International Conference on Industrialized Construction
Gainesville, Florida, United States | August 3-5, 2026

ConstructionQA: Evaluating Multimodal Large Language Models on Industrial Construction Drawing Set

Obiora Odugu[1]*, Mohammad Hadi Foroozandeh Dehkordi[2], Asif Faisal Chowdhury[3], Muhammad Tahir Khan[4], Chao Wang[5]

[1] Graduate Student, Department of Construction Management, Louisiana State University, USA, oodugu1@lsu.edu

[2] Graduate Student, Department of Construction Management, Louisiana State University, USA, MohammadHadi.ForoozandehDehkordi@lsu.edu

[3] Graduate Student, Department of Construction Management, Louisiana State University, USA, achowd6@lsu.edu

[4] Graduate Student, Department of Construction Management, Louisiana State University, USA, mkhan49@lsu.edu

[5] Graduate Student, Department of Construction Management, Louisiana State University, USA, chaowang@lsu.edu

ABSTRACT

Construction drawing sets are legally binding documents requiring complex multi-sheet, multi-discipline interpretation skills from practitioners. Despite the potential of multimodal large language models (MLLMs) to support plan review workflows, no systematic benchmark has evaluated their comprehension of complete construction document sets. This paper presents ConstructionQA, applied to a real 100% Construction Documents (CD) drawing set for an industrial warehouse facility, comprising approximately over 150 sheets across eight engineering disciplines. ConstructionQA structures evaluation around three progressively complex tasks: Information Extraction (IE), which tests value retrieval from single sheets using F1-BoW scoring; Sheet Reference Retrieval (SR), which tests navigation of the callout-and-reference system using F1-Sheets scoring; and Cross-Drawing Compliance (CDC), which tests multi-sheet, multi-discipline coordination reasoning using a composite score of YN Accuracy and ROUGE-L. Five MLLMs are evaluated under three input conditions: a distractor condition, in which plausible but irrelevant sheets augment the ground-truth context, a guaranteed-context condition, in which only relevant sheets are provided and a native-PDF condition. Models evaluated include GPT-5.4, Gemini 2.5 Flash, GPT-4.1 Nano, Gemma3:12b, and Qwen3-VL:8B. Results show that frontier models substantially outperform smaller models across all tasks, but performance remains insufficient for deployment in all categories. Sheet Reference Retrieval emerges as a critical bottleneck, with open-source models achieving near-zero SR scores across both conditions. These findings highlight the need for hybrid architectures that separate sheet retrieval from compliance reasoning in practical AEC deployment context

KEYWORDS

multimodal large language models; construction drawings; benchmark evaluation; information extraction; cross-drawing compliance

INTRODUCTION

Construction drawing sets serve as the primary contract document for the architecture, engineering, and construction (AEC) industry, legally binding all parties involved in building delivery (American Institute of Architects, 2017). A single commercial project may encompass more than 100 sheets spanning architectural, structural, civil, mechanical, electrical, plumbing, and fire protection disciplines. Interpreting these documents requires practitioners to navigate cross-referenced callout systems, reconcile discipline-specific annotations, and synthesize spatially distributed information. This is a cognitively demanding process whose errors cost the U.S. industry an estimated $31.3 billion annually in rework (PlanGrid and FMI Corporation, 2018).

The AEC industry has seen growing AI adoption across scheduling, cost estimation, compliance checking, and site monitoring (Pan & Zhang, 2021). Transformer-based language models have been applied to construction planning documents, with He et al. (2024) demonstrating that GPT-4 and BERT models can identify and classify construction planning constraints from meeting transcripts. Sulbaran (2024) found that MPNET outperforms BERT for construction schedule activity classification. Code compliance checking using LLM-assisted rule extraction and design verification has been explored, though reliability remains insufficient for autonomous decision-making (Zhang & El-Gohary, 2017). Despite this progress, relatively few studies address the visual interpretation of engineering drawings, which require models to process specialized graphical languages with symbols, annotations, and cross-referenced callout systems.

