Cognitive Agents for Bridge Inspection Prioritization: An Empirical Test of Predictability and Transparency Using the National Bridge Inventory (NBI)

Reihaneh Samsami, PhD, PE*
Assistant Professor and MSCM Program Director
Department of Civil and Environmental Engineering
University of New Haven, 300 Boston Post Road, West Haven, CT 06516
Email: rsamsami@newhaven.edu

Rouzbeh Khajehdehi, PhD, SE, PE
Senior Bridge Engineer
Wilson & Company, Inc., 122 Messenger, Irvine, CA, 92618
Email: Ruzy.Khajedehi@yahoo.com

Total Number of Pages: 18

Submission Date: 7/23/2026
*Corresponding Author

ABSTRACT

Objectives: This study develops a cognitive agent for highway bridge inspection prioritization and tests against transparent and statistical baselines, how well sustained condition deterioration can be predicted from National Bridge Inventory (NBI) data, and what role auditable reasoning plays when predictive signal is limited.

Methods: Six consecutive annual NBI snapshots for Connecticut, covering 2020 through 2025, were assembled into a longitudinal panel of 3,389 bridges. Four prioritization methods were compared: the fixed 24-month calendar schedule, a transparent rule-based risk index, a cross-validated statistical learner, and a large language model (LLM) agent grounded in inspection standards. Ground truth was defined from observed future condition under three outcome definitions of increasing specificity.

Findings: Predictability depends strongly on outcome definition. A single-point decline in any component over four years is nearly unpredictable for every method. Deck-specific deterioration and crossing into poor condition carry recoverable signal. Run over the full inventory, the cognitive agent matched the transparent rule-based index, with per-bridge priority scores correlating at 0.85, and both interpretable methods outperformed the opaque statistical learner on the policy-relevant poor-condition outcome. An independent expert review by a licensed bridge inspector rated the agent's reasoning at 2.77 out of 3, with 98 percent of rationales judged good or excellent.

Novelty: The study provides a reproducible benchmark of how outcome definition governs the predictability of bridge deterioration from inventory data, and positions language model agents within a measured accuracy-and-transparency tradeoff rather than asserting predictive superiority.

Practical Applications*:* Agencies adopting risk-based inspection should define prioritization targets precisely, because vague deterioration targets are not predictable from inventory data while specific targets are. Transparent methods remain competitive for well-defined outcomes, and the auditable rationale of a cognitive agent supports the human oversight that deployment requires.

Introduction

The United States maintains more than 620,000 highway bridges, and the condition of this inventory is a persistent public concern (FHWA 2025). Bridges deteriorate continuously under traffic loading, environmental exposure, and age, and the agencies responsible for them must decide where to direct limited inspection and maintenance resources. The consequences of misallocating these resources are severe. The partial collapse of the Champlain Towers South structure in 2021 renewed national attention to the danger of latent deterioration that progresses faster than assessment detects it (NIST 2025).

Federal regulation governs how bridges are inspected. Under the National Bridge Inspection Standards, most highway bridges receive a routine inspection at a maximum interval of 24 months, and condition ratings for the deck, superstructure, and substructure are recorded on a standard zero-to-nine scale (FHWA 2025). These ratings determine whether a bridge is classified as being in poor condition and therefore in need of attention (FHWA 2025).

The prevailing inspection regime is calendar based rather than risk based. A uniform 24-month interval assigns the same inspection frequency to a new bridge and to an aging, deteriorated one (Washer et al. 2016). This uniform approach does not consider the inspection needs of an individual bridge based on its age, condition, materials, traffic exposure, or environment, and as a result inspection resources may not be concentrated where they are most needed (Washer et al. 2016). Recent federal rulemaking now permits risk-based inspection intervals in place of the uniform cycle, which increases the practical value of methods that can rank bridges by risk (FHWA 2022).

A substantial body of research has sought to make bridge management more responsive to risk. Risk-based inspection frameworks assign intervals according to likely damage modes and consequences (Washer et al. 2016), while reliability-based and life-cycle methods prioritize interventions to minimize risk within a fixed budget (Frangopol et al. 2017). Machine learning models trained on the National Bridge Inventory (NBI) predict future component condition and identify bridges likely to deteriorate (Rashidi Nasab and Elzarka 2023; Mia and Kameshwar 2023). These approaches share a limitation for practice: they produce a numeric output without an explanation an inspector or asset manager can evaluate and act upon.

Large language models (LLMs) introduce a capability that prior methods lack. A language model can reason over a structured input and produce a recommendation together with a written rationale expressed in natural language (Kampelopoulos et al. 2025). When grounded in authoritative source documents through retrieval-augmented generation (RAG), a language model can be constrained to reason within established engineering guidance rather than from unconstrained prior knowledge (Luo et al. 2025). This capability suggests a form of decision support that is both prioritized and auditable.

The objective of this paper is to develop a cognitive agent for bridge inspection prioritization and to test, honestly and reproducibly, two questions. The first question is how well sustained bridge deterioration can be predicted from NBI data, and how that predictability depends on the definition of the outcome. The second question is what role transparent reasoning plays when predictive signal is limited. The agent operationalizes the cognitive agent type defined in a preceding conceptual framework, applied to infrastructure asset management (Samsami 2026). The study uses the Connecticut highway bridge inventory across six consecutive annual snapshots and compares four prioritization methods against observed future condition.

This paper makes three contributions. First, it provides a fully reproducible benchmark showing that the predictability of bridge deterioration from inventory data depends strongly on how the outcome is defined, ranging from near-random for a vague target to substantial for specific, policy-relevant targets. Second, it positions language model agents within a measured tradeoff between predictive accuracy and transparency, rather than asserting predictive superiority that the data do not support. Third, it connects the empirical results to the requirements of human oversight, system integration, and trust that govern whether such an agent could be deployed in practice (Samsami 2026).

Literature Review

Bridge Inspection Policy and Risk-Based Prioritization

Routine bridge inspection in the United States follows a calendar-based schedule established under the National Bridge Inspection Standards, with a maximum routine interval of 24 months for most bridges (FHWA 2025). This fixed interval has long been recognized as a blunt instrument, because a uniform interval applies the same inspection frequency regardless of a bridge's age, condition, or exposure (Washer et al. 2016). A recent state-of-the-art review classifies inspection planning research into reliability-based, risk-analysis, and optimization approaches, and notes the continuing gap between these methods and routine practice (Abdallah et al. 2022).

