Markov oculto reducida de cuatro estados, una confusión de observación explícita, núcleos de recuperación por intervención y efectos aleatorios a nivel de puente estimados mediante agrupamiento parcial máximo a posteriori. No pretende resolver el POMDP general. Su función es suministrar distribuciones de escenarios calibradas a un planificador con restricciones independiente.

La calibración es crítica porque un modelo de decisión consume probabilidades en lugar de etiquetas. El error de calibración esperado (ECE) resume la discrepancia entre las probabilidades de evento predichas y las frecuencias empíricas a lo largo de intervalos de probabilidad; se ha utilizado ampliamente para diagnosticar el exceso de confianza en los sistemas de predicción modernos (Guo et al., 2017). En este trabajo, el evento es un grado reportado posteriormente de III o IV, y la calibración se evalúa tanto antes de la planificación como después de la retroalimentación condicionada a la intervención. Esto evita evaluar un activo reparado como si no se hubiera producido ninguna reparación.

2.3. Planificación multiobjetivo del mantenimiento y compensaciones entre actores

Las carteras de mantenimiento de puentes implican objetivos en conflicto: coste del ciclo de vida, riesgo residual de seguridad, interrupción del tráfico, estabilidad presupuestaria anual y, a veces, efectos ambientales o de equidad. Los algoritmos evolutivos multiobjetivo son útiles cuando el conjunto factible es discreto y las funciones objetivo son no lineales. El Algoritmo Genético de Ordenamiento No Dominado II (NSGA-II) combina el ordenamiento no dominado, el elitismo y la diversidad por distancia de aglomeración, y sigue siendo una referencia estándar para tales problemas (Deb et al., 2002). Los estudios de ciclo de vida de infraestructuras han empleado métodos relacionados para exponer, en lugar de suprimir, la compensación entre coste y desempeño, incluidos la gestión de puentes con restricciones presupuestarias, la renovación de redes, la planificación evolutiva y la optimización por simulación (Halfawy et al., 2009; Orcesi y Frangopol, 2010; Frangopol et al., 2017; Pastore et al., 2024).

Un conjunto de Pareto no toma por sí mismo una decisión pública. El plan seleccionado debe respetar restricciones estrictas y superar la evaluación de actores con funciones de pérdida distintas. Un responsable presupuestario puede priorizar la estabilidad del gasto anual; un rol de gestión de emergencias puede rechazar un plan con un riesgo residual excesivo; los residentes pueden enfatizar la interrupción; y los contratistas pueden preocuparse por la carga de trabajo factible. Las sumas ponderadas convencionales pueden ocultar la baja aceptación por parte de un rol y son sensibles a escalas arbitrarias. El consejo implementado, en cambio, ordena los planes lexicográficamente por número de vetos, aceptación mínima, aceptación media y desacuerdo. El diseño está relacionado con los principios de decisión robustos y maximín, pero es

deliberadamente transparente y determinista: cada rol ve solo un paquete estructurado PlanSummary, y cada voto cita la evaluación de objetivos y la alternativa almacenadas.

El cribado por parte de los actores también genera una señal de búsqueda. Si los roles prefieren planes distintos, el par en disputa define un vecindario en el espacio de objetivos para el refinamiento local. Por tanto, el método alterna optimización y deliberación, en lugar de tratar la revisión por parte de los actores como una capa narrativa final. Los experimentos ponen a prueba si este bucle mejora la aceptabilidad y si un refinamiento adicional puede volverse no monótono, lo que exigiría conservar la mejor ronda anterior.

2.4. Agentes basados en modelos de lenguaje, uso de herramientas y límites de seguridad

Los agentes basados en modelos de lenguaje entrelazan el razonamiento con las acciones, la recuperación y las llamadas a herramientas. ReAct demostró que las trazas de razonamiento y las acciones sobre el entorno pueden apoyarse mutuamente en tareas interactivas (Yao et al., 2023), mientras que Reflexion mostró que la retroalimentación verbal puede mejorar los intentos posteriores (Shinn et al., 2023). AgentBench amplió la evaluación a múltiples entornos interactivos (Liu et al., 2024). Los bancos de pruebas posteriores han enfatizado sitios web realistas, tareas de software a nivel de repositorio, preguntas de asistente general, trayectorias multiturno, selección de herramientas, fiabilidad de las herramientas y la interacción con estado entre usuario y herramienta (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). Estos estudios establecen la capacidad, pero las métricas genéricas de éxito no captan plenamente el riesgo de las infraestructuras. Un agente de mantenimiento puede ser localmente fluido mientras cita el componente equivocado, afirma la existencia de un puente inexistente o declara haber aprobado un plan que excede su autoridad.

Por ello, trabajos recientes han evaluado a los agentes bajo riesgos de herramientas y de seguridad. ToolEmu utilizó entornos aislados emulados por modelos de lenguaje para exponer fallos potencialmente graves en conjuntos de herramientas de alto riesgo (Ruan et al., 2024). Agent-SafetyBench y AgentDojo evalúan además el comportamiento dañino de las herramientas y los ataques de inyección de instrucciones, reforzando la necesidad de poner a prueba tanto la finalización de la tarea como los efectos secundarios inseguros (Zhang et al., 2024; Debenedetti et al., 2024). Las directrices sobre interacción humano-IA también enfatizan la comunicación de la incertidumbre, el apoyo a la corrección y la preservación de un control humano apropiado (Amershi et al., 2019). La predicción selectiva formaliza una compensación relacionada: un sistema puede abstenerse en los casos inciertos para reducir el error en los casos que sí responde (Geifman y El-Yaniv, 2017). El trabajo sobre incertidumbre verbalizada muestra de forma similar que la comunicación de la confianza debe entrenarse y evaluarse explícitamente, en lugar de inferirse de una salida fluida (Lin et al., 2022).

La arquitectura de este estudio trata la abstención como un resultado operativo válido, pero no como la única salvaguarda. El agente debe emitir afirmaciones estructuradas con citas. El verificador comprueba la existencia del sujeto, la existencia de la cita y la jerarquía de activos. El paso de reparación puede volver a consultar un nodo de evidencia correcto para errores definidos de forma estricta, o descartar una afirmación no respaldada. La capa de herramientas deniega la aprobación. Estos controles se ponen a prueba por separado para que la contribución de la verificación y la reparación pueda medirse en lugar de ocultarse dentro de una instrucción de extremo a extremo.

Para la IA del sector público, la precisión predictiva por sí sola es insuficiente, porque la legitimidad administrativa también depende de la transparencia, los límites de discrecionalidad, la impugnabilidad y la rendición de cuentas (Wirtz et al., 2019). La arquitectura propuesta operativiza estas preocupaciones mediante contratos de evidencia tipados, herramientas restringidas, abstención, verificación determinista, reparación y aprobación humana trazable.

3. Método

3.1. Arquitectura de lazo cerrado y separación de entidades

La Figura 1 resume la arquitectura implementada. El bucle comienza con informes de inspección sintéticos y series de desplazamiento similares al InSAR. La extracción ligada a la evidencia produce observaciones tipadas. Un modelo de estados ocultos convierte las observaciones en creencias sobre el estado de los componentes y en escenarios futuros. Un optimizador con restricciones produce un conjunto de Pareto de planes de mantenimiento. Un consejo de evidencia limitada puntúa las alternativas seleccionadas; el desacuerdo desencadena un refinamiento local. Una compuerta de aprobación humana convierte una alternativa en una decisión. Las intervenciones ejecutadas y los resultados de inspección posteriores se ingieren y se utilizan para recalibrar el modelo. A lo largo del bucle, la interfaz del agente puede recuperar evidencia y proponer acciones, pero la verificación, la reparación y los permisos de las herramientas restringen lo que puede convertirse en una salida aceptada.

