Grounding Large Language Models (LLMs) in Agency Standards: A Controlled Comparison of Closed-Book, Retrieval-Augmented, and Full-Document Question Answering for Construction Inspection

Reihaneh Samsami*, Ph.D., P.E.
Assistant Professor and MSCM Program Coordinator
Department of Civil and Environmental Engineering
University of New Haven, West Haven, CT 06516
Email: rsamsami@newhaven.edu

Khaled Sayed, Ph.D.
Assistant Professor
Department of Electrical & Computer Engineering and Computer Science
University of New Haven, West Haven, CT, 06517
Email: ksayed@newhaven.edu

Total pages: 17

Submission date: 7/9/2026
*Corresponding Author

ABSTRACT

Objectives: Construction inspectors on transportation projects must verify work against a fragmented body of specifications under schedule pressure. Transportation agencies are now deciding how, and whether, to deploy Large Language Models (LLMs) to support this task. This study examines whether an LLM must be grounded in agency documents at all, and if so, whether grounding should take the form of Retrieval-Augmented Generation (RAG) or simply providing the full source document within the model's context window.

Methods: The effect of document grounding was isolated through a controlled, within-question comparison. Three contemporary LLMs from different model families each answered 70 validated inspection questions, drawn from seven public Connecticut Department of Transportation (CTDOT), Occupational Safety and Health Administration (OSHA), and Federal Highway Administration (FHWA) documents, under three conditions: closed-book, RAG, and full-document context. Responses were scored with token-level F1, ROUGE-L, and semantic similarity, together with an LLM-judge rubric validated against blinded ratings by a licensed professional engineer. Consistency across systems was quantified with the Gini coefficient.

Findings: Grounding proved decisive: relative to closed-book prompting, RAG raised the judge composite by 1.31 to 1.48 points on the five-point scale and full-document context by 1.40 to 1.66 points ($p < 0.000001$), while 56 to 64 percent of closed-book answers were judged factually unacceptable. The two grounding strategies were comparable, with full-document context ahead by average 0.14 points despite requiring roughly 35 times the input tokens. Grounding gains were largest for project-specific documents and smallest for widely published standards ($r = -0.83$).

Novelty: This is the first controlled study to isolate the causal effect of document grounding on transportation inspection question answering, and to compare RAG against full-document prompting for agency standards.

Practical Applications: The findings offer agencies evidence-based guidance on when retrieval infrastructure is warranted for inspection support and when simpler deployment suffices.

INTRODUCTION

The condition of transportation infrastructure in the U.S. continues to demand more effective inspection and compliance practices. The American Society of Civil Engineers (ASCE) assigned the nation's infrastructure an overall grade of "C" in its most recent report card, citing persistent needs across roads and bridges (ASCE 2025). Construction inspection is one of the principal quality assurance mechanisms available to transportation agencies, yet the information environment in which inspectors operate is fragmented. A single highway project may be governed simultaneously by state standard specifications, supplemental project specifications, federal regulations, and safety requirements, and inspectors frequently resolve questions in the field by manually searching across these documents under time pressure. Prior work by the authors demonstrated that a RAG framework grounded in agency documents can outperform general-purpose chatbots on this task (Sayed and Samsami 2026).

Since that work, the deployment landscape has shifted in two ways that motivate the present study. First, general-purpose LLMs have improved rapidly, raising the question of whether document grounding still provides a measurable advantage over a capable model answering from parametric knowledge alone. Second, context windows have expanded to the point where entire specification documents can be supplied directly in a prompt, which offers agencies a grounding strategy that requires no retrieval infrastructure, no vector database, and no chunking design decisions. General-domain studies comparing RAG with long-context prompting have reached mixed conclusions, finding that long-context prompting can match or exceed RAG in accuracy while RAG retains a substantial cost advantage (Li et al. 2024; Li et al. 2025). Whether those findings transfer to the specification-driven, terminology-dense question answering that characterizes construction inspection has not been established.

The earlier comparison in Sayed and Samsami (2026), like most system-versus-chatbot evaluations in the construction literature, confounded two variables: the underlying model and the presence of grounding. A RAG system built on one model compared against different models without retrieval cannot attribute performance differences to retrieval itself. This study removes that confound with a within-question, within-model design: the same three LLMs, drawn from three different model families, answer the same 70 inspection questions under three information conditions, namely closed-book (question only), RAG (question plus top-ranked retrieved passages), and full-document (question plus the complete source document). Because every question is answered nine times, the marginal effect of each grounding strategy can be measured as a paired difference within each model.

This study makes four contributions to construction inspection practice and to the growing literature on LLMs in transportation. First, it isolates the causal effect of document grounding on inspection question answering through a controlled factorial design rather than a system-level comparison. Second, it contributes a benchmark of 70 expert-validated question and reference-answer pairs spanning seven public agency documents that cover materials specifications, safety standards, quality assurance procedures, and project-specific drawings. Third, it pairs reproducible automated metrics with an LLM-judge rubric whose reliability is verified against blinded ratings by a licensed professional engineer, providing a template for defensible evaluation in agency settings. Fourth, it translates the results into deployment guidance for transportation agencies weighing retrieval infrastructure against simpler long-context prompting.