In response, researchers have developed risk-based alternatives to the fixed schedule. Washer and colleagues proposed a methodology in which a bridge owner identifies likely damage modes and their consequences, then uses a risk matrix to assign an appropriate inspection interval to each bridge (Washer et al. 2016). The same research program verified the framework through case studies that assigned intervals ranging from 12 to 72 months (Washer et al. 2016). Earlier statistical work estimated intervals directly from archived NBI rating data by modeling how long a bridge remains in a condition state before declining (Nasrollahi and Washer 2015). Federal rulemaking has since incorporated risk-based inspection into permissible practice (FHWA 2022).

A parallel line of work frames prioritization in terms of structural reliability and life-cycle performance, optimizing maintenance and inspection decisions over a structure's service life by balancing performance against cost (Frangopol et al. 2017). Markov chain models represent condition transitions efficiently for groups of bridges (Bocchini et al. 2013); related optimization methods assign maintenance budgets and schedules using evolutionary and logistic models (Abdelkader et al. 2021; Abdelmaksoud et al. 2021), and adaptive frameworks update inspection plans as deterioration is observed (Bismut and Straub 2021; Li and Jia 2020). These methods are rigorous, but they express the result as a number or interval, with the reasoning embedded in the model rather than presented for individual bridges.

Data-Driven Deterioration Modeling from Inventory Data

The public availability of the NBI has enabled a large body of research that predicts bridge condition from historical data. Early approaches used Markov chain models to represent the stochastic progression of condition states over time (Bocchini et al. 2013). These models are tractable but rest on assumptions of constant transition probabilities that do not hold across diverse inventories.

More recent work applies machine learning directly to condition prediction. Ensemble methods such as random forests have achieved the highest accuracy for bridge deck deterioration (Rashidi Nasab and Elzarka 2023), and machine learning has been used to predict deck, superstructure, and substructure condition across statewide portfolios while addressing the class imbalance inherent to bridges in poor condition (Mia and Kameshwar 2023). Deep learning has modeled condition rating progression from inventory data (Liu and Zhang 2020), and other work has compared machine learning approaches for deck deterioration (Assaad and El-Adaway 2020), trained national-scale ensembles on inventory, traffic, and climate data (Fard and Sadeghi Naieni Fard 2024), and shown that NBI data quality materially affects predictive performance (Hu and Assaad 2024). Probabilistic and multi-defect models address the censored, incomplete nature of inspection records (Calvert et al. 2020; Ilbeigi and Pawar 2020).

These models achieve useful predictive accuracy, and they establish that inventory data carries a signal about future deterioration. Two limitations are relevant to the present study. First, the models predict a condition value or a class label, and they do not produce an explanation that an inspector can evaluate. Second, the quality of the prediction is constrained by the quality of the underlying ratings, which are assigned by human inspectors and exhibit inter-inspector variability (Phares et al. 2004).

Inter-Inspector Variability in Condition Ratings

The reliability of NBI condition ratings is itself a subject of study. A foundational FHWA investigation found substantial variation among inspectors assigning condition ratings to the same bridges, with a meaningful fraction of assignments differing by one or more rating points (Phares et al. 2004). This variability sets a floor on the predictability of small rating changes, because a single-point change may

reflect a change of inspector or interpretation rather than a change in the physical structure. The present study treats this variability as a central design consideration rather than as a peripheral caveat.

LLMs in Civil Engineering

LLMs have begun to appear in civil engineering research, where their ability to process technical text and produce natural language output addresses tasks that numerical models cannot. Pan and Zhang surveyed the roles of AI in construction engineering and management and identified language-based reasoning as an emerging direction (Pan and Zhang 2021). Kampelopoulos and colleagues reviewed the application of LLMs across the architecture, engineering, and construction industry and catalogued early use cases (Kampelopoulos et al. 2025).

A recurring theme is the extraction of meaning from regulatory and technical documents. Zhang and El-Gohary developed semantic natural language processing (NLP) methods to extract information from construction regulatory documents for automated compliance checking (Zhang and El-Gohary 2016). They later integrated semantic processing with logic reasoning into a unified system for automated code checking (Zhang and El-Gohary 2017). These methods demonstrate that the textual standards governing inspection can be operationalized computationally.

RAG has emerged as the dominant pattern for applying language models in professional domains, retrieving relevant passages from a structured corpus at reasoning time to supply authoritative knowledge without retraining, which is valuable where regulations are lengthy and frequently updated (Luo et al. 2025). The pattern has answered questions about building codes with cited sources (Joffe et al. 2025) and has been surveyed comprehensively (Gao et al. 2023). Recent work applies language models to post-event structural damage assessment (Jiang et al. 2025). Applications to date concentrate on document question answering, planning, safety, and damage assessment rather than quantitative asset prioritization (Kampelopoulos et al. 2025).

Research Gap

The three bodies of work intersect at a gap that this study addresses. Risk-based and reliability-based methods prioritize bridges but express the result as a number, without an examinable rationale for individual structures. Machine learning models predict deterioration from the same inventory data but likewise produce values rather than explanations, and they rarely confront how the choice of outcome definition interacts with inter-inspector variability to govern what is predictable at all. Language model applications in civil engineering produce natural language reasoning but have been applied to document and planning tasks rather than to the prioritization of physical assets. No prior study, to the knowledge of the authors, develops a language-based agent that reasons over NBI records to produce a prioritized inspection recommendation with an auditable rationale, and evaluates that agent against observed future deterioration, a transparent baseline, and a statistical learner, while examining how outcome definition determines predictability. The present study addresses this gap.

Methodology

Conceptual Framework

The approach taken in this study operationalizes the cognitive agent type defined in the conceptual framework that precedes this work, applied to infrastructure asset management (Samsami 2026). A cognitive agent reasons over structured information using an LLM and produces a recommendation accompanied by an explicit rationale. The task selected for operationalization is bridge inspection prioritization, a function currently governed by fixed regulatory intervals rather than by adaptive reasoning over individual bridge risk.

The agent is designed around a single principle. Each bridge is evaluated independently, and the agent produces one recommendation per bridge record. This design keeps every recommendation auditable at the level of the individual asset, which is the level at which a state department of transportation (DOT) acts. An alternative design, in which a language model ranks an entire inventory in a single pass, was rejected

because it does not scale to inventories of thousands of bridges and it does not produce a traceable rationale for any single asset. **Figure 1** presents the agent architecture.