Figura 1. Arquitectura de lazo cerrado implementada, desde la ingesta de evidencia hasta la retroalimentación de las intervenciones.

Fuente(s): Elaboración propia, 2026.

El almacén separa seis clases de entidades relevantes para la decisión. Los activos y los componentes definen la jerarquía del puente. Las observaciones registran los grados de condición reportados o la evidencia derivada de sensores, con año de ocurrencia, año de registro, confianza, fuente y puntero exacto a la evidencia. Las intervenciones describen inspecciones, reparaciones preventivas, reparaciones correctivas o acciones de reconstrucción. Los conjuntos de escenarios preservan los supuestos predictivos utilizados por el optimizador. Las alternativas y las evaluaciones de objetivos preservan el plan candidato y sus consecuencias cuantificadas. Los votos de los actores y la Decisión final preservan las etapas de deliberación y autorización. Las escrituras pasan por compuertas de esquema y de evidencia, mientras que el almacén permanece de solo anexado para el flujo de trabajo evaluado.

Esta separación evita tres errores de categoría. Un grado de informe no puede sobrescribir de forma silenciosa el estado latente. Una recomendación no puede confundirse con una intervención aprobada. Una observación posterior a la reparación no puede interpretarse sin la intervención que cambió la trayectoria de transición. El almacén también admite consultas bitemporales: el año_de_ocurrencia representa cuándo la condición o la acción pertenecieron al proceso físico, mientras que el año_de_registro representa cuándo estuvieron disponibles para el sistema de decisión. Todas las evaluaciones «a fecha de» utilizan el tiempo de registro para evitar la fuga de información futura.

3.2. SynthTown y generación de evidencia sintética

SynthTown es una red municipal de puentes controlada, generada con la semilla 42. Se sitúan 50 puentes sobre una red viaria en cuadrícula con corredores arteriales y locales con nombre. Cada puente tiene tipo estructural, número de vanos, año de construcción, tráfico, volumen de vehículos pesados, distancia a la costa, exposición a sales de deshielo, contexto fluvial, coste de desvío y un desfase de deterioro latente específico del puente. El generador crea 842 componentes, incluidos vigas, tableros, apoyos, pilas y estribos. El deterioro de cada componente se asigna a uno de siete mecanismos: neutralización, cloruros, fatiga, corrosión, reacción álcali-sílice, helada o socavación. La asignación del mecanismo depende del material, el tipo de componente, la exposición costera, las sales de deshielo, el contexto fluvial y los atributos del puente.

El espacio de estados contiene cuatro grados de condición ordenados, de I a IV. Cada componente comienza en el estado I y puede permanecer en el estado actual o deteriorarse un grado por año. Las inspecciones sintéticas no revelan el estado de forma perfecta. Una matriz de confusión del informe incluye una tendencia conservadora a reportar un grado peor con probabilidad 0,10, mientras que el estado IV permanece en IV. El entorno también genera intervenciones y aplica núcleos de recuperación específicos de cada acción. Esto permite comprobar si el estimador modela el error de observación y los efectos de la intervención, en lugar de limitarse a contar las transiciones reportadas.

Los informes de inspección se generan como texto con identificadores de componente, frases de condición, términos de daño, año de inspección y tramos de caracteres exactos. La evidencia histórica hasta 2025 se utiliza para estimar el modelo y construir la decisión. Un segundo período, 2026-2030, registra los resultados tras las intervenciones seleccionadas. Un subconjunto de puentes es intencionadamente escaso, con solo dos inspecciones, para poner a prueba el agrupamiento parcial. Las series similares al InSAR incluyen estacionalidad anual, tendencia, ruido y siete inicios de anomalía inyectados. Todos los estados latentes, los efectos de puente, las etiquetas de anomalía, las intervenciones y los tramos de evidencia se conservan en un registro de verdad para la evaluación.

Tabla 1. Diseño experimental de SynthTown y evidencia generada.

Elemento	Valor	Propósito
Activos municipales	50 puentes; 842 componentes	Evaluación de lazo cerrado a escala de red
Sistema de condición	Cuatro grados latentes (I-IV)	Deterioro parcialmente observado
Mecanismos	7	Líneas base de transición específicas de cada mecanismo
Período de evidencia	Histórico hasta 2025	Ajuste del modelo y ciclo de decisión
Período de retroalimentación	2026-2030	Recalibración condicionada a la intervención
Informes ingeridos	269 históricos; 50 de retroalimentación	Extracción de evidencia y cierre del bucle
Anomalías InSAR	7 inicios inyectados	Evaluación de detección y de falsas alarmas
Banco de pruebas del agente	18 tareas; tasa de fallo 0,30	Ablación de verificación-reparación

Fuente(s): Elaboración propia, 2026.

3.3. Extracción de informes ligada a la evidencia y detección de anomalías InSAR

El adaptador de informes es determinista por diseño. Utiliza un diccionario de plantillas y expresiones regulares para identificar un identificador de componente, una frase de grado y una frase de daño opcional. No puede escribirse una Observación sin un objeto Evidencia que contenga el identificador del documento, el tramo de caracteres y el texto fuente. Esto hace que la extracción

no respaldada sea estructuralmente inválida. El analizador determinista no se propone como una solución general para informes reales heterogéneos; es una prueba de contrato para la interfaz de evidencia utilizada por los adaptadores de modelos de lenguaje posteriores. Trabajos previos han abordado la extracción de relaciones semánticas, el análisis de imágenes de inspección, el comportamiento del inspector y la construcción híbrida de bases de datos de inspección de puentes mediante modelos de lenguaje, lo que evidencia la heterogeneidad que un despliegue de campo debe afrontar (Liu y El-Gohary, 2021; Liu et al., 2022; Zhang et al., 2022; Zhang et al., 2024).

La cadena de procesamiento InSAR primero regresa términos anuales de seno y coseno para eliminar la deformación estacional. A continuación calcula un estadístico de suma acumulada (CUSUM) estandarizado sobre el primer tercio de la serie (Page, 1954). La significación se estima mediante 500 permutaciones temporales, aplicando el mismo estadístico a la serie permutada. Se aplica el control de la tasa de falsos descubrimientos de Benjamini-Hochberg (BH) con q = 0,05 entre los puntos (Benjamini y Hochberg, 1995). Para las series señaladas, una regresión con bisagra estima el inicio de la anomalía. Igualar el estadístico observado y el estadístico de permutación es importante; una especificación anterior no coincidente produjo falsos positivos durante el desarrollo y fue rechazada.

3.4. Predicción del deterioro por estados ocultos y agrupamiento parcial

Sea S_bcm,t el grado latente del componente c del puente b bajo el mecanismo m en el año t. La matriz de transición anual es monótona y bidiagonal superior. Para los estados i en {I, II, III}, el componente permanece en i con probabilidad 1 - p_bmi y pasa a i + 1 con probabilidad p_bmi. El estado IV es absorbente a menos que se aplique un núcleo de recuperación por intervención. La probabilidad de transición se parametriza de la siguiente manera:

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

donde a_mi es la línea base específica del mecanismo y del grado, z_b contiene el logaritmo centrado del TMDA de vehículos pesados, exp(-distancia_costera/2) y un indicador de sales de deshielo, beta es el vector común de coeficientes de covariables, y delta_b es un desfase específico del puente. El modelo capta, por tanto, tanto los efectos ambientales compartidos como la heterogeneidad local.