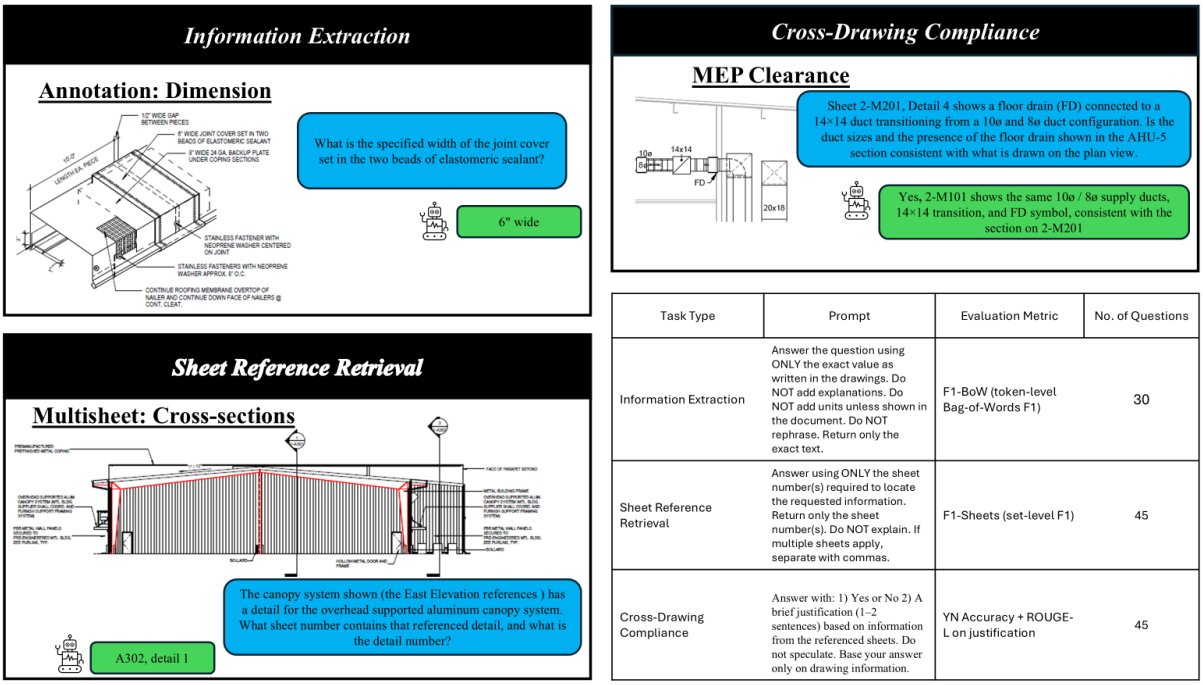
The emergence of multimodal large language models (MLLMs) has heightened interest in automating document-intensive workflows in the AEC industry (Alayrac et al., 2022). However, progress has been constrained by limited trust in model reliability and the high-cost implications of errors in construction, where misinterpretation of documents can result in substantial rework and safety risks. Despite the importance of this, there have been few study attempting to systematic evaluate MLLM comprehension of complete construction document sets. Existing general-purpose benchmarks such as MME, and MMBench treat each sheet as a standalone input and do not require models to navigate across sheets or synthesize cross-discipline information (Fu et al., 2023; Liu et al., 2023). Other domain-specific benchmarks including DesignQA (Doris et al., 2024), CEQuest (Wu et al., 2025), and AECV-Bench (Kondratenko et al., 2026) address engineering document understanding but do not evaluate the multi-sheet coordination reasoning that characterizes professional construction drawing review.

This paper introduces ConstructionQA and applies it to a100% Construction drawing set for an industrial building. ConstructionQA evaluates five MLLMs on three tasks: Information Extraction (IE), Sheet Reference Retrieval (SR), and Cross-Drawing Compliance (CDC) under distractor and guaranteed-context conditions. This design helps quantify the retrieval bottleneck separately from the reasoning capability, yielding actionable insights for deploying MLLMs in AEC workflows.

THE CONSTRUCTIONQA BENCHMARK

Drawing and Benchmark Structure

The dataset is grounded in a 100% Construction Documents drawing set for an Industrial warehouse and training facility. The facility is a two-building complex of approximately 33,000 SF comprising over 150 sheets across eight engineering disciplines: Architectural, Structural, Civil, Mechanical, Electrical, Plumbing, Interior, and Fire/Life Safety.



Task Type	Prompt	Evaluation Metric	No. of Questions
Information Extraction	Answer the question using ONLY the exact value as written in the drawings. Do NOT add explanations. Do NOT add units unless shown in the document. Do NOT rephrase. Return only the exact text.	F1-BoW (token-level Bag-of-Words F1)	30
Sheet Reference Retrieval	Answer using ONLY the sheet number(s) required to locate the requested information. Return only the sheet number(s). Do NOT explain. If multiple sheets apply, separate with commas.	F1-Sheets (set-level F1)	45
Cross-Drawing Compliance	Answer with: 1) Yes or No 2) A brief justification (1–2 sentences) based on information from the referenced sheets. Do not speculate. Base your answer only on drawing information.	YN Accuracy + ROUGE-L on justification	45

Figure 1. ConstructionQA benchmark task structure and examples. The figure illustrates the three evaluation sections: (top left) Information Extraction (IE), (bottom left) Sheet Reference Retrieval (SR), and (top right) Cross-Drawing Compliance (CDC). (Bottom right) summarizes prompts, evaluation metrics, and question counts for each section.