LITERATURE REVIEW

LLMs in Construction and Transportation

LLMs have moved quickly from exploratory demonstrations to a substantial research area within construction informatics. A recent systematic review of the field maps applications across compliance and governance, coordination, design and planning, operations, and education, and identifies domain grounding and evaluation practice as recurring challenges (Gao et al. 2026). Earlier natural language processing (NLP) work in construction targeted regulatory text specifically: deep learning models were used to extract procedural constraints from construction regulations (Zhong et al. 2020), and customized part-of-speech tagging improved the interpretation of building codes whose language defeats general-purpose tools (Xue and Zhang 2021). A critical review of GPT models across the construction project lifecycle identified regulatory compliance verification and inspection support among the most promising application areas while cautioning that reliability in domain-specific use remains unproven (Saka et al. 2024). LLM-based approaches have since been applied to literature synthesis in off-site construction (Jeong et al. 2024) and, in the transportation domain, to inspection support through the SMART-Inspect framework, which grounded an open-weight model in Connecticut Department of Transportation (CTDOT), Occupational Safety and Health Administration (OSHA), and Federal Highway Administration (FHWA) documents and outperformed general-purpose chatbots on rubric-scored inspection queries (Sayed and Samsami 2026). In the bridge domain specifically, a comprehensive review of AI across the bridge management lifecycle positioned LLMs as the connective layer between inspection narratives, standards, and decision support, while its own demonstration documented hallucinated maintenance recommendations for components absent from the input, emphasizing the reliability problem for ungrounded models in inspection contexts (Kumar and Agrawal 2025). Complementing such reviews, a bridge engineering benchmark and domain-specialized model showed that even a fine-tuned domain LLM answered maintenance questions more professionally when combined with retrieval over agency documents (Guo et al. 2025).

Grounding Strategies: Retrieval Augmentation and Long Context

RAG was introduced as a way to combine parametric knowledge with non-parametric access to an external corpus, improving performance on knowledge-intensive tasks while providing provenance for generated content (Lewis et al. 2020). In construction, the choice between grounding strategies has begun to receive direct study: a comparison of RAG against fine-tuning for construction safety knowledge retrieval found retrieval augmentation to be the more practical adaptation route for safety management question answering (Lee et al. 2024), and a hybrid search engine combining keyword and embedding retrieval improved question-answering-based construction quality checks over standard RAG pipelines (He et al. 2025). Within transportation agencies specifically, a multi-agent RAG assistant grounded in more than 500 state DOT technical documents achieved high retrieval precision on pavement management queries, illustrating institutional interest in document-grounded assistants for agency knowledge management and workforce training (Amaram et al. 2026). The alternative to retrieval that has emerged most recently is long-context prompting, in which the source material is supplied wholesale to models whose context windows now accommodate hundreds of pages. Benchmarking studies in the general domain report that sufficiently resourced long-context prompting can consistently outperform RAG on average while costing substantially more per query (Li et al. 2024), whereas subsequent work with stronger benchmark controls concluded that neither approach dominates and that the better choice depends on task type, document length, and model (Li et al. 2025). These mixed findings, obtained on general corpora such as novels and financial filings, leave open how the trade-off resolves for agency specifications, whose distinguishing features include dense cross-references, tabulated requirements, and terminology whose precise usage carries compliance weight.

Evaluating Generated Answers

Evaluation of generated text against reference answers has traditionally relied on lexical overlap measures, including token-level F1 as popularized by reading comprehension benchmarks (Rajpurkar et al. 2016) and Recall Oriented Understudy for Gisting Evaluation (ROUGE), which measures longest common subsequence agreement (Lin 2004). These metrics are reproducible but penalize legitimate paraphrase, which motivates complementing them with embedding-based measures such as BERTScore that credits meaning-preserving rephrasings (Zhang et al. 2019), and with the LLM-as-a-judge paradigm, in which a strong model scores responses against a rubric. Systematic study of LLM judges found agreement with human experts exceeding eighty percent on open-ended tasks, while documenting position, verbosity, and self-enhancement biases that require mitigation (Zheng et al. 2023). Rubric-conditioned judging with chain-of-thought prompting further improved alignment with human judgment (Liu et al. 2023). Two mitigations from this literature are adopted in the present study: the judge model is drawn from outside the set of compared systems to avoid self-enhancement bias, and judge reliability is verified against blinded ratings from a domain expert.

Research Gap

Three gaps emerge from this literature. First, existing construction and transportation studies typically compare a proposed grounded system against different ungrounded systems, leaving the contribution of grounding itself unidentified. Second, the RAG-versus-long-context question, actively debated in the general NLP literature, has not been examined on transportation agency documents, where the answer directly determines whether agencies need retrieval infrastructure. Third, evaluation in the construction LLM literature rarely validates automated or LLM-based scoring against domain experts, limiting the defensibility of reported results for agency decision making. The present study addresses all three gaps.

METHODOLOGY

The study is a within-question factorial experiment. Three LLMs, one from each of three model families, answer each of 70 inspection questions under three information conditions, producing 630 responses. Because conditions vary within model and within question, the effect of grounding is estimated as a paired difference, controlling for both question difficulty and model capability. All experimental artifacts, including every raw response with its retrieval provenance, are logged to support reproduction. The workflow comprises (1) corpus and question bank preparation, (2) response generation under the three conditions, and (3) multi-metric evaluation with expert validation, described in turn below.