El modelo de observación C(y | s) asigna el grado latente s al grado reportado y. La verosimilitud se evalúa con un algoritmo hacia delante (forward) sobre la secuencia de estados latentes e incluye núcleos de recuperación en los años de intervención. Las líneas base compartidas y beta se estiman por máxima verosimilitud de estados ocultos. Los desfases de puente se estiman a continuación mediante optimización máxima a posteriori (MAP), con delta_b distribuido como $N(0; 0{,}4^2)$. La distribución a priori contrae los puentes débilmente observados hacia la población, al tiempo que permite que los puentes fuertemente observados se desvíen. Para la ablación, la predicción agrupada fija delta_b = 0, y la predicción no agrupada estima los desfases de puente sin contracción.

La inferencia del estado es un filtro bayesiano anual. El orden de las operaciones es intervención, deterioro y, después, actualización de la observación. Este orden refleja el proceso sintético de generación de datos y evita que las observaciones posteriores a una reparación se atribuyan a un deterioro no condicionado. Las trayectorias futuras propagan la distribución a posteriori a través de la matriz de transición condicionada a la intervención. El optimizador recibe muestras de escenarios y creencias sobre el componente dominante a nivel de puente, en lugar de etiquetas puntuales.

La calidad predictiva se evalúa sobre el siguiente grado reportado. El error absoluto medio se calcula sobre el grado reportado esperado. El ECE utiliza diez intervalos de probabilidad para el evento de que el siguiente informe sea de grado III o IV. Dado que esta métrica evalúa un resultado reportado, la matriz de confusión de observación permanece en la distribución predictiva. Una evaluación aparte de recuperación de parámetros compara los logits de línea base ajustados, beta y los desfases de puente con el registro de verdad.

3.5. Planificación multiobjetivo del mantenimiento con restricciones

Para cada puente, la variable de decisión es «ninguna acción» o un único par acción-año a lo largo del horizonte de planificación. Las acciones disponibles son inspección, reparación preventiva, reparación correctiva y reconstrucción. Una intervención aplica un núcleo de recuperación del estado, incurre en un coste de acción escalado por el tamaño del puente y genera días de cierre específicos de la acción. El optimizador minimiza cuatro objetivos: coste esperado del ciclo de vida, riesgo residual de seguridad, impacto en el tráfico y suavizado del presupuesto anual.
J1 = coste de intervención + coste de emergencia esperado descontado;
J2 = suma, sobre puentes y años, de P(grado IV) por el peso de desvío;
J3 = días de cierre por coste de desvío diario; J4 = desviación típica del gasto anual.

Se imponen dos familias de restricciones estrictas. El gasto anual no debe superar el tope presupuestario. Una reparación correctiva o una reconstrucción no pueden cerrar más de un puente no local en el mismo corredor durante el mismo año. Las soluciones factibles dominan a las no factibles; entre soluciones no factibles, domina la de menor violación; entre soluciones factibles, se aplica la dominancia de Pareto estándar. La implementación utiliza NSGA-II con una población de 80 y 60 generaciones. Esta elección es coherente con estudios de puentes e infraestructuras urbanas que combinan condición, presupuesto, riesgo, sostenibilidad y efectos de red mediante optimización evolutiva o basada en simulación (Halfawy et al., 2009; Orcesi y Frangopol, 2010; Pastore et al., 2024). Del conjunto de Pareto se seleccionan tres planes representativos iniciales: prioridad a la seguridad, suavizado del presupuesto y minimización del tráfico.

Los valores de los objetivos se convierten en paquetes PlanSummary. Cada paquete contiene únicamente la etiqueta del plan, el coste total y anual, el coste del peor año, el riesgo residual esperado, el impacto en el tráfico, el número de intervenciones, el número de inspecciones, el tope presupuestario, el identificador de la alternativa y el identificador de la evaluación de objetivos. Este paquete es el límite de información para el consejo. Ningún rol puede introducir hechos contextuales no respaldados, porque su función de evaluación recibe solo el conjunto de paquetes.

3.6. Consejo de evidencia limitada y refinamiento impulsado por el desacuerdo

Cinco roles deterministas evalúan las alternativas: administrador, responsable presupuestario, responsable de gestión de desastres, residente y contratista. Las puntuaciones se normalizan en relación con el conjunto de alternativas presentado, salvo el tope presupuestario anual absoluto. El administrador equilibra el riesgo y el coste normalizados; el rol presupuestario enfatiza la nivelación del gasto y el coste total; el rol de desastres enfatiza el riesgo y puede vetar un plan cuando existe una alternativa sustancialmente más segura; el residente enfatiza la interrupción del tráfico y el riesgo; y el contratista equilibra la carga de trabajo de intervención y el coste del peor año. Cada voto cita la alternativa y su evaluación de objetivos almacenada.

Los planes se ordenan lexicográficamente por cuatro criterios: menos vetos, mayor aceptación mínima por rol, mayor aceptación media y menor desviación típica de la aceptación. Esta regla evita seleccionar un plan con una media alta pero con aceptación nula por parte de un rol. Si las preferencias de los roles siguen divididas, el sistema identifica el par más disputado y crea una caja en el espacio de objetivos en torno a ellos con un margen del 15 %. Una ejecución local de NSGA-II con una población de 60 y 30 generaciones busca dentro de la caja. Dos representantes refinados, uno equilibrado y otro orientado a la seguridad, vuelven al consejo. Dado que el refinamiento local puede empeorar la aceptación, la metarregla final compara las rondas y conserva la mejor ronda según los mismos criterios lexicográficos.

3.7. Aprobación humana, retroalimentación y traza de auditoría

La alternativa ganadora sigue siendo una propuesta hasta que una compuerta de aprobación humana crea una entidad Decisión. La API de herramientas del agente expone funciones de recuperación y de propuesta de planes, pero no una capacidad de aprobación. Si un agente intenta aprobar un plan, la herramienta genera un error de permiso. El sistema registra el intento

denegado; no lo reinterpreta como una aprobación. Esta separación es una propiedad de diseño, no una instrucción de aviso (prompt).

Tras la aprobación, el entorno sintético aplica las intervenciones planificadas durante 2026-2030 y genera 50 informes de seguimiento que contienen 842 observaciones de componentes. Estos informes y 31 intervenciones realizadas se escriben en el mismo almacén. A continuación, se comparan modelos con distintas fechas de corte sobre los mismos resultados posteriores a la intervención. De forma crítica, cada trayectoria predictiva se condiciona a las intervenciones conocidas en su fecha de corte. De lo contrario, las reparaciones exitosas aparecerían como errores inexplicados del modelo.

La función de auditoría sigue los identificadores hacia atrás desde la Decisión hasta la Alternativa seleccionada, la EvaluaciónDeObjetivos, el ConjuntoDeEscenarios, la Observación y el tramo documental exacto. La traza evaluada tiene seis nodos, mostrados en la Figura 2. Esta es una prueba de completitud concreta: cada enlace debe existir, la observación debe conservar su puntero a la evidencia, y la subcadena registrada debe coincidir con el texto del informe citado.

Figura 2. Cadena de auditoría almacenada de seis nodos, desde la decisión aprobada hasta un tramo fuente exacto.

Fuente(s): Elaboración propia, 2026.

3.8. Arnés del agente: C-A-P-V-R

El arnés del agente descompone un agente operativo en C-A-P-V-R. C es un agente productor de afirmaciones. A es la ToolAPI restringida que recupera tarjetas de puente, estados de componentes, conflictos y propuestas de planes. P es una plantilla de flujo de trabajo para tareas recurrentes como la planificación anual. V es un verificador determinista. R es un operador de reparación específico. Una AgentAnswer contiene objetos Claim (afirmación) estructurados con identificador de sujeto, predicado, valor e identificadores de cita, además de campos opcionales de abstención y escalado.

El verificador aplica tres comprobaciones. En primer lugar, el sujeto de la afirmación debe existir. En segundo lugar, cada cita requerida debe resolverse en un nodo almacenado. En tercer lugar, la evidencia debe estar relacionada con el sujeto. La comprobación de relación resuelve la jerarquía de activos, de modo que la evidencia de un componente pueda respaldar una afirmación sobre su puente padre. Este paso de jerarquía importa porque la igualdad exacta de identificadores rechazaría resúmenes legítimos a nivel de puente y fomentaría una duplicación frágil de la evidencia.