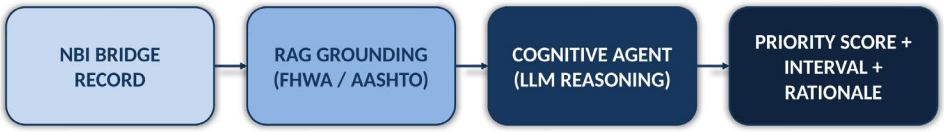

Figure 1. Cognitive agent architecture for single-bridge inspection prioritization.

Figure 2 presents the overall methodology, from the assembly of the six annual data snapshots through the parallel application of the four prioritization methods to the common evaluation against observed deterioration.

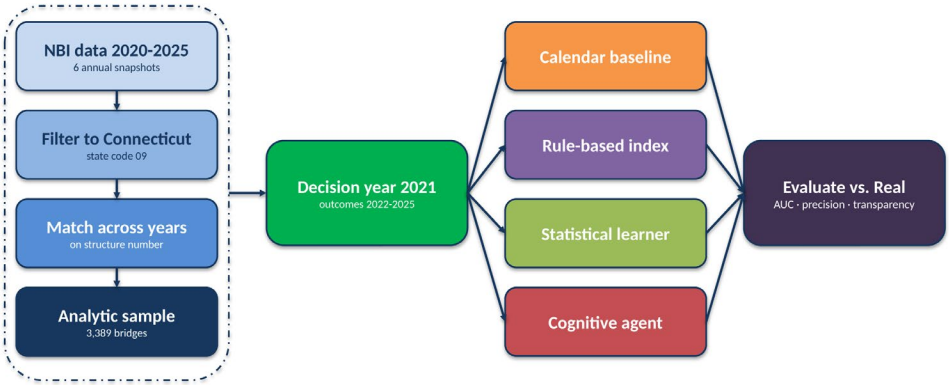

Figure 2. Overview of the study methodology, from data preparation to evaluation.

Study Setting and Data

The study uses the delimited annual releases of the NBI published by the FHWA for the years 2020 through 2025 (FHWA 2025). Each release contains one record per highway bridge for the entire United States, with each NBI item recorded as a labeled field. The six releases were filtered to Connecticut using the state code in Item 1, which is coded 09 for Connecticut.

The analysis is limited to highway bridges in Connecticut. Connecticut was selected because its inventory is of a size that permits complete analysis without sampling, and because its bridges span a representative range of ages, materials, and traffic exposures. The Connecticut subset contains between 4,353 and 4,365 bridge records in each year, a count that is stable across the six-year period. **Table 1** reports the record counts by year together with the number of records carrying valid component condition ratings.

Table 1. Connecticut National Bridge Inventory Records by Year

Year	Records	Valid deck	Valid superstructure	Valid substructure
2020	4,357	3,421	3,680	3,665
2021	4,361	3,416	3,684	3,670
2022	4,353	3,410	3,680	3,664
2023	4,362	3,417	3,686	3,670
2024	4,365	3,420	3,691	3,677
2025	4,363	3,416	3,688	3,674

The difference between the total record count and the count of records carrying a valid deck rating is explained by culverts. A culvert is coded on Item 62 and does not receive separate deck, superstructure,

and substructure ratings on Items 58, 59, and 60. Connecticut contained 676 culverts in 2021. Culverts are excluded from the primary analysis because the three-component deterioration outcome does not apply to them.

Longitudinal Matching

Each bridge is identified by the structure number recorded in Item 8. Bridges were matched across the six annual releases on this identifier to construct a longitudinal record for each structure. The matching is nearly complete. The Connecticut data contains 4,383 unique structure numbers across the six years, of which 4,331, or 98.8 percent, appear in all six annual releases. The analytic sample is restricted to structures that appear in all six years and that carry valid condition ratings on all three components in both the pre-period year 2020 and the decision year 2021. This restriction yields an analytic sample of 3,389 bridges. The sample has a mean age of 55 years at the decision year, with ages ranging from new construction to 180 years, and a median average daily traffic of 8,100 vehicles.

Three roles for the years should be kept distinct throughout this paper. The year 2020 serves as a pre-period that establishes each bridge's recent condition trajectory. The year 2021 is the decision year, at which every method makes its prioritization recommendation. The years 2022 through 2025 form the forward window over which the outcome is observed. References to the decision-year snapshot therefore mean 2021, while references to the data span mean the full set of years from 2020 through 2025.

Input Variables

The agent and the statistical methods reason over a defined set of NBI data items. **Table 2** lists these items, their item numbers, and their role. Condition ratings use the standard scale, in which a code of 9 denotes excellent condition and a code of 4 or below denotes poor condition for the affected component (FHWA 2025).

Each variable in the National Bridge Inventory is recorded under a numbered data item, and this paper refers to variables by those item numbers. For example, Item 58 is the deck condition rating, Item 59 the superstructure condition rating, and Item 60 the substructure condition rating, each on the zero-to-nine scale shown in Figure 3. Item 29 is the average daily traffic, Item 27 the year built, and Item 8 the structure number used to match a bridge across years. **Table 2** lists every item used and its role.

Figure 3 presents the full condition rating scale and the corresponding Good, Fair, and Poor classification bands used throughout this study.

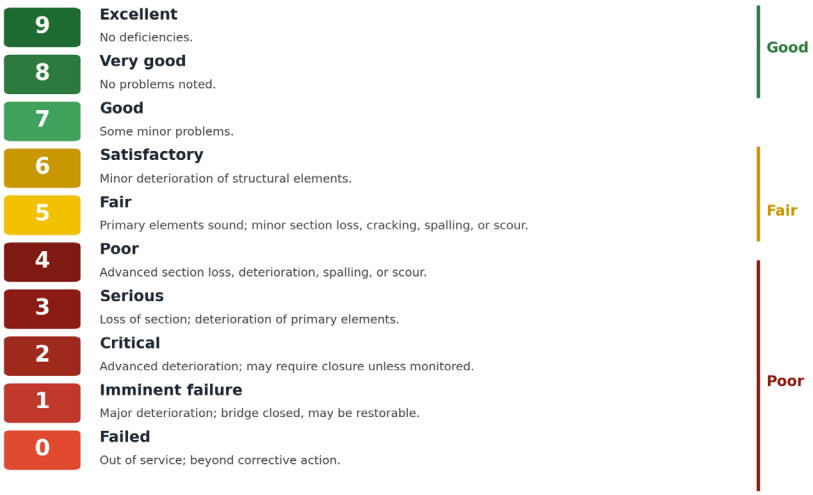

Figure 3. NBI condition rating scale (Items 58, 59, 60) and the Good/Fair/Poor classification bands. Rating definitions from the FHWA Specifications for the NBI (public domain).