Corpus and Question Bank

The corpus consists of seven publicly available documents used in transportation construction inspection practice, published by CTDOT, OSHA, and FHWA, and summarized in **Table 1** (FHWA 2012; CTDOT 2012; CTDOT 2019; CTDOT n.d.; OSHA 2002; OSHA 2015; OSHA n.d.). The documents span materials specifications, safety standards and training instruments, quality assurance procedures, and project-specific engineering drawings, providing coverage of the document genres an inspector consults. One document, a set of fall protection drawings from a Connecticut bridge project, contains no machine-readable text layer; its content was recovered with optical character recognition (OCR) at 300 dpi and manually verified against the source before inclusion.

For each document, ten question and reference-answer pairs were developed and verified against the source text, with the source page recorded for each pair, yielding 70 pairs in total. Questions were labeled by query type as explicit, where the answer is stated directly in the document (for example, a required overlap dimension or a testing frequency), or interpretive, where answering requires synthesis across provisions or conditional reasoning. This labeling supports analysis of whether grounding benefits depend on question character. Questions were written to reflect the targeted, standards-driven inquiries an inspector poses during active construction, ranging from single-value lookups such as a required plank overlap or a fall-protection height threshold to procedural questions such as the responsibilities of satellite testing laboratories. Reference answers were drawn verbatim or near-verbatim from the source text so that scoring rewards fidelity to the governing document rather than plausible-sounding generalities.

TABLE 1 Corpus Documents and Question Bank Composition (10 Questions per document)

Document	Agency	Genre
Traffic Sign Retroreflective Sheeting Identification Guide	FHWA	Materials identification guide
Hot Mix Asphalt Project Specifications (Meriden, CT)	CTDOT	Project materials specification
Acceptance and Assurance Testing Policies and Procedures	CTDOT	Quality assurance procedure

TABLE 1 Corpus Documents and Question Bank Composition (10 Questions per document) (Continued)

Document	Agency	Genre
Before/After Fall Protection Drawings (Project 304-008)	CTDOT	Project engineering drawings
Fall Protection Pre/Post Test	OSHA	Safety training instrument
Construction Industry Digest (OSHA 2202)	OSHA	Safety standards digest
Guide to Scaffolds in the Construction Industry (OSHA 3150)	OSHA	Safety standard guide

Models and Information Conditions

Three commercially available LLMs were selected, one from each of three major model families: Claude Haiku 4.5 (Anthropic), GPT-4o mini (OpenAI), and Gemini 2.5 Flash (Google). Small and mid-tier models were selected deliberately: they represent the price point at which agency-scale deployment is plausible, and the grounding question is most consequential precisely where parametric knowledge is weakest. All models were queried through their public application programming interfaces in June - July 2026 with temperature 0.1 and a shared system prompt instructing the model to answer as an assistant to a transportation construction inspector and to state when an answer is unavailable.

Each question was posed to each model under three conditions. In the closed-book condition, the model receives only the question. In the RAG condition, the source documents are segmented into chunks of up to 700 characters with 300-character overlap, chunk boundaries preferring sentence breaks; chunks are embedded with a sentence-transformer encoder (Reimers and Gurevych 2019), and the five chunks most similar to the question by cosine similarity, restricted to the question's source document to mirror an inspector querying a known specification, are supplied as context in the prompt format of Sayed and Samsami (2026). In the full-document condition, the complete text of the source document is supplied as context; the longest corpus document, the 247-page acceptance and assurance manual, produced prompts of roughly 175,000 tokens that exceeded practical context limits during piloting and was therefore excluded from this condition for all three models, a constraint revisited in the Discussion. Every response is recorded in an append-only provenance log containing the model identifier, condition, retrieved chunk identifiers and similarity scores where applicable, and timestamps. An automated integrity validator screens the completed log for duplicated response text across distinct questions, empty responses, and interface errors before any scoring is permitted, a safeguard adopted after data assembly errors were identified in a predecessor dataset.

Evaluation Design

Each response is evaluated against its reference answer with three automated metrics and one rubric-based judgment. The automated metrics are token-level F1 (Rajpurkar et al. 2016), ROUGE-L (Lin 2004), and cosine similarity between sentence-transformer embeddings of the response and the reference (Reimers and Gurevych 2019; Zhang et al. 2019).

Each automated metric is defined over tokens shared between the model response and the reference answer. Token F1 is the harmonic mean of precision P, the fraction of response tokens found in the reference, and recall R, the fraction of reference tokens found in the response (**Equation 1**). ROUGE-L applies the same form to the longest common subsequence (LCS) between response X and reference Y, with precision and recall normalized by response and reference length (**Equation 2**). Semantic similarity is the cosine of the angle between the sentence-transformer embedding vectors a and b of the response and reference (**Equation 3**).

$$F_1 = \frac{2PR}{P+R} \quad (1)$$

$$ROUGE - L = \frac{2PR}{P+R}, \text{ where } P = \frac{LCS(X,Y)}{|X|} \text{ and } R = \frac{LCS(X,Y)}{|Y|} \quad (2)$$

$$\cos(a, b) = (a \cdot b) / (\|a\| \, \|b\|) \quad (3)$$

The rubric-based judgment applies the weighting of the authors' prior evaluation framework: an LLM judge assigns integer scores from 1 to 5 for accuracy, relevance, clarity, and terminology, combined as shown in **Equation 4**.

$$Score = 0.5(Accuracy) + 0.25(Relevance) + 0.15(Clarity) + 0.10(Terminology) \quad (4)$$