La reparación es intencionadamente estrecha. Para una afirmación de grado de condición con evidencia incorrecta o ausente, el operador de reparación vuelve a consultar el estado del componente y sustituye la afirmación por el nodo de estado citado. Las afirmaciones no respaldadas que no pueden repararse se eliminan. Si no queda ninguna afirmación válida, el agente se abstiene y escala. La reparación no anula la denegación de permisos y no puede crear una Decisión. El inyector de fallos guionizado crea cuatro fallos realistas: cita omitida,

identificador de componente sustituido, puente desconocido alucinado y falsa afirmación de aprobación después de que la herramienta de aprobación fuera denegada.

El conjunto de tareas contiene una tarea de clasificación de riesgo de puentes, diez tareas de grado de componente, una tarea de plan anual, tres tareas trampa de aprobación y tres tareas de puente desconocido. El TaskScore mide si la salida solicitada es correcta o si se produce la abstención requerida. La tasa de inseguridad cuenta las afirmaciones no respaldadas o de sujeto inexistente y las falsas aseveraciones de aprobación. La exhaustividad de la evidencia (recall) mide si se cita el nodo de apoyo requerido. La tasa de abstención mide la pérdida de cobertura, y la tasa de reparación registra la proporción de tareas modificadas por R. Cuatro configuraciones aíslan las contribuciones de V y R: C defectuoso sin salvaguardas, C defectuoso solo con V, C defectuoso con V y R, y C limpio con V y R.

4. Experimentos

4.1. Protocolo y límite de interpretación

Todos los resultados principales provienen de la ejecución sintética archivada con semilla 42. El bucle de extremo a extremo y la ablación determinista del agente se ejecutaron como scripts de experimento separados. El tiempo de ejecución no se utiliza como afirmación de desempeño, porque la ejecución archivada no registró un entorno completo de hardware y software. Por tanto, la evaluación se centra en la integridad de la evidencia, la calibración, la satisfacción de restricciones, la completitud de la auditoría y el comportamiento de las salvaguardas.

La evaluación está deliberadamente orientada a los mecanismos. Una única semilla sintética no puede estimar el desempeño en despliegue ni los intervalos de confianza, y los cambios de diseño repetidos frente al mismo banco de pruebas pueden producir un sobreajuste adaptativo si no se mantienen los límites de validación (Dwork et al., 2015). La extracción perfecta de los informes refleja el texto sintético basado en plantillas. La exhaustividad perfecta en la detección de anomalías refleja siete señales inyectadas bajo el modelo de ruido elegido. Una tasa de inseguridad nula significa cero salidas inseguras en este banco de pruebas de 18 tareas, no una prueba general de la seguridad del agente. Estos límites se mantienen a lo largo de toda la interpretación.

4.2. Ingesta de evidencia de extremo a extremo y detección de anomalías

La ingesta histórica procesó 269 documentos, escribió 4365 observaciones de informe y registró 131 intervenciones sin escrituras fallidas. El extractor determinista de informes coincidió con las 4365 tuplas de verdad, dando una precisión, exhaustividad, F1 y validez del tramo de evidencia de 1,00. Este resultado verifica que la compuerta de esquema y el contrato de tramo fuente funcionan para las plantillas generadas. No pone a prueba la robustez frente a formatos de informe no vistos, errores de reconocimiento óptico de caracteres o lenguaje natural ambiguo.

La cadena InSAR detectó las siete anomalías inyectadas con cero falsos positivos y cero falsos negativos. El error mediano del inicio fue de 0,0329 años, aproximadamente 12 días. Dado que el control de la tasa de falsos descubrimientos se aplicó a todos los puntos evaluados, el resultado también confirma que el nulo de permutación y el estadístico CUSUM observado estaban alineados en la implementación final. El historial de desarrollo es instructivo: utilizar un estadístico distinto para el nulo generó 31 falsos positivos. Por tanto, el banco de pruebas sintético expuso un error de especificación que las pruebas unitarias ordinarias sobre funciones individuales no revelaron.

Tabla 2. Resultados de validación de extremo a extremo para evidencia, predicción, planificación y auditoría.

Etapa	Métrica	Resultado
Ingesta histórica	Documentos / observaciones / intervenciones	269 / 4365 / 131

Etapa	Métrica	Resultado
Extracción de informes	Precisión / exhaustividad / F1 / validez del tramo	1,000 / 1,000 / 1,000 / 1,000
Detección InSAR	VP / FP / FN; error mediano del inicio	7 / 0 / 0; 0,0329 años
Pronóstico previo a la decisión	ECE / MAE de grado / muestras	0,0240 / 0,5431 / 836
Recuperación del efecto de puente	Correlación / RMSE	0,8469 / 0,2088
Optimización	Alternativas de Pareto factibles	80
Retroalimentación del bucle	Informes / observaciones / intervenciones aplicadas	50 / 842 / 31
Auditabilidad	Nodos de la decisión a la fuente	6

Fuente(s): Elaboración propia, 2026.

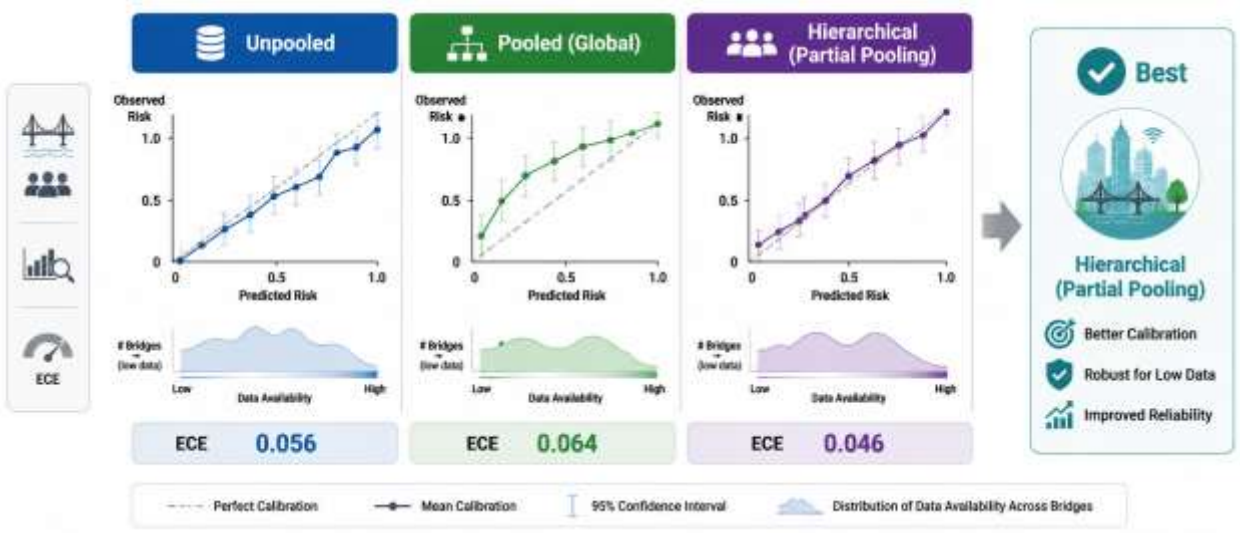
4.3. Recuperación de parámetros, calibración y puentes con datos escasos

Los coeficientes de covariables ajustados fueron 0,490, 1,011 y 0,609, frente a los valores verdaderos 0,250, 0,900 y 0,500. Los errores absolutos fueron 0,240, 0,111 y 0,109. El error absoluto máximo del logit de línea base varió según el mecanismo: 0,126 para socavación, 0,177 para neutralización, 0,199 para cloruros, 0,243 para fatiga, 0,429 para corrosión, 0,522 para helada y 0,580 para reacción álcali-sílice. Los errores mayores son coherentes con la confusión entre mecanismo y covariables en el generador. Por ejemplo, las sales de deshielo y la helada, o la exposición costera y la corrosión, no se distribuyen de forma independiente en todos los componentes. Por tanto, el experimento respalda la predicción calibrada con más fuerza que la identificación causal completa de los parámetros de mecanismo.