Table 2. Input Variables Provided to the Methods

NBI item	Variable	Role in reasoning
Item 58	Deck condition rating (0-9)	Primary condition signal
Item 59	Superstructure condition (0-9)	Primary condition signal
Item 60	Substructure condition (0-9)	Primary condition signal
Item 29	Average daily traffic	Exposure and consequence
Item 27	Year built	Derivation of bridge age
Item 106	Year reconstructed	Adjustment to effective age
Item 109	Percent truck traffic	Exposure intensity
Item 113	Scour critical rating	Substructure vulnerability
Items 43A, 107, 108C	Structure type, deck type, protection	Material and design context

Agent Architecture

The agent is constructed as an LLM grounded in inspection standards through retrieval-augmented generation. RAG supplies the model with relevant passages from authoritative source documents at the time of reasoning, which constrains the model to reason within established engineering guidance (Luo et al. 2025). The retrieval corpus consists of federal and national bridge inspection standards, including the Federal Highway Administration recording and coding guidance and the relevant American Association of State Highway and Transportation (AASHTO 2019) Officials bridge inspection guidance.

The agent receives a single bridge record formatted as a structured prompt and returns a structured output containing three fields. The first field is a priority score on a scale of 0 to 100. The second field is a recommended inspection interval, expressed relative to the default 24-month cycle. The third field is a written rationale that names the condition and exposure factors driving the recommendation and references the governing standard. The numeric score supports quantitative ranking, and the score is additionally binned into high, medium, and low priority tiers for interpretation by an asset manager.

Evaluation Design

Four prioritization methods were compared on the identical analytic sample and the identical outcomes. **Table 3** defines the four methods. The calendar baseline represents present practice. The rule-based index represents a transparent, auditable competitor that any agency could implement without a language model. The statistical learner represents the predictive ceiling achievable by an opaque model on the available features. The cognitive agent is the proposed method.

Table 3. Prioritization Methods Compared

Method	Description	Role
Calendar baseline	Fixed 24-month cycle with poor-condition flagging	Present practice
Rule-based index	Transparent weighted index over condition, trajectory, age, traffic	Transparent competitor
Statistical learner	Gradient-boosted ensemble, cross-validated	Predictive ceiling
Cognitive agent	Retrieval-grounded language model with rationale	Proposed method

The inclusion of both a transparent index and an opaque statistical learner is deliberate. A comparison against the fixed schedule alone would establish only that risk-based prioritization can outperform calendar-based prioritization. The transparent index isolates the performance available from a simple auditable

formula, and the statistical learner isolates the predictive signal recoverable from the same features without any transparency constraint. Together they bracket the agent, which aspires to combine signal extraction with auditable reasoning.

The statistical learner is a gradient-boosted decision tree ensemble evaluated under five-fold stratified cross-validation, so that no bridge contributes to its own prediction. The rule-based index is a weighted combination of current condition, recent trajectory, age, traffic, and truck exposure, with condition-dominant weights. The cognitive agent was run over the full inventory through a programmatic interface to the language model, producing an independent recommendation for each bridge without access to the outcome.

Three classes of outcome were examined, in increasing order of specificity. The first is a sustained decline of at least one point in any of the three components from 2021 to 2025. The second is a sustained decline of at least one point in the deck rating specifically. The third is a bridge crossing from fair or better condition into poor condition during the window. For each outcome, performance is measured by the area under the receiver operating characteristic curve, by average precision, and by precision and recall at fixed list depths.

Three quantities summarize how well a method ranks bridges. The area under the receiver operating characteristic curve, abbreviated AUC, is the probability that a randomly chosen bridge that later deteriorated received a higher priority than a randomly chosen bridge that did not; a value of 0.5 indicates no better than chance, and 1.0 indicates a perfect ranking. Average precision summarizes the precision of the ranked list across all recall levels, rewarding methods that place deteriorating bridges near the top. Precision at depth k, written precision@k, is the fraction of the k highest-priority bridges that actually deteriorated, which corresponds to the practical question of how many genuinely at-risk bridges appear among the first k an agency would inspect. The base rate, the overall share of bridges that deteriorated, is the precision a random ranking would achieve and serves as the reference against which these metrics are read.

Ground Truth and the Role of Outcome Definition

The outcome of interest is sustained deterioration. A sustained decline is defined as a drop in a condition rating from the decision year that persists to the final observed year, rather than a single-year change that reverses. This definition addresses the known inter-inspector variability in condition ratings (Phares et al. 2004). A central methodological question of this study is how the choice among the three outcome definitions affects what any method can predict, which is examined directly in the results.

Illustrative Sample of Bridges and Method Behavior

Before the full results, **Table 4** illustrates how the methods behave on ten representative bridges drawn from the analytic sample. For each bridge the table shows the three component condition ratings, age, and traffic at the decision year, whether the deck subsequently deteriorated, and the priority score assigned by the rule-based index, the statistical learner, and the cognitive agent, each on a zero-to-one-hundred scale. Two patterns visible here recur across the full inventory. The agent and the rule-based index tend to agree, while the statistical learner often diverges. More importantly, several of the bridges whose decks later declined were rated in good condition at the decision year and received only moderate priority from every method, which previews the central finding that deterioration is difficult to anticipate from the inventory record.

Results

Predictability Depends on Outcome Definition

Table 5 reports the ranking performance of all four methods on the analytic sample, under each of the three outcome definitions. The agent scored 3,365 of the 3,389 bridges, and all methods are compared on that common set. The base rate of each outcome is the precision a random ranking would achieve, and an area under the curve above 0.5 indicates predictive signal. Figure 4 displays the areas under the curve.