The weighting prioritizes accuracy at fifty percent, reflecting the safety-critical and compliance-critical nature of inspection decisions, followed by relevance, clarity, and correct use of technical terminology (Sayed and Samsami 2026). The judge model is Claude Sonnet 4.6, which is not among the three compared systems, mitigating the self-enhancement bias documented for LLM judges (Zheng et al. 2023). Judge reliability was assessed against blinded human ratings. A stratified random sample of 75 responses, drawn evenly across model and condition cells and supplemented after the full run to cover all documents, was independently rated by a licensed professional engineer blinded to the generating model and condition. Agreement between the expert and the judge was strong: Spearman rank correlation of 0.77 ($p < 0.001$), Pearson correlation of 0.80, mean absolute difference of 0.60 points, and 82.7 percent of items within one point, with the judge averaging 0.41 points stricter than the expert, a conservative direction for the safety-critical use case. Worked examples showing rated model responses appear in **Tables 5 through 7**.

Finally, because inspection decision making requires consistency in addition to average quality, the dispersion of composite scores across the three models is quantified per question and condition with the Gini coefficient (Dorfman 1979), following the interpretation framework of the authors' prior work: values near zero indicate that systems agree in quality, and larger values indicate that the choice of system materially affects the guidance an inspector receives. For each question and condition, the coefficient is computed over the three models' composite scores sorted in ascending order (**Equation 5**).

$$G = \left[\frac{2 \cdot \Sigma_i \, i \cdot x_i}{n \cdot \Sigma_i \, x_i}\right] - \frac{n+1}{n} \quad (5)$$

RESULTS

The completed experiment comprises 600 scored responses: 70 questions answered by three models under closed-book and RAG conditions, and 60 questions (excluding the acceptance and assurance manual) under the full-document condition. Results are reported in four parts: overall performance, the paired grounding effect, performance by document, and cross-system consistency. Judge composite scores on the five-point scale of **Equation 4** serve as the primary metric; the automated lexical and semantic metrics agree with the judge on every ordinal comparison reported below and are tabulated alongside it.

Overall Performance by Model and Condition

Table 2 reports mean scores by model and condition, and **Figure 1** displays the judge composite. The dominant structure in the data is the condition effect, not the model effect: the three models cluster within 0.11 points of one another in the closed-book condition (2.64 to 2.74), within 0.22 points under RAG (3.95 to 4.17), and within 0.23 points under full-document context (4.22 to 4.45), while the gap between conditions is roughly 1.5 points. The lexical metrics show one divergence worth noting: Gemini scores markedly higher on token F1 and ROUGE-L than the other models at equal judge composite, indicating that it reproduces reference wording more literally while Claude and GPT paraphrase more, a stylistic difference the semantic and judge measures largely discount.

TABLE 2 Mean Scores by Model and Information Condition

Model	Condition	Judge Composite (1-5)	Token F1	ROUGE-L	Semantic Similarity
Claude Haiku 4.5	Closed-book	2.64	0.14	0.10	0.55
	RAG	4.12	**0.27**	**0.23**	**0.65**
	Full document	**4.42**	0.23	0.20	**0.65**
GPT-4o mini	Closed-book	2.74	0.19	0.14	0.58
	RAG	4.17	**0.37**	**0.32**	0.68
	Full document	**4.45**	0.36	0.31	**0.69**
Gemini 2.5 Flash	Closed-book	2.64	0.23	0.18	0.57
	RAG	3.95	0.47	0.43	0.71
	Full document	**4.22**	**0.49**	**0.45**	**0.72**

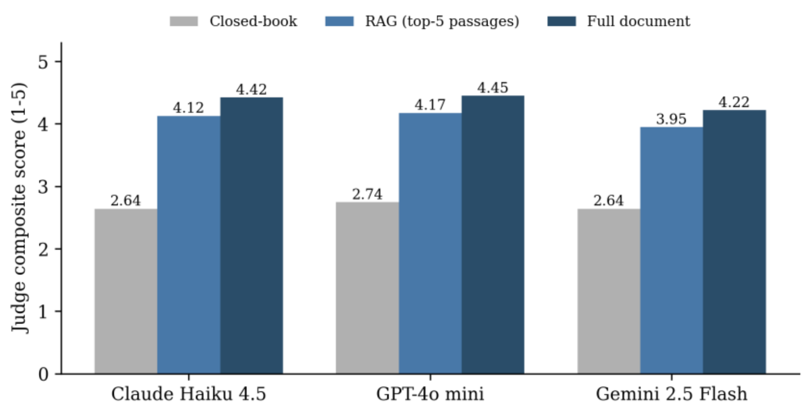

Figure 1 Judge composite score by model and information condition (n = 70 questions for closed-book and RAG; n = 60 for full document).

The Grounding Effect

Table 3 reports the paired contrasts. Grounding produced large, consistent, and statistically decisive gains for every model: RAG improved the judge composite by 1.31 to 1.48 points over closed-book prompting, improving 77 to 87 percent of individual questions, and full-document context improved it by 1.40 to 1.66 points, improving 77 to 93 percent, with Wilcoxon signed-rank p-values below 0.0000002 for all six contrasts. The practical meaning of the closed-book baseline deserves emphasis: 56 to 64 percent of closed-book responses received judge accuracy scores of 2 or below, meaning a majority of ungrounded answers to these specification-driven questions were factually unacceptable, frequently delivered in fluent and confident prose. Grounding with RAG reduced that unacceptable-answer rate to 7 percent for Claude, 10 percent for GPT, and 19 percent for Gemini.