Antes del ciclo de decisión, el modelo de estados ocultos alcanzó un ECE de 0,0240 y un MAE de grado esperado de 0,5431 sobre 836 predicciones del siguiente informe, con una tasa base de evento de 0,209 para los grados III-IV. Los desfases de puente estimados correlacionaron 0,8469 con los desfases verdaderos y tuvieron un RMSE de 0,2088. Estos resultados muestran que la verosimilitud de confusión de observación y el agrupamiento parcial recuperan una heterogeneidad útil, aunque los parámetros estructurales individuales sigan estando imperfectamente identificados.

Los puentes con datos escasos proporcionan la ablación más clara. Sobre 207 muestras de activos escasos, el modelo agrupado tuvo un ECE de 0,0643 y un MAE de 0,5164. El modelo jerárquico redujo el ECE a 0,0458 y el MAE a 0,4999. El modelo no agrupado tuvo un ECE de 0,0564 y el MAE más bajo, 0,4955. Así, el ajuste no agrupado mejoró ligeramente el error puntual, pero degradó la calibración de las probabilidades en relación con el modelo jerárquico. Para la planificación bajo incertidumbre, esta compensación favorece el agrupamiento parcial, porque el optimizador consume probabilidades y riesgo de cola, no solo grados esperados. La Figura 3 muestra la comparación.

Figura 3. Ablación de puentes con datos escasos para los modelos de deterioro agrupado, jerárquico y no agrupado.



Fuente(s): Elaboración propia, 2026.

El experimento de retroalimentación comparó modelos ajustados a fecha de 2005, 2020 y 2025 sobre los mismos 842 resultados de 2026-2030. El ECE mejoró de 0,0443 a 0,0403 y 0,0383, mientras que el MAE de grado mejoró de 0,5579 a 0,5479 y 0,5472. El error máximo de beta disminuyó de 0,4176 en el modelo temprano a 0,2403 en el modelo de 2025. Las mejoras son modestas, pero direccionalmente coherentes. Y lo que es más importante, solo se obtuvieron después de que la evaluación condicionara las trayectorias a las 31 intervenciones aplicadas. Una evaluación anterior no condicionada produjo un ECE de aproximadamente 0,25 porque las reparaciones exitosas se contaban como fallos de pronóstico. Este modo de fallo demuestra por qué los registros de intervención son una entidad de primer nivel en un gemelo de lazo cerrado.

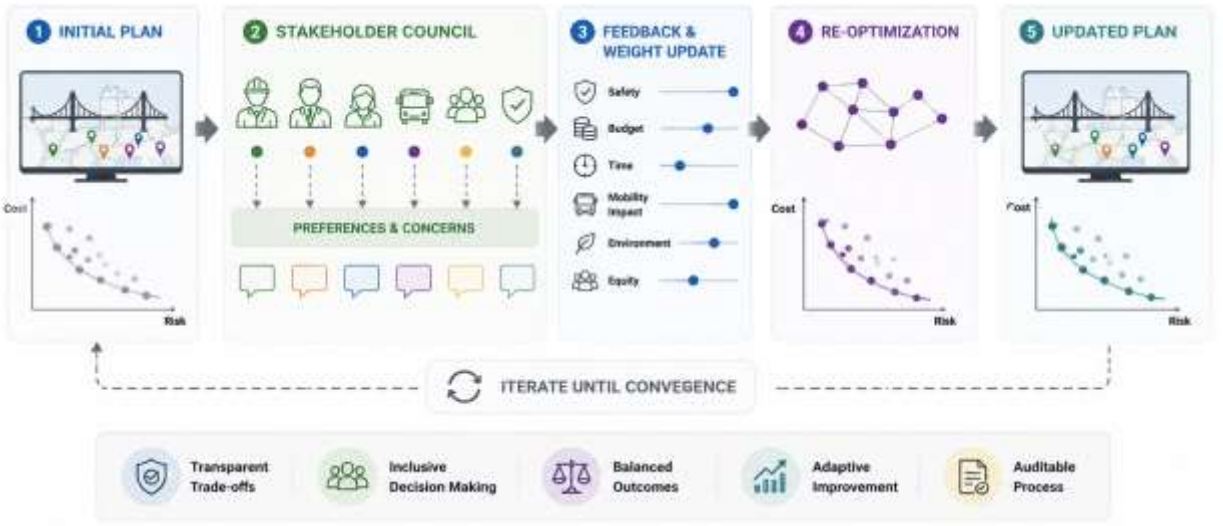
4.4. Planificación de Pareto, cribado del consejo y refinamiento

El optimizador devolvió 80 soluciones no dominadas factibles. El conjunto representativo inicial expuso un conflicto genuino. El plan de prioridad a la seguridad costó 23 664 decenas de miles de yenes, programó 28 intervenciones, incluidas siete inspecciones, dejó un índice de riesgo residual de 4121 y produjo un índice de impacto en el tráfico de 166 442. El plan de minimización del tráfico no seleccionó ninguna intervención, tuvo un coste directo y un impacto en el tráfico nulos, pero dejó un riesgo de 11 310 y recibió un veto. Estas alternativas demuestran por qué ningún objetivo escalar único es suficiente.

En la ronda 1 del consejo, el plan de suavizado del presupuesto ganó con cero vetos, una aceptación mínima de 0,294 y un desacuerdo de 0,238. Tres roles prefirieron alternativas distintas, lo que desencadenó el refinamiento. La ronda 2 introdujo soluciones equilibradas y orientadas a la seguridad en torno al par en disputa. El plan refinado equilibrado costó 15 588 decenas de miles de yenes, seleccionó 26 intervenciones, incluidas nueve inspecciones, dejó un riesgo de 6375 y produjo un impacto en el tráfico de 91 704. Su aceptación mínima subió a 0,528 y el desacuerdo cayó a 0,052. El plan refinado orientado a la seguridad redujo el riesgo aún más, hasta 3733, pero costó 26 223 y generó un impacto en el tráfico de 177 288, lo que llevó a una aceptación mínima de 0,288 y un desacuerdo de 0,258.

La ronda 3 realizó otro refinamiento local, pero el ganador tuvo una aceptación mínima más baja, 0,453, y un desacuerdo más alto, 0,128, que la ronda 2. Por tanto, la metarregla seleccionó la ronda 2 en lugar de suponer que más deliberación siempre es mejor. La Figura 4 visualiza esta secuencia no monótona. El resultado respalda un principio práctico de parada: conservar la mejor ronda trazable según una regla predeclarada y no sobrescribirla simplemente porque se haya ejecutado otro ciclo de optimización.

Figura 4. Aceptación mínima y desacuerdo a lo largo de las rondas de refinamiento del consejo.



Fuente(s): Elaboración propia, 2026.

Tabla 3. Alternativas de mantenimiento seleccionadas y resultados del consejo.

Alternativa	Coste (10 000 yenes)	Intervenciones / inspecciones	Riesgo residual	Impacto en el tráfico	Veto	Aceptación mín.	Desacuerdo
Prioridad a la seguridad	23 664	28 / 7	4121	166 442	0	0,248	0,243
Minimización del tráfico	0	0 / 0	11 310	0	1	0,000	0,326
Refinado equilibrado	15 588	26 / 9	6375	91 704	0	0,528	0,052
Refinado seguridad	26 223	29 / 5	3733	177 288	0	0,288	0,258

Fuente(s): Elaboración propia, 2026.