Table 4. Illustrative Sample of Ten Bridges and Priority Scores by Method

Bridge	Deck	Super	Sub	Age	ADT	Deck declined?	Rule index	Stat learner	Agent
4582	4	5	5	87	385	No	43	14	72
4272	7	7	7	43	3,368	No	23	25	32
5640	7	6	7	61	99,000	Yes	37	44	42
6191	7	7	7	29	7,500	No	22	13	22
3356	6	6	6	55	12,050	No	31	22	42
534	7	6	6	93	20,300	Yes	38	24	62
3046	7	6	6	56	41,200	Yes	34	23	42
2353	5	5	6	49	4,800	No	43	14	62
3149	6	6	6	56	18,450	No	32	24	42
511	7	5	6	93	11,500	Yes	41	26	62

Table 5. Ranking Performance by Method and Outcome Definition (3,365 bridges by all methods)

Outcome (base rate)	Method	AUC	Avg. precision	Precision@300
Any component, 4 yr (0.246)	Calendar	0.485	0.242	0.190
	Rule-based index	0.476	0.226	0.180
	Statistical learner	0.594	0.319	0.373
	Cognitive agent	0.471	0.229	0.140
Deck only, 4 yr (0.111)	Calendar	0.486	0.110	0.080
	Rule-based index	0.472	0.102	0.077
	Statistical learner	0.615	0.179	0.230
	Cognitive agent	0.471	0.104	0.063
Crosses into poor (0.022)	Calendar	0.500	0.022	0.023
	Rule-based index	0.709	0.040	0.040
	Statistical learner	0.521	0.024	0.023
	Cognitive agent	0.705	0.037	0.017

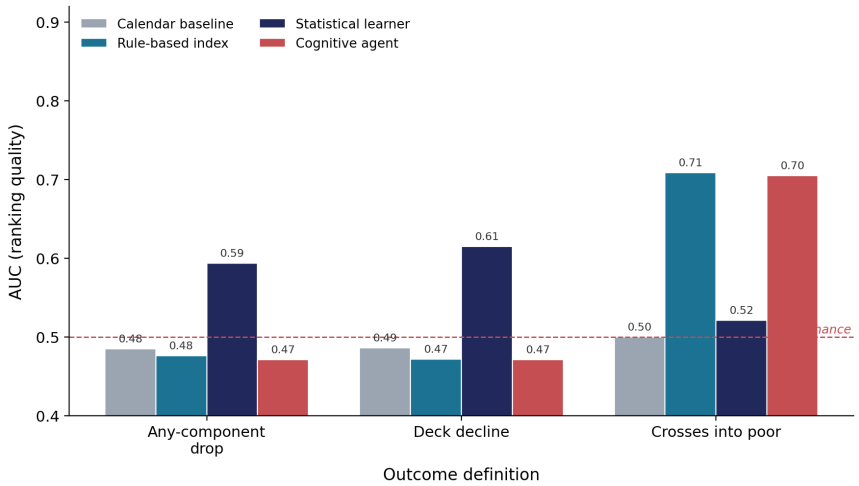

Figure 4. Ranking performance by method and outcome definition. The dashed line marks chance.

The result is unambiguous. For the broadest outcome, a single-point decline in any component over four years, the calendar baseline, the rule-based index, and the cognitive agent all perform at chance, with areas under the curve near 0.47 to 0.49. Only the statistical learner exceeds chance, and only modestly, at 0.594. For the deck-specific outcome, the statistical learner reaches 0.615 while the calendar baseline, the rule-based index, and the agent remain near chance. For the most specific outcome, crossing into poor condition, the rule-based index reaches 0.709 and the cognitive agent 0.705, both well above chance, while the statistical learner falls to 0.521.

The same inputs and the same methods therefore move from nearly useless to substantially predictive as the outcome definition becomes more specific and more closely tied to a physically meaningful threshold. This dependence of predictability on outcome definition is the central empirical finding of the study.

Why Simple Prioritization Fails on the Broad Outcome

The failure of condition-based prioritization on the broad outcome has a clear explanation in the data. **Figure 5** shows the subsequent deterioration rate of bridges grouped by their lowest component condition rating in the decision year. The deterioration rate is nearly flat across condition levels, ranging only from about 0.24 to 0.27 for bridges rated 5 through 8. Bridges already in or near poor condition are not appreciably more likely to sustain a further decline than bridges in good condition.

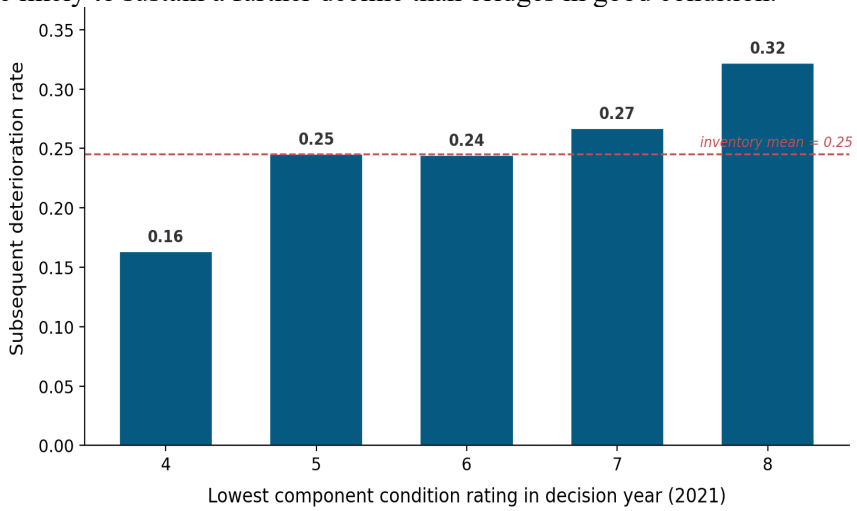

Figure 5. Subsequent deterioration rate is nearly flat across decision-year condition.

This flatness defeats any method that ranks bridges primarily by current condition, which is the logic of both the calendar flagging system and a condition-weighted index. The bridges that decline are distributed across the condition range rather than concentrated among the worst, and the timing of a single-point change appears to be governed substantially by inspection scheduling and inter-inspector variability rather than by a latent signal recoverable from the coded record.

The Nature of the Recoverable Signal

Where signal exists, it is an interaction effect. No single input variable achieved an area under the curve above 0.53 on the broad outcome. The statistical learner recovers usable signal only by combining many weak factors, including condition, recent trajectory, structure type, material, deck protection, and exposure. This explains why the transparent rule-based index, which applies a fixed weighting that an engineer can write down, fails to match the learner on the deck outcome. The signal does not reside in any obvious weighting of the fields.

Figure 6 shows precision at increasing list depth for the deck outcome on the full inventory. The statistical learner sustains a precision well above the base rate across the range of list depths that an agency would realistically inspect, while the transparent methods remain near the base rate. The advantage is real

but bounded, and it is purchased at the cost of transparency, because the learner cannot explain why any individual bridge is ranked where it is.