The comparison between the two grounding strategies is far closer. Full-document context outperformed RAG by 0.17 points for Claude ($p = 0.019$), 0.11 for GPT ($p = 0.19$), and 0.15 for Gemini ($p = 0.52$) on the 60-question subset where both were feasible, a difference that reached significance for one model of three. Against this marginal quality edge stand two costs: full-document prompts averaged roughly 35 times the input tokens of RAG prompts in this corpus (approximately 31,000 versus 900 tokens of context per query), and the strategy was infeasible for the longest document at all, whereas RAG answered the acceptance manual's questions at a composite of 3.25, its lowest per-document score but far above the 1.83 closed-book baseline.

TABLE 3 Paired Condition Contrasts on Judge Composite Score

Model	Contrast	n	Mean Difference	Standard Deviation	Questions Improved	Wilcoxon p
Claude Haiku 4.5	RAG vs. closed-book	70	+1.48	1.00	87%	< 0.000001
	Full doc vs. closed-book	60	+1.66	1.12	92%	< 0.000001
	Full doc vs. RAG	60	+0.17	0.65	33%	0.019
GPT-4o mini	RAG vs. closed-book	70	+1.43	1.13	83%	< 0.000001
	Full doc vs. closed-book	60	+1.58	0.99	93%	< 0.000001
	Full doc vs. RAG	60	+0.11	0.54	23%	0.19
Gemini 2.5 Flash	RAG vs. closed-book	70	+1.31	1.62	77%	< 0.000001
	Full doc vs. closed-book	60	+1.40	1.53	77%	< 0.000001
	Full doc vs. RAG	60	+0.15	1.08	32%	0.52

Performance by Document

Table 4 disaggregates performance by corpus document, averaging over the three models, and Figure 2 plots the RAG gain against the closed-book baseline. The pattern is strongly systematic: the documents on which ungrounded models performed worst are the documents on which grounding helped most ($r = -0.83$, $p = 0.020$). Closed-book performance was highest for the widely published OSHA safety materials, whose content saturates public training corpora, and lowest for the project-specific documents, the fall protection drawings for a specific Connecticut bridge project (1.70) and the state acceptance and assurance procedures (1.83), whose contents no model could plausibly have memorized. RAG lifted the drawings document by 2.37 points, the largest single-document gain in the study, and it bears noting that this document entered the corpus only as OCR text recovered from scanned engineering sheets. The acceptance manual remained the hardest document even when grounded, suggesting that long procedural documents with dense cross-references challenge retrieval quality itself, not merely context capacity.

A query-type analysis was planned but proved underpowered: the question bank is dominated by explicit questions whose answers are stated directly in the source (66 of 70), leaving only four interpretive questions. Mean RAG composites for the two types were nearly identical (4.09 versus 4.02), but no inference is drawn from so imbalanced a split, and the imbalance is noted as a limitation.

TABLE 4 Judge Composite by Document and Condition (Mean Across Models)

Document	Closed-book	RAG	Full Document	RAG Gain
Fall Protection Drawings (project-specific)	1.70	4.07	4.45	+2.37
Acceptance & Assurance Procedures(CTDOT)	1.83	3.25	excluded	+1.42
Sign Sheeting Guide (FHWA)	2.30	4.43	4.38	+2.13
HMA Specifications (Meriden project)	2.80	4.29	4.37	+1.49
Scaffolds Guide (OSHA 3150)	2.78	3.87	3.94	+1.09
Construction Industry Digest (OSHA 2202)	3.49	4.04	4.32	+0.55
Fall Protection Quiz (OSHA)	3.82	4.61	4.71	+0.79

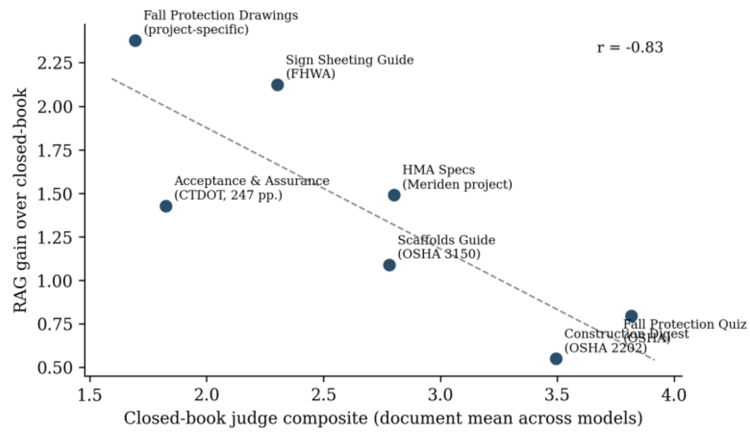

Figure 2 RAG gain versus closed-book baseline by corpus document (document means across the three models; r = -0.83, p = 0.020).

A Worked Example: How Grounding Changes Model Answers

The document on which grounding helped most, a set of before-and-after fall protection drawings for a specific Connecticut bridge rehabilitation (CTDOT Project 304-008), illustrates the mechanism behind the aggregate results. Its content is entirely project-specific, so no model could have seen it during training, and it entered the corpus only as text recovered through OCR from scanned engineering sheets. **Tables 5 through 7** show the verbatim responses of all three models, closed-book and then RAG-grounded, to three of its ten questions, alongside the reference answer and the expert accuracy score each response received.