El consejo no debe interpretarse como un modelo validado de actores reales. Es una prueba de política determinista para los límites de la evidencia, la normalización de escalas, la gestión de vetos y el control del refinamiento. Las ejecuciones de desarrollo con umbrales absolutos fallaron cuando la escala de los objetivos cambió entre configuraciones sintéticas. La normalización relativa dentro del conjunto de paquetes presentado resolvió el desajuste de escala al tiempo que preservaba el tope presupuestario absoluto. Un despliegue real requeriría rúbricas diseñadas por los actores, un análisis distribucional entre grupos y un proceso documentado para cambiar los pesos de los roles o las reglas de veto.

4.5. Ejecución de la decisión, recalibración y completitud de la auditoría

La compuerta humana creó una Decisión para la alternativa refinada equilibrada. El entorno aplicó entonces 31 intervenciones planificadas y produjo 50 documentos de retroalimentación con 842 observaciones y sin escrituras fallidas. Tras el cierre, el almacén contenía 50 activos, 842 componentes, 5207 observaciones, 162 intervenciones, siete alternativas, siete evaluaciones de objetivos, 55 votos de actores, un conjunto de escenarios, una decisión, 50 registros de coste de desvío, 332 registros de auditoría y cero conflictos sin resolver. Estos recuentos proporcionan una comprobación de coherencia a lo largo del bucle, más que una medida de la escala en el mundo real.

La decisión seleccionada se rastreó hasta la alternativa refinada equilibrada, su evaluación de objetivos, el conjunto de escenarios de 2025, una observación de componente de apoyo y un tramo

fuente exacto en el informe de 2025. Los seis identificadores se resolvieron. Esta prueba es más estricta que adjuntar un enlace genérico a un documento: el nodo final incluye desplazamientos de caracteres y el texto exacto de la evidencia. Un revisor puede, por tanto, distinguir la observación reportada original del estado inferido y del plan que la utilizó.

4.6. Ablación de las salvaguardas del agente

La Tabla 4 y la Figura 5 muestran la ablación del agente. Con una probabilidad de fallo inyectado de 0,30 y sin verificador ni reparación, el TaskScore fue de 0,556, la tasa de inseguridad de 0,278, la abstención de 0,167 y la exhaustividad de la evidencia de 0,500. Las salidas inseguras incluyeron afirmaciones de condición con sujeto equivocado, activos desconocidos alucinados y falsas aseveraciones de aprobación después de que la herramienta de aprobación fuera denegada. La compuerta de herramientas bloqueó el efecto secundario intentado, pero sin verificación de la salida el agente aún podía afirmar que la aprobación se había producido. Esta distinción entre la seguridad de la acción y la seguridad de la comunicación es operativamente importante.

Añadir el verificador redujo la tasa de inseguridad a cero y elevó el TaskScore a 0,667, pero la abstención aumentó a 0,444 porque las respuestas inválidas se rechazaron. Añadir la reparación preservó las cero salidas inseguras, elevó el TaskScore a 0,722, aumentó la exhaustividad de la evidencia a 0,600 y redujo la abstención a 0,333. La tasa de reparación fue de 0,278. Con un generador de afirmaciones limpio y ambas salvaguardas, el TaskScore alcanzó 0,944 y la exhaustividad de la evidencia 1,00, mientras que la abstención permaneció en 0,333 porque las solicitudes de aprobación y de activos desconocidos requieren correctamente un escalado. La abstención residual es, por tanto, en parte un comportamiento conforme a la política, no un fallo de capacidad.

Figura 5. Desempeño en las tareas, salidas inseguras y abstención bajo las ablaciones de verificación-reparación.

Fuente(s): Elaboración propia, 2026.

Tabla 4. Ablación del agente con una tasa de fallo inyectado de 0,30.

Configuración	TaskScore	Tasa de inseguridad	Abstención	Exhaustividad de evidencia	Tasa de reparación	Intentos de aprobación denegados
Sin V / sin R	0,556	0,278	0,167	0,500	0,000	1
V / sin 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

Configuración	TaskScore	Tasa de inseguridad	Abstención	Exhaustividad de evidencia	Tasa de reparación	Intentos de aprobación denegados
C limpio + V + R	0,944	0,000	0,333	1,000	0,000	0

Fuente(s): Elaboración propia, 2026.

La ablación respalda dos conclusiones dentro del banco de pruebas. En primer lugar, la verificación determinista fue el mecanismo que eliminó las afirmaciones finales inseguras; la compuerta de permisos por sí sola fue insuficiente, porque el agente podía informar erróneamente de una acción denegada. En segundo lugar, la reparación recuperó cobertura sin debilitar el verificador. Esto sugiere un orden útil para el diseño de agentes de alto riesgo: restringir las herramientas, verificar el resultado estructurado, intentar una reparación acotada y abstenerse cuando la afirmación no pueda fundamentarse. La autocorrección basada solo en avisos (prompts) no es necesaria para la propiedad de seguridad evaluada aquí.

4.7. Análisis de fallos, validez y limitaciones

El entorno sintético expuso varios fallos de especificación que serían difíciles de diagnosticar solo a partir de la precisión agregada de las tareas. La Tabla 5 resume las revisiones más consecuentes. La primera fue el ruido de observación. El recuento ingenuo de transiciones trataba los grados reportados como verdad y producía transiciones sesgadas y una mala calibración. Incorporar la matriz de confusión a la verosimilitud de los estados ocultos redujo este error. La segunda fue el condicionamiento a la intervención. La recalibración sin núcleos de recuperación juzgaba los componentes reparados frente a una trayectoria sin reparación y penalizaba incorrectamente el modelo. La tercera fue un desajuste entre el estadístico de prueba InSAR y el nulo de permutación, que creó 31 falsos positivos antes de la corrección.

Otros fallos concernían a la lógica institucional. Los umbrales absolutos de los actores no se transferían entre escalas de objetivos, por lo que se adoptó la normalización relativa, salvo para el tope presupuestario real. El refinamiento local era no monótono, lo que motivó una metarregla de mejor ronda. La comprobación de citas basada solo en la igualdad exacta del sujeto rechazaba afirmaciones válidas a nivel de puente respaldadas por evidencia de componentes, por lo que el verificador se modificó para resolver la jerarquía de activos. Por último, bloquear la herramienta de aprobación no impedía que el agente afirmara falsamente la aprobación, lo que motivó la verificación a nivel de salida de los predicados de aprobación.

Tabla 5. Fallos de especificación identificados mediante la validación sintética de lazo cerrado.

Fallo observado	Consecuencia	Revisión implementada
Grado reportado tratado como verdad latente	Transiciones sesgadas y mala calibración	Verosimilitud de confusión de observación e inferencia hacia delante
Retroalimentación evaluada sin intervenciones	Reparación exitosa contada como error de pronóstico	Trayectorias condicionadas a la intervención y núcleos de recuperación
El estadístico CUSUM difería del nulo de permutación	31 falsos positivos en desarrollo	Estadístico idéntico para las series observada y permutada, más control BH
Umbrales absolutos del consejo	Preferencias dependientes de la escala e inestables	Normalización relativa por conjunto de paquetes; tope presupuestario absoluto conservado
El refinamiento local repetido se supuso monótono	Una ronda posterior redujo la aceptación	Metaselección lexicográfica entre rondas
La cita exigía igualdad exacta del sujeto	Afirmaciones válidas sobre el puente padre rechazadas	Relación de evidencia consciente de la jerarquía de activos

Fallo observado	Consecuencia	Revisión implementada
Solo denegación de la herramienta	El agente podía informar falsamente de una aprobación	Verificación de predicados más compuerta obligatoria de aprobación humana

Fuente(s): Elaboración propia, 2026.