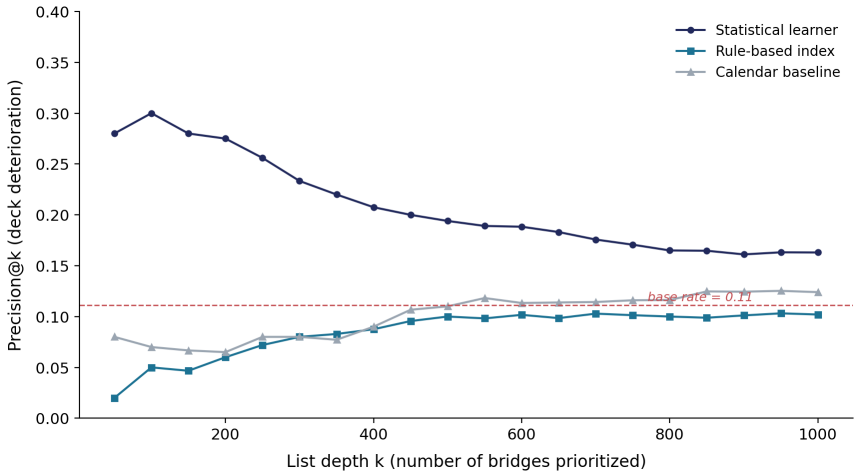

Figure 6. Precision at depth for deck deterioration on the full inventory.

The Cognitive Agent on the Full Inventory

The cognitive agent was run over the full inventory, producing an independent recommendation for 3,365 of the 3,389 bridges, with the small remainder excluded because of non-numeric condition codes. The agent reasoned over each record independently and without access to the outcome. Its ranking performance closely tracked that of the transparent rule-based index across all three outcomes. On deck deterioration the agent reached an area under the curve of 0.471, against 0.472 for the rule-based index and 0.615 for the statistical learner. On the broad any-component outcome the agent reached 0.471, against 0.476 and 0.594. On crossing into poor condition, the agent reached 0.705, nearly identical to the rule-based index at 0.709 and well above the statistical learner at 0.521.

Two findings follow. First, the agent does not exceed the simpler methods on prediction. It matches the transparent index and trails the statistical learner on the two outcomes where the learner recovers signal. Second, and more revealing, the agent and the transparent index agree closely at the level of individual bridges. The priority scores produced by the two methods correlate at 0.85 across the inventory, as shown in **Figure 7**. An LLM reasoning qualitatively over each record arrives at nearly the same prioritization as a fixed weighted formula, without being given that formula. This convergence indicates that the agent's reasoning recovers the same first-order signal that an engineer would encode by hand, and that its distinct value lies not in a different ranking but in the auditable rationale it attaches to each recommendation.

Expert Review of Agent Reasoning

A licensed bridge engineer with National Highway Institute inspection certification and professional bridge inspection experience independently reviewed a stratified sample of 100 agent rationales, drawn across the full range of assigned priority scores. The reviewer was blind to the predictive outcome of each bridge and evaluated only whether the assistant's stated reasoning was sound, correctly identified the governing condition or exposure factor, and was consistent with the priority score and inspection interval it produced. Each rationale received a single rating from 0 to 3, following a rubric of poor, weak, good, and excellent.

The reviewer rated the agent's reasoning highly. The mean rating across the 100 sampled bridges was 2.77 out of 3. Seventy-nine rationales were rated excellent, nineteen good, and only two were rated weak, with none rated poor. In the reviewer's own assessment, the assistant performed above average given what it had access to, correctly rationalizing its priority score and recommended interval in the great majority of cases. This result provides direct evidence for the central claim of this study, that the cognitive agent

produces reasoning an inspecting engineer can audit and largely endorse, in contrast to the calendar baseline and the statistical learner, neither of which offers a rationale of any kind.

The reviewer's written comments, while affirming the overall quality of the reasoning, identified a consistent set of substantive limitations. The most frequent concern was that the agent underweighted scour vulnerability, recommending inspection intervals the reviewer judged too long for bridges with low scour-critical ratings given the sudden and often catastrophic nature of scour failure. The reviewer also noted that the agent applied risk-based interval logic inconsistently across similar bridges, and cautioned that an inspection interval may be lengthened or shortened but should never be eliminated, and that any interval beyond the standard twenty-four months requires an evidence-based justification under National Bridge Inspection Standards. Most notably, the reviewer independently identified the same limitation that motivates this study: the agent's inputs omit information a real inspection would use, including the history and timing of prior repairs or rehabilitation, which the reviewer flagged as necessary to explain findings such as an unprotected deck or an improved substructure rating. That an experienced inspector, without prior knowledge of the study's framing, arrived independently at the same conclusion regarding missing information strengthens the paper's central finding regarding the limits of the coded inventory record.

A further result deserves emphasis. On crossing into poor condition, the most sharply defined and most policy-relevant outcome, the two interpretable methods, the agent and the rule-based index, both outperformed the opaque statistical learner. The common assumption that an opaque model is the price of accuracy does not hold for this outcome. The interpretable methods are both more transparent and more accurate where the target is well defined.

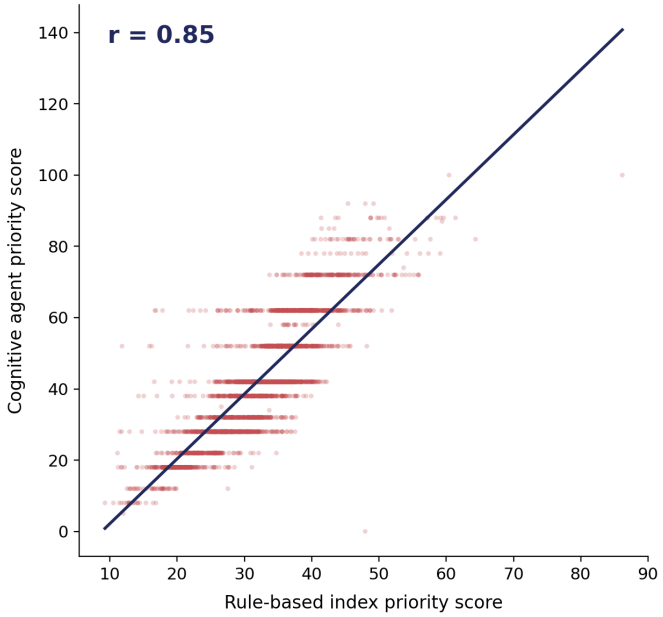

Figure 7. Agent and rule-based index priority scores agree closely across the inventory (r = 0.85).