TABLE 5 Sample Model Responses and Scores: Tie-Off Cable Diameter

Question: What diameter tie off cable is used for fall protection?
Reference answer: a 3/8" diameter tie off cable.

Model	Condition	Response	Score
Claude Haiku 4.5	Closed-book	Fall Protection Tie-Off Cable Diameter The standard diameter for fall protection tie-off cables is 5/8 inch (16 mm). This diameter is specified in ANSI/ASSE standards and OSHA...	1
Claude Haiku 4.5	RAG	Fall Protection Tie Off Cable Diameter According to the CTDOT bridge fall protection documentation, a 3/8" diameter tie off cable is used for fall protection. Additional Requi...	5
GPT-4o mini	Closed-book	The diameter of tie-off cables used for fall protection typically ranges from 5/16 inch to 3/8 inch. However, the specific diameter may vary based on the manufacturer's specif...	3
GPT-4o mini	RAG	The diameter of the tie-off cable used for fall protection is 3/8 inches.	5
Gemini 2.5 Flash	Closed-book	There is no single standard diameter for a "tie-off cable" used for fall protection. The appropriate diameter is determined by the engineered design of the specific fall prote...	1

TABLE 5 Sample Model Responses and Scores: Tie-Off Cable Diameter (Continued)

Model	Condition	Response	Score
Gemini 2.5 Flash	RAG	A 3/8-inch diameter tie-off cable is used for fall protection.	5

TABLE 6 Sample Model Responses and Scores: Engineer of Record

Question: What is the name of the Engineer that is listed on the second drawing? **Reference answer:** Lochner.			
Model	**Condition**	**Response**	**Score**
Claude Haiku 4.5	Closed-book	I don't have access to any drawings or documents to review. To answer your question about the Engineer listed on the second drawing, please share the drawing or document with ...	1
Claude Haiku 4.5	RAG	Based on the context provided, the Engineer listed on the drawings is Lochner. This information appears consistently across the multiple drawing sheets referenced in the ConnD...	5
GPT-4o mini	Closed-book	I'm sorry, but I do not have access to specific drawings or documents, so I cannot provide the name of the Engineer listed on the second drawing.	1
GPT-4o mini	RAG	The name of the Engineer listed on the second drawing is Lochner.	5
Gemini 2.5 Flash	Closed-book	The answer is not available.	1
Gemini 2.5 Flash	RAG	The information provided does not list the Engineer for the second drawing.	1

TABLE 7 Sample Model Responses and Scores: Plank Load Rating

Question: What is the PSF load rating for the planks with a 7-foot spacing? **Reference answer:** 50PSF			
Model	**Condition**	**Response**	**Score**
Claude Haiku 4.5	Closed-book	I don't have access to specific plank load rating data or specifications. To answer this question accurately, I would need: 1. The plank material type (wood species, composite...	1
Claude Haiku 4.5	RAG	PSF Load Rating for Planks with 7 Foot Spacing Based on the provided documentation, the allowable load rating for planks with a 7 foot (7') spacing is 50 PSF. This specificati...	5
GPT-4o mini	Closed-book	The PSF (pounds per square foot) load rating for planks with a 7-foot spacing can vary based on several factors, including the type of material, the dimensions of the planks, ...	1

TABLE 7 Sample Model Responses and Scores: Plank Load Rating (Continued)

Model	Condition	Response	Score
GPT-4o mini	RAG	The PSF load rating for the planks with a 7-foot spacing is 50 PSF.	5
Gemini 2.5 Flash	Closed-book	The PSF load rating for planks cannot be determined without additional information. To calculate the allowable load, the following details are required: Plank Material: (e.g.,...	1
Gemini 2.5 Flash	RAG	The PSF load rating for the planks with a 7-foot spacing is 50 PSF.	5

Three patterns recur across these tables and hold across the full corpus. First, closed-book grounding failures take two forms, and the tables show both: outright fabrication, as when a model states a specific but incorrect cable diameter or load figure it cannot possibly know, and appropriate refusal, as when a model reports that the information is unavailable. Refusal earns a low accuracy score because the question went unanswered, but it is the safe failure; fabrication is the dangerous one, because a fluent, confident, wrong answer about a tie-off capacity or a load rating is precisely what could mislead an inspector. In the tie-off cable question, the three ungrounded models produced three different wrong answers, one asserting five-eighths inch, one hedging across a range, one denying any standard exists, a divergence that itself signals unreliability. Second, grounding largely resolves both failure modes: supplied with the drawings, the models converge on the correct three-eighths-inch diameter and the correct fifty-pound-per-square-foot rating, and accuracy scores jump from 1 to 5. Third, retrieval is imperfect, visible in the engineer-of-record question, where two models recovered the name Lochner from the retrieved passages but the third did not, because the relevant text was not among its top-ranked chunks; full-document context, which places the entire drawing set before the model, recovered it in every case. This is the concrete form of the small average advantage full-document context holds over RAG.

These tables also demonstrate the scoring itself, and thus the basis for trusting the LLM judge. A reviewer can read any response above and form an independent judgment of its accuracy; the expert did exactly this for 75 such responses, blind to model and condition. The correspondence was strong, Spearman 0.77 and Pearson 0.80 against the judge, with a mean absolute difference of 0.60 points and 82.7 percent of items within one point, the judge running 0.41 points stricter on average, a conservative bias for a safety-critical application. The complete set of rated responses accompanies this paper as supplementary data.