Varias limitaciones acotan las afirmaciones. En primer lugar, SynthTown es un único municipio generado, y los resultados principales reportados utilizan una sola semilla. El experimento demuestra una coherencia ejecutable, no una generalización estadística. En segundo lugar, las plantillas de informe y los tramos de evidencia se generan a partir del mismo esquema conocido. Por tanto, un F1 de extracción de 1,00 debería leerse como una prueba de contrato, no como un banco de pruebas de documentos reales. En tercer lugar, el proceso de deterioro de cuatro estados y los siete mecanismos están simplificados; el estimador jerárquico es un Bayes empírico por MAP, en lugar de un tratamiento posterior completo de la incertidumbre de los parámetros.

En cuarto lugar, el planificador utiliza una intervención por puente a lo largo del horizonte, cuatro objetivos y NSGA-II. No prueba la optimalidad global y no modela cuadrillas, plazos de aprovisionamiento, interacciones detalladas entre componentes ni equidad entre barrios. En quinto lugar, los roles del consejo son sustitutos deterministas, no actores humanos. Su valor reside en exponer el comportamiento de agregación y refinamiento, no en predecir la aceptación pública. En sexto lugar, el banco de pruebas del agente contiene 18 tareas con una única tasa de fallo. Una tasa de inseguridad nula se aplica solo a esas tareas y clases de fallo. Los avisos adversarios, la inyección indirecta de instrucciones, las herramientas coludidas y las entradas corruptas del almacén quedan fuera del alcance.

La evaluación sintética no establece la validez externa. Por ello, la Sección 5 presenta una implementación aparte para la ciudad de Utsunomiya, con el fin de comprobar si los mismos contratos de evidencia, el modelo de procedencia, la lógica de confianza en las observaciones, las interfaces de priorización y las salidas de planificación pueden operar sobre datos municipales heterogéneos a escala. Esa implementación no resuelve las brechas estadísticas e institucionales restantes. Múltiples semillas sintéticas deberían cuantificar la varianza y la probabilidad de fallo. Los documentos de inspección reales deberían poner a prueba el formato, el OCR, la ambigüedad y la evidencia ausente. El modelo de deterioro debería compararse con líneas base semi-markovianas o POMDP más ricas. La planificación debería contrastarse con solucionadores exactos de instancias pequeñas y con la sensibilidad a los pesos de presupuesto y tráfico. Expertos humanos deberían evaluar las tarjetas de evidencia de Utsunomiya y definir las rúbricas del consejo. Las pruebas del agente deberían ampliar las taxonomías de fallos e incluir resultados de herramientas adversarios antes de cualquier conexión con sistemas de aprobación o de órdenes de trabajo.

5. Implementación municipal en el mundo real en Utsunomiya

5.1. Propósito, alcance y límite de despliegue

Los experimentos controlados de la Sección 4 establecen la validez interna, porque se conocen los estados latentes, los efectos de intervención y los fallos del agente inyectados. Por sí solos, no muestran que los contratos de evidencia y las interfaces de planificación puedan operar sobre la heterogeneidad, la ausencia de datos y la escala de datos reales de infraestructura pública. Por ello, se preparó una segunda implementación para los registros de puentes de la ciudad de Utsunomiya. El propósito de esta implementación era validar el prototipo en un entorno relevante de datos municipales, no afirmar que el sistema hubiera entrado en operaciones rutinarias de la ciudad. Se ejecutó en modo sombra: el software produjo tarjetas de evidencia, candidatos de inspección y patrulla, orientación sobre observaciones por satélite y calendarios de mantenimiento, mientras que la autoridad final de ingeniería y administrativa permaneció fuera del software.

El piloto instanció las mismas distinciones rectoras utilizadas por el agente de lazo cerrado: los registros fuente se almacenaron como Observaciones; las condiciones derivadas de puentes o componentes se trataron como Estados inferidos; y las inspecciones, la monitorización, las reparaciones y las opciones de renovación se representaron como Intervenciones o acciones propuestas. Cada recomendación generada conservó campos de procedencia y confianza. El piloto con datos reales no activó todos los componentes del banco de pruebas sintético. En particular, el consejo determinista de actores y el arnés de inyección de fallos C-A-P-V-R se evaluaron en SynthTown, mientras que la implementación de Utsunomiya ejercitó la vinculación de identidad de activos, la ingesta de evidencia, la confianza en las observaciones por satélite, la generación de prioridades, la programación consciente de la capacidad y la planificación del ciclo de vida.

5.2. Integración de datos municipales

El inventario base contenía 1733 registros de puentes de Utsunomiya extraídos del catálogo nacional de puentes xROAD. El piloto responde al contexto político japonés más amplio de infraestructuras envejecidas, recursos de mantenimiento limitados y la necesidad de una gestión sistemática documentada por el Ministerio de Territorio, Infraestructuras, Transporte y Turismo (2024). Cada registro conservaba el identificador del puente, el nombre, la ruta, el gestor, el año de construcción, las dimensiones, el año de inspección, el grado de condición, el estado de reparación, las coordenadas y el registro fuente original. El catálogo proporcionó la identidad de activo común utilizada para conectar las observaciones por satélite, la evidencia de inspección, los modelos a nivel de componente y las salidas de planificación. Cuando los identificadores no se compartían entre fuentes, la coincidencia utilizó el nombre del puente, la ruta, la ubicación y comprobaciones de distancia, en lugar de una asignación silenciosa por vecino más cercano.

La monitorización por satélite se instanció para 22 puentes objetivo. El muestreo produjo 1592 puntos de InSAR de dispersores persistentes, incluidos 282 puntos clasificados como estacionales. El prototipo almacenó las series de desplazamiento a nivel de punto y los resúmenes a nivel de puente como observaciones, en lugar de como diagnósticos estructurales directos. Un módulo de visibilidad utilizó el contexto urbano tridimensional circundante y la geometría de visión del radar para producir características de visibilidad, sombra y riesgo de solapamiento (layover). El manifiesto del piloto registró 22 evaluaciones de activos de alta visibilidad y dos indicadores de reserva sobre el terreno. Por tanto, una ausencia de anomalía con baja visibilidad no se interpretó como evidencia de seguridad; en cambio, la orientación de acción mantuvo o aumentó la necesidad de patrulla sobre el terreno o de otro modo de observación.

El catálogo de evidencia también conectó 106 PDF de inspección, 107 modelos de puente tridimensionales orientados al mantenimiento y 4182 registros de componentes. Estos recuentos describen el catálogo compartido del proyecto utilizado por el piloto y no implican que cada tipo de dato estuviera disponible para cada puente de Utsunomiya. La evidencia ausente se mantuvo explícita. Los PDF de inspección se vincularon por nombre, ruta y ubicación, y expusieron los documentos fuente y las fotografías extraídas. Los modelos tridimensionales se utilizaron como referencias de mantenimiento con nivel de detalle (LOD) 2.5 y con identificadores de componente, no como modelos de información de construcción (BIM) de grado constructivo. Esta distinción impidió que la disponibilidad geométrica se confundiera con una condición estructural verificada.

5.3. Salidas de decisión y ejecución operativa

La tubería municipal combinó la vulnerabilidad estructural, la evidencia de deformación, el cambio del entorno, la criticidad de uso, las señales de patrulla, el valor de la incertidumbre y el tiempo transcurrido desde la inspección en una puntuación de prioridad de mantenimiento. Generó 500 tarjetas de prioridad para su revisión. Cada tarjeta presentaba los atributos de activo disponibles, la evidencia satelital y su orientación de visibilidad, la evidencia de inspección vinculada, las referencias a modelos y la categoría de acción propuesta. La distribución de salida fue de 15 puentes en la categoría A, 83 en la B, uno en la C, 10 en la D y 391 en la E. Estas categorías eran etiquetas operativas producidas por el prototipo; no eran calificaciones de condición reglamentarias y no sustituían el diagnóstico de un ingeniero.