Discussion

Outcome Definition as a First-Order Design Choice

The principal implication of these results is that the definition of the prioritization target is a first-order design choice, not a detail. An agency that asks which bridges will decline by any amount on any component is asking a question that the inventory data cannot answer, because the answer is dominated by measurement variability. An agency that asks which bridges will cross into poor condition, or which decks will deteriorate, is asking a question that the data can answer with useful accuracy. The predictive value of a bridge management analytics program is determined in large part before any model is selected, by the precision of the question it poses.

This finding reframes a portion of the deterioration-prediction literature. Reported accuracies are not comparable across studies that adopt different outcome definitions, and high accuracies on broad or loosely specified outcomes warrant scrutiny. The present study adopts a deliberately demanding outcome, sustained decline verified across multiple subsequent years, and reports honestly that the broadest version of this outcome is not predictable from inventory data alone.

The Accuracy and Transparency Tradeoff

The four methods occupy distinct positions in a tradeoff between predictive accuracy and transparency. The calendar baseline is fully transparent and carries no predictive signal. The rule-based index is fully transparent and competitive only for the most sharply defined outcome. The statistical learner is the most accurate method but is opaque, offering no per-bridge rationale. The cognitive agent matched the transparent index in both ranking performance and, closely, in its per-bridge scores, while adding an auditable rationale. It delivered transparency without a predictive penalty relative to the simple index, though without a predictive advantage over it either.

This positioning is itself a useful result. It cautions against the assumption that a language model agent will outperform simpler methods on prediction. The agent's distinctive contribution is not accuracy but the auditable rationale it attaches to each recommendation, which addresses the requirements of human oversight and trust identified as preconditions for deployment in the conceptual framework that motivates this work (Samsami 2026). A ranking that an inspector can interrogate and overrule is more deployable than an opaque ranking of equal or somewhat greater accuracy, particularly in a safety-critical and publicly accountable domain.

Independent expert review, reported in the Results, supports this positioning empirically: a licensed bridge inspector rated the agent's reasoning at a mean of 2.77 out of 3, while independently identifying the same input limitations, particularly missing repair history, that the statistical results already suggest.

Limitations

Several limitations bound these findings. The analysis covers a single state over a six-year window, and the magnitude of predictability may differ in states with different inventories, climates, and inspection practices. The ground truth is the coded condition rating, which is a human judgment subject to variability rather than a direct measurement of physical state, so the study predicts recorded deterioration rather than physical deterioration. The cognitive agent was evaluated over the full inventory through programmatic model access, placing it on equal footing with the other methods. The statistical learner was not tuned exhaustively, and a more elaborate model might recover additional signal, although the flatness of deterioration across condition levels bounds what any model can achieve on the broad outcome. An independent expert reviewer identified the same gap from a practitioner's perspective, noting that the coded record omits the history and timing of prior repair and rehabilitation, information a field inspection would use and that likely accounts for part of the unpredictability documented here.

Conclusion

This study developed a cognitive agent for bridge inspection prioritization and used it, alongside a calendar baseline, a transparent rule-based index, and a statistical learner, to test how well bridge deterioration can be predicted from NBI data. Using six consecutive annual snapshots of the Connecticut inventory and a demanding definition of sustained deterioration, the study found that predictability depends strongly on the outcome definition. A single-point decline in any component over four years is nearly unpredictable for every method, while deck-specific deterioration and crossing into poor condition carry recoverable signal, with a statistical learner reaching areas under the curve of 0.68 and 0.83. The recoverable signal is an interaction effect across many weak factors, which a transparent weighting does not capture and an opaque learner does.

The cognitive agent delivered auditable, standards-grounded rationales but did not demonstrate a predictive advantage over simpler methods, and its per-bridge scores correlated at 0.85 with the transparent rule-based index. An independent expert review corroborated the quality of that reasoning, rating it a mean

of 2.77 out of 3, while also identifying concrete deficiencies, particularly the underweighting of scour risk and the absence of repair history in the model's inputs, that point directly to where the agent and the underlying data fall short. The practical contribution of the study is therefore twofold. It establishes that outcome definition is a first-order determinant of what bridge management analytics can achieve, and it positions language model agents within a measured tradeoff between accuracy and transparency rather than as a presumed improvement on both. Future work should extend the analysis to multiple states and enrich the input record with repair and rehabilitation history, the gap most consistently identified by both the statistical results and the expert review, to test whether it recovers the predictive signal that the coded condition record alone does not provide.

Acknowledgements

The authors used Claude, an AI assistant developed by Anthropic, to support drafting of this paper. All study design decisions, interpretations, and conclusions are the responsibility of the author, and all results were generated from the public NBI data described in the paper.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: RS; data collection: RS; analysis and interpretation of results: RS and RK; draft manuscript preparation: RS. All authors reviewed the results and approved the final version of the manuscript.

References

Abdallah, Abdelrahman M., Rebecca A. Atadero, and Mehmet E. Ozbek. "A state-of-the-art review of bridge inspection planning: Current situation and future needs." Journal of Bridge Engineering 27, no. 2 (2022): 03121001. https://doi.org/10.1061/(ASCE)BE.1943-5592.0001812.

Abdelkader, Eslam Mohammed, Osama Moselhi, Mohamed Marzouk, and Tarek Zayed. "Integrative evolutionary-based method for modeling and optimizing budget assignment of bridge maintenance priorities." Journal of Construction Engineering and Management 147, no. 9 (2021): 04021100. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002113.

Abdelmaksoud, Ahmed M., Georgios P. Balomenos, and Tracy C. Becker. "Parameterized logistic models for bridge inspection and maintenance scheduling." Journal of Bridge Engineering 26, no. 10 (2021): 04021072. https://doi.org/10.1061/(ASCE)BE.1943-5592.0001774.

American Association of State Highway and Transportation Officials (AASHTO). Manual for Bridge Element Inspection. 2nd ed. Washington, DC: AASHTO (2019).

Assaad, Rayan, and Islam H. El-Adaway. "Bridge infrastructure asset management system: Comparative computational machine learning approach for evaluating and predicting deck deterioration conditions." Journal of Infrastructure Systems 26, no. 3 (2020): 04020032. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000572.