Cross-System Consistency

Consistency across the three systems was quantified with the per-question Gini coefficient over judge composite scores. Disparity was modest in all conditions and decreased with grounding: mean Gini fell from 0.062 closed-book to 0.053 under RAG and 0.047 under full-document context, values in the near-equality band of the interpretation framework of Sayed and Samsami (2026). Win counts tell the same story from another angle: closed-book, Claude was the best-scoring system on 36 of 70 questions, more than the other two combined, whereas under grounding the wins spread more evenly (RAG: GPT 31, Claude 22, Gemini 17). Grounding thus not only raised all systems but compressed the differences among them.

DISCUSSION AND PRACTICAL IMPLICATIONS

Three questions structure the interpretation of these results, each with direct consequences for transportation agencies.

First, does grounding matter? The paired design permits a direct answer: decisively, for every model tested. The roughly 1.5-point composite gain is the difference between a system whose majority of answers are factually unacceptable and one whose answers are largely usable, and the failure mode on the wrong side of that line is the dangerous one: closed-book models rarely declined to answer, instead producing fluent, confidently worded, and wrong responses to questions about tie-off capacities, testing frequencies, and material specifications, a failure mode documented even in purpose-built inspection assistants (Kumar and Agrawal 2025). The improvement of general-purpose models has not closed this gap; parametric knowledge simply does not contain project specifications and state procedures. Agencies should treat ungrounded chat interfaces as unsuitable for specification-driven inspection questions regardless of how capable the underlying model appears.

Second, which grounding strategy should agencies choose? Retrieval infrastructure carries development and maintenance costs but restricts model attention to relevant passages; full-document prompting requires no infrastructure but incurs higher per-query cost and hard capacity limits. The results place the quality difference between them at 0.11 to 0.17 points, significant for one model of three, while RAG delivered 93 to 94 percent of full-document quality at roughly 3 percent of the context tokens and was the only strategy able to ground the 247-page acceptance manual at all. The practical translation is that full-document prompting is a legitimate lightweight pattern for short, self-contained documents and ad hoc use, but RAG is the appropriate default for agency-scale corpora, both because specification libraries routinely exceed even current context windows and because per-query token costs compound across an inspection workforce. These findings extend the mixed general-domain evidence (Li et al. 2024; Li et al. 2025) with a domain-specific answer: for transportation standards, the strategies are near-equivalent in quality where both are feasible, and feasibility, not quality, is the deciding factor.

Third, where does grounding help most? The strong inverse relationship between closed-book performance and grounding gain ($r = -0.83$) gives agencies a prioritization rule: grounding investment pays most where content is least represented in public training data, which is precisely the project-specific material, drawings, project specifications, and state procedures, that carries the most operational and contractual weight. Widely published federal safety standards showed the smallest gains because models already partially know them, though even there grounding improved every document. The fall protection drawings result is particularly encouraging for practice: a document that exists only as scanned engineering sheets, recovered through optical character recognition, went from the worst closed-book score in the corpus to a usable 4.07 under RAG, indicating that even non-digital legacy documents can be brought into a grounded assistant with modest preprocessing.

The consistency analysis speaks to procurement. Inter-model disparity was modest to begin with and shrank under grounding, and best-system honors spread across all three vendors once documents were supplied. For agencies, this decouples the deployment decision from the model-selection decision: the choice that determines outcome quality is whether and how to ground, while the choice among major model families can be driven by cost, data governance terms, and integration fit, subject to periodic re-evaluation as models evolve. Practical adoption will additionally depend on data governance, integration with inspection management systems, and inspector trust in AI-assisted recommendations, considerations consistent with those identified in the authors' prior work (Sayed and Samsami 2026). Limitations of the present study include a corpus drawn from a single state's practice environment, reference answers authored by the research team, evaluation at a single point in the rapid evolution of commercial models, a question bank dominated by explicit factual queries that leaves interpretive questions underpowered, and a text-only treatment of one drawing-based document.

CONCLUSION

This study isolated the effect of document grounding on LLM-based construction inspection question answering through a within-question factorial design: three models from three families answered 70 expert-validated questions from seven public transportation agency documents under closed-book, RAG, and full-document conditions, with evaluation combining reproducible automated metrics, an expert-validated LLM judge, and Gini-based consistency analysis. Grounding was decisive, raising composite quality by 1.3 to 1.7 points on a five-point scale and cutting the rate of factually unacceptable answers from a majority of responses to a small minority, while the choice between grounding strategies proved secondary: full-document context held a marginal quality edge where it fit, but RAG matched 93 to 94 percent of its quality at a small fraction of the cost and was the only strategy able to serve the longest document. Gains concentrated on project-specific documents, giving agencies a clear prioritization rule for grounding investment, and grounding compressed inter-model differences, decoupling deployment decisions from vendor selection. Future work will extend the question bank toward interpretive and multi-document reasoning, evaluate the framework with practicing inspectors in the field, treat drawing-based documents multimodally rather than through optical character recognition alone, and re-evaluate the grounding trade-off longitudinally as context windows and model knowledge continue to expand.

AUTHOR CONTRIBUTIONS

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

DECLARATION OF CONFLICTING INTERESTS

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The authors disclosed no financial support for the research, authorship, and/or publication of this article.