La clasificación se convirtió después en un programa de inspección y patrulla consciente de la capacidad. Bajo la capacidad anual configurada, se seleccionaron 108 elementos, y 21 puentes portaban señales de informe de patrulla. Un planificador aparte a nivel de municipio evaluó las opciones de no hacer nada, monitorización, reparación preventiva, reparación mayor y renovación a lo largo de 30 años. Generó 900 acciones seleccionadas respetando los límites configurados de presupuesto anual y de número de proyectos. Por tanto, el piloto puso a prueba algo más que la visualización en un mapa: la evidencia heterogénea se transformó en candidatos de acción revisables y en un plan escalonado en el tiempo que podía inspeccionarse en un panel de evidencia basado en Cesium y exportarse como artefactos JSON, GeoJSON, CSV y HTML.

Tabla 6. Datos del piloto a escala municipal de Utsunomiya y salidas generadas.

Elemento del piloto	Recuento o resultado	Función en el prototipo
Inventario de puentes xROAD	1733 puentes	Catálogo de activos y población de planificación
Monitorización PS-InSAR	22 objetivos; 1592 puntos; 282 estacionales	Observaciones de deformación e interpretación estacional
Confianza en la observación SAR	22 de alta visibilidad; 2 indicadores de reserva	Fiabilidad de la observación y reserva sobre el terreno
Evidencia de inspección	106 PDF vinculados en el catálogo compartido	Documentos, registros extraídos y fotografías
Referencias de mantenimiento 3D	107 modelos; 4182 componentes	Visualización de evidencia vinculada a componentes
Salidas de prioridad	500 tarjetas de evidencia	Revisión trazable de los candidatos clasificados
Categorías de acción	A/B/C/D/E = 15/83/1/10/391	Solo enrutamiento operativo
Selección consciente de la capacidad	108 elementos de inspección o patrulla	Carga de trabajo anual ejecutable
Plan de ciclo de vida	900 acciones a lo largo de 30 años	Planificación restringida por presupuesto y número

Fuente(s): Elaboración propia, 2026.

5.4. Interpretación TRL5, salvaguardas y brecha restante

La implementación de Utsunomiya es coherente con una interpretación acotada de Nivel de Madurez Tecnológica (TRL) 5: se implementaron y pusieron a prueba elementos de software de extremo a extremo en un entorno relevante de datos municipales, mientras que el despliegue operativo permaneció fuera del alcance del estudio (NASA Earth Science and Technology Office, s. f.). Esta clasificación concierne a la validación técnica de un prototipo. No implica adopción municipal, certificación, aprovisionamiento, conexión con sistemas de aprobación en producción, ni evidencia de que las recomendaciones fueran ejecutadas por la ciudad de Utsunomiya. La ejecución técnica y el despliegue institucional siguen siendo objetivos de validación distintos.

El piloto también acota la interpretación de los resultados sintéticos. SynthTown sigue siendo la fuente de evidencia de verdad de referencia para la recuperación de estados latentes, las comprobaciones de retroalimentación causal, el comportamiento de refinamiento de los actores y las ablaciones de seguridad de verificación-reparación. Utsunomiya aporta evidencia externa de que las interfaces de activos, observación, procedencia, visibilidad, priorización y planificación pueden aceptar datos reales a escala municipal sin colapsar la información ausente en una falsa certeza. Por el contrario, el piloto aún no aporta verdad de referencia para los pronósticos de deterioro ni evidencia prospectiva de que el calendario generado reduzca fallos, costes o interrupciones.

El siguiente paso de madurez es una evaluación en modo sombra supervisada por expertos. Los gestores de infraestructuras deberían revisar una muestra estratificada de tarjetas de prioridad, comparar los elementos seleccionados con los planes de inspección y reparación existentes, registrar la evidencia ausente o engañosa y medir el tiempo necesario para reconstruir cada recomendación. Las principales métricas de aceptación deberían ser la resolución de los enlaces de evidencia, las violaciones de restricciones, la suficiencia de la explicación valorada por expertos, la abstención apropiada ante evidencia ausente y la estabilidad de la carga de trabajo seleccionada frente a cambios plausibles en el presupuesto y la fiabilidad de las observaciones. Solo después de esta etapa debería conectarse el agente a sistemas de aprobación o de órdenes de trabajo.

6. Conclusión

Este estudio implementó una capa de decisión de IA auditable de lazo cerrado para gemelos digitales de puentes urbanos, validó sus mecanismos internos en SynthTown e instanció sus interfaces de evidencia y planificación con datos reales en un piloto a escala municipal en Utsunomiya. La contribución es la composición de la ingesta de observaciones ligada a la evidencia, la predicción del deterioro por estados ocultos y condicionada a la intervención, la planificación multiobjetivo con restricciones, el cribado por parte de actores con evidencia limitada, la aprobación humana, la retroalimentación de resultados y las salvaguardas de verificación-reparación. El sistema mantiene las observaciones distintas de los estados inferidos, las recomendaciones distintas de las decisiones y las decisiones distintas de las intervenciones ejecutadas.

En SynthTown, el bucle procesó 50 puentes y 842 componentes, logró pronósticos calibrados previos a la decisión, recuperó la heterogeneidad a nivel de puente, generó 80 soluciones de Pareto factibles, mejoró la aceptación mínima del consejo mediante el refinamiento local, ingirió los resultados posteriores a la intervención y preservó una ruta de seis nodos desde la decisión final hasta un tramo documental exacto. En la ablación del agente, la verificación redujo las salidas inseguras de 0,278 a cero, mientras que la reparación acotada mejoró la cobertura de tareas sin debilitar ese resultado. En Utsunomiya, el prototipo procesó 1733 registros de puentes y 1592 puntos PS-InSAR, vinculó la evidencia de inspección y de modelos de componentes, generó 500 tarjetas de prioridad, seleccionó 108 elementos de inspección o patrulla restringidos por la capacidad y produjo un plan de mantenimiento a 30 años con 900 acciones.

Los hallazgos respaldan un principio de diseño y una afirmación de madurez acotada. Los agentes de infraestructuras de alto riesgo no deberían evaluarse solo por la precisión de las respuestas o la fluidez del modelo; deberían evaluarse como componentes de un sistema de decisión cerrado, con observaciones inciertas, restricciones ejecutables, límites de permisos, abstención, reparación, procedencia y retroalimentación. SynthTown revela defectos de especificación bajo verdad conocida, mientras que el piloto de Utsunomiya demuestra la ejecución en un entorno relevante de datos municipales, coherente con un prototipo de TRL5. No demuestra una adopción administrativa en producción ni una reducción prospectiva del riesgo. La siguiente etapa es la validación con múltiples semillas, las pruebas de extracción con documentos reales y la evaluación en modo sombra supervisada por expertos de las salidas municipales.

6.1. Disponibilidad de datos y de código

Un paquete de replicación anonimizado que contiene el generador de SynthTown, la configuración de semilla fija, las definiciones de esquema, los scripts de evaluación y los archivos de resultados derivados se proporcionará a través del sistema de envío para la revisión por pares. La publicación del banco de pruebas sintético y del código no sensible está prevista tras la aceptación. Los registros fuente municipales que ya son públicos siguen estando disponibles a través de sus proveedores originales; los artefactos derivados que contengan información de vinculación local se compartirán solo cuando las condiciones de licencia, privacidad y administrativas lo permitan.

Referencias

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

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