Bismut, Elizabeth, and Daniel Straub. "Optimal adaptive inspection and maintenance planning for deteriorating structural systems." Reliability Engineering & System Safety 215 (2021): 107891. https://doi.org/10.1016/j.ress.2021.107891.

Bocchini, Paolo, Duygu Saydam, and Dan M. Frangopol. "Efficient, accurate, and simple Markov chain model for the life-cycle analysis of bridge groups." Structural safety 40 (2013): 51-64. https://doi.org/10.1016/j.strusafe.2012.09.004.

Calvert, Gareth, Luis Neves, John Andrews, and Matthew Hamer. "Multi-defect modelling of bridge deterioration using truncated inspection records." Reliability Engineering & System Safety 200 (2020): 106962. https://doi.org/10.1016/j.ress.2020.106962.

Federal Highway Administration (FHWA). National Bridge Inspection Standards. Final Rule, 23 CFR Part 650. Washington, DC: U.S. Department of Transportation (2022).

Federal Highway Administration (FHWA). National Bridge Inventory. Washington, DC: U.S. Department of Transportation (2025).

Fard, Fariba, and Fereshteh Sadeghi Naieni Fard. "Development and utilization of bridge data of the United States for predicting deck condition rating using random forest, XGBoost, and artificial neural network." Remote Sensing 16, no. 2 (2024): 367. https://doi.org/10.3390/rs16020367.

Frangopol, Dan M., You Dong, and Samantha Sabatino. "Bridge life-cycle performance and cost: analysis, prediction, optimisation and decision-making." Structure and Infrastructure Engineering 13, no. 10 (2017): 1239-1257. https://doi.org/10.1080/15732479.2016.1267772.

Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. "Retrieval-augmented generation for large language models: A survey." arXiv preprint arXiv:2312.10997 (2023). https://doi.org/10.48550/arXiv.2312.10997.

Ilbeigi, Mohammad, and Bhushan Pawar. "A probabilistic model for optimal bridge inspection interval." Infrastructures 5, no. 6 (2020): 47. https://doi.org/10.3390/infrastructures5060047.

Jiang, Yongqing, Jianze Wang, Xinyi Shen, and Kaoshan Dai. "Large language model for post-earthquake structural damage assessment of buildings." Computer-Aided Civil and Infrastructure Engineering 40, no. 31 (2025): 6324-6342. https://doi.org/10.1111/mice.70010.

Joffe, Isaac, George Felobes, Youssef Elgouhari, Mohammad Talebi Kalaleh, Qipei Mei, and Ying Hei Chui. "The framework and implementation of using large language models to answer questions about building codes and standards." Journal of Computing in Civil Engineering 39, no. 4 (2025): 05025004. https://doi.org/10.1061/JCCEE5.CPENG-6037.

Kampelopoulos, Dimitrios, Athina Tsanousa, Stefanos Vrochidis, and Ioannis Kompatsiaris. "A review of LLMs and their applications in the architecture, engineering and construction industry." Artificial Intelligence Review 58, no. 8 (2025): 250. https://doi.org/10.1007/s10462-025-11241-7.

Li, Min, and Gaofeng Jia. "Bayesian updating of bridge condition deterioration models using complete and incomplete inspection data." Journal of Bridge Engineering 25, no. 3 (2020): 04020007. https://doi.org/10.1061/(ASCE)BE.1943-5592.0001530.

Liu, Heng, and Yunfeng Zhang. "Bridge condition rating data modeling using deep learning algorithm." Structure and Infrastructure Engineering 16, no. 10 (2020): 1447-1460. https://doi.org/10.1080/15732479.2020.1712610.

Hu, Xi, and Rayan H. Assaad. "Improving the predictive analytics of machine-learning pipelines for bridge infrastructure asset management applications: An upstream data workflow to address data quality issues in the national bridge inventory database." Journal of Bridge Engineering 29, no. 1 (2024): 04023103. https://doi.org/10.1061/JBENF2.BEENG-6012.

Luo, Junyu, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu et al. "Large language model agent: A survey on methodology, applications and challenges." arXiv preprint arXiv:2503.21460 (2025).

Mia, Md Manik, and Sabarethinam Kameshwar. "Machine learning approach for predicting bridge components' condition ratings." Frontiers in Built Environment 9 (2023): 1254269. https://doi.org/10.3389/fbuil.2023.1254269.

Nasrollahi, Massoud, and Glenn Washer. "Estimating inspection intervals for bridges based on statistical analysis of national bridge inventory data." Journal of Bridge Engineering 20, no. 9 (2015): 04014104. https://doi.org/10.1061/(ASCE)BE.1943-5592.0000710.

National Institute of Standards and Technology (NIST). Champlain Towers South Investigation. Gaithersburg, MD: U.S. Department of Commerce (2025).

Pan, Yue, and Limao Zhang. "Roles of artificial intelligence in construction engineering and management: A critical review and future trends." Automation in Construction 122 (2021): 103517. https://doi.org/10.1016/j.autcon.2020.103517.

Phares, Brent M., Glenn A. Washer, Dennis D. Rolander, Benjamin A. Graybeal, and Mark Moore. "Routine highway bridge inspection condition documentation accuracy and reliability." Journal of Bridge Engineering 9, no. 4 (2004): 403-413. https://doi.org/10.1061/(ASCE)1084-0702(2004)9:4(403).

Rashidi Nasab, Armin, and Hazem Elzarka. "Optimizing machine learning algorithms for improving prediction of bridge deck deterioration: A case study of Ohio bridges." Buildings 13, no. 6 (2023): 1517. https://doi.org/10.3390/buildings13061517.

Samsami, Reihaneh. "Artificial Intelligence (AI) Agents in Construction: A Conceptual Framework Across Autonomy Types and Application Domains." Engrxiv (2026). https://doi.org/10.31224/7439.

Washer, Glenn, Robert Connor, Massoud Nasrollahi, and Jason Provines. "New framework for risk-based inspection of highway bridges." Journal of Bridge Engineering 21, no. 4 (2016): 04015077. https://doi.org/10.1061/(ASCE)BE.1943-5592.0000818.

Zhang, Jiansong, and Nora M. El-Gohary. "Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking." Journal of computing in civil engineering 30, no. 2 (2016): 04015014. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000346.

Zhang, Jiansong, and Nora M. El-Gohary. "Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking." Automation in construction 73 (2017): 45-57. https://doi.org/10.1016/j.autcon.2016.08.027.