REFERENCES

Amaram, Divija, Lu Gao, Gowtham Reddy Gudla, and Tejaswini Sanjay Katale. "Developing an AI Assistant for Knowledge Management and Workforce Training in State DOTs." arXiv preprint arXiv:2603.03302. (2026).

American Society of Civil Engineers (ASCE). “Report Card for America's Infrastructure.” American Society of Civil Engineers. (2025). https://infrastructurereportcard.org.

Connecticut Department of Transportation (CTDOT). “Hot Mix Asphalt (HMA) Specifications, Meriden, Connecticut Project.” Connecticut Department of Transportation. (2012)

Connecticut Department of Transportation (CTDOT). “Acceptance and Assurance Testing Policies and Procedures.” Connecticut Department of Transportation. (2019)

Connecticut Department of Transportation (CTDOT). "Before/After Fall Protection Drawings, ConnDOT Project 304-008, Town of Seymour, Connecticut." Connecticut Department of Transportation. (n.d.)

Dorfman, Robert. "A formula for the Gini coefficient." The review of economics and statistics (1979): 146-149. https://doi.org/10.2307/1924845.

Federal Highway Administration (FHWA). “Traffic Sign Retroreflective Sheeting Identification Guide.” Federal Highway Administration. (2012)

Gao, Yu, Tak Wing Yiu, Xuesong Shen, and Vivian WY Tam. "Large language models in smart construction: a systematic review of implementation strategies, applications and future directions." Engineering, Construction and Architectural Management 33, no. 15 (2026): 159-181. https://doi.org/10.1108/ECAM-10-2025-1668.

Guo, Cheng, Hu Kai, Shuxian Liang, Yiyang Jiang, Yi Gao, Xian-Sheng Hua, and Wei Dong. "SDBench: A Survey-based Domain-specific LLM Benchmarking and Optimization Framework." In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 13492-13506. (2025).

He, Chuanni, Weilin He, Min Liu, Shaolong Leng, and Song Wei. "Enriched construction regulation inquiry responses: A hybrid search approach for large language models." Journal of Management in Engineering 41, no. 3 (2025): 04025001. https://doi.org/10.1061/JMENEA.MEENG-6444.

Jeong, Jaemin, Daeyoung Gil, Daeho Kim, and Jaewook Jeong. "Current research and future directions for off-site construction through LangChain with a large language model." Buildings 14, no. 8 (2024): 2374. https://doi.org/10.3390/buildings14082374.

Kumar, Deepak, and Anil Agrawal. "Advancing bridge infrastructure management through artificial intelligence: a comprehensive review." International Journal of Bridge Engineering, Management and Research 2, no. 3 (2025): 214250021-1. https://doi.org/10.70465/ber.v2i3.45.

Lee, Jungwon, Seungjun Ahn, Daeho Kim, and Dongkyun Kim. "Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval." Automation in Construction 168 (2024): 105846. https://doi.org/10.1016/j.autcon.2024.105846

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks." Advances in neural information processing systems 33 (2020): 9459-9474.

Li, Kuan, Liwen Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Shuai Wang, and Minhao Cheng. "LaRA: Benchmarking Retrieval-Augmented Generation and Long-Context LLMs--No Silver Bullet for LC or RAG Routing." arXiv preprint arXiv:2502.09977. (2025).

Li, Zhuowan, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. "Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach." In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 881-893. (2024). https://doi.org/10.18653/v1/2024.emnlp-industry.66

Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." In Text summarization branches out, pp. 74-81. (2004).

Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. "G-eval: NLG evaluation using gpt-4 with better human alignment." In Proceedings of the 2023 conference on empirical methods in natural language processing, pp. 2511-2522. (2023).

Occupational Safety and Health Administration (OSHA). “Construction Industry Digest.” Occupational Safety and Health Administration. (2015)

Occupational Safety and Health Administration (OSHA). “A Guide to Scaffold Use in the Construction Industry (OSHA 3150).” Occupational Safety and Health Administration. (2002)

Occupational Safety and Health Administration (OSHA). “Fall Protection Pre- and Post-Test and Answer Key.” Occupational Safety and Health Administration. (n.d.)

Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp. 3982-3992. (2019).

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. "Squad: 100,000+ questions for machine comprehension of text." In Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 2383-2392. (2016).

Saka, Abdullahi, Ridwan Taiwo, Nurudeen Saka, Babatunde Abiodun Salami, Saheed Ajayi, Kabiru Akande, and Hadi Kazemi. "GPT models in construction industry: Opportunities, limitations, and a use case validation." Developments in the Built Environment 17 (2024): 100300.

Sayed, Khaled, and Reihaneh Samsami. “SMART-Inspect: Standards Management and Retrieval Technology for Inspection”. In Proceedings of the Construction Research Congress 2026. Reston, VA: American Society of Civil Engineers. (2026).

Xue, Xiaorui, and Jiansong Zhang. "Part-of-speech tagging of building codes empowered by deep learning and transformational rules." Advanced Engineering Informatics 47 (2021): 101235.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675 (2019).

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin et al. "Judging llm-as-a-judge with mt-bench and chatbot arena." Advances in neural information processing systems 36 (2023): 46595-46623.

Zhong, Botao, Xuejiao Xing, Hanbin Luo, Qirui Zhou, Heng Li, Timothy Rose, and Weili Fang. "Deep learning-based extraction of construction procedural constraints from construction regulations." Advanced Engineering Informatics 43 (2020): 101003. https://doi.org/10.1016/j.aei.2019.101003.