The benchmark comprises 120 question–answer pairs. Each question references one or more ground-truth sheets that contain the information required to answer it. Questions are organized into three task types of increasing cognitive complexity. Figure 1 shows each question task type, sample questions and response including prompt design. The task type attempts to mirror the hierarchical reasoning workflow of professional drawing review. They include:

- **Information Extraction (IE)** which tests atomic value retrieval from a single sheet (e.g., project number from a title block, room area from a floor plan annotation, equipment capacity from a mechanical schedule). Responses are scored with F1-BoW, a token-level bag-of-words F1 that awards partial credit for minor formatting variation. Let p and g denote the predicted and ground-truth token multisets respectively:

$$\text{Precision}_{\text{BoW}} = \frac{|p \cap g|}{|p|}, \quad \text{Recall}_{\text{BoW}} = \frac{|p \cap g|}{|g|}, \quad \text{F1}_{\text{BoW}} = 2 \times \frac{(\text{P}_{\text{BoW}} \times \text{R}_{\text{BoW}})}{(\text{P}_{\text{BoW}} + \text{R}_{\text{BoW}})} \qquad (1)$$

- **Sheet Reference Retrieval (SR)** which tests identification of destination sheets for callout symbols (detail bubbles, section cuts, door tags). Responses are scored with F1-Sheets, a set-level F1 over predicted sheet identifiers that rewards partial retrieval of multi-sheet answers.

$$\text{Precision}_{\text{sheets}} = \frac{|p \cap g|}{|p|}, \quad \text{Recall}_{\text{sheets}} = \frac{|p \cap g|}{|g|}, \quad \text{F1}_{\text{sheets}} = 2 \times \frac{(\text{P}_{\text{sheets}} \times \text{R}_{\text{sheets}})}{(\text{P}_{\text{sheets}}\ \text{R}_{\text{sheets}})} \quad (2)$$

- **Cross-Drawing Compliance (CDC)** which tests multi-sheet, multi-discipline coordination reasoning, requiring a Yes/No determination and written justification. CDC responses are scored with a composite metric that prioritizes the binary compliance determination while rewarding justifications that cite specific values and sheet references:

$$\text{CDC}_{\text{comb}} = 0.7 \times \text{YN} - \text{Accuracy} + 0.3 \times \text{ROUGE} - \text{L} \quad (3)$$

  where ROUGE-L is computed as the F1 score of the longest common subsequence (LCS) between the predicted and ground-truth justification texts.

Models and Input Representation

Five MLLMs are evaluated: GPT-5.4 (OpenAI, frontier proprietary), Gemini 2.5 Flash (Google, frontier proprietary), GPT-4.1 Nano (OpenAI, lightweight proprietary), Gemma3:12b (Ollama, open-weight, 12B parameters), and Qwen3-VL:8B (Ollama, open-weight, 8B parameters). All models are evaluated in a zero-shot setting under deterministic conditions (temperature = 0, top-p = 1.0, seed = 42) with no few-shot examples, chain-of-thought prompting, or retrieval augmentation. Each model is provided with the question and the corresponding sheets required to answer the question. Each sheet is rasterized at 300 DPI, yielding one PNG image per sheet at approximately. This is sufficient to resolve annotation text, dimension strings, and callout symbols for MLLM review. The sheets provided are determined using three conditions. First is the distractor condition, where ground-truth sheets are augmented with plausible but irrelevant sheets selected by a four-phase algorithm that prioritizes sheets from the same discipline and hundred-series sub-group. The number of distractors is inversely proportional to the ground-truth sheet count, holding total input size to approximately 6–9 sheets per question. The second condition is the guaranteed-context condition. Here, only the exact sheets required to answer each question are provided, isolating reasoning capability from the retrieval challenge. Finally, the native-PDF condition provides the original PDF documents directly to the model without conversion to a standardized image format, leaving rendering and interpretation of document structure entirely to the model. This experiment is conducted using the web-based interfaces of GPT-5.4 and Gemini 2.5 Flash, where native PDF upload and parsing are supported. Due to context size and performance constraints, the native-PDF condition is evaluated only on these two frontier models as smaller models are unable to receive large documents.

Table 1. Results under distractor and guaranteed-context conditions. Bold values indicate best performance per metric per condition.

Model	Distractor Condition					Guaranteed Context				
	IE $F1_{Bow}$	SR $F1_{sheets}$	CDC Acc	CDC R-L	CDC Comb	IE F1-BoW	SR F1	CDC Acc	CDC R-L	CDC Comb
GPT-5.4	**0.563**	**0.868**	**0.756**	0.111	**0.563**	0.595	0.902	**0.610**	0.112	**0.461**
Gemini 2.5	0.553	0.801	0.293	**0.130**	0.244	**0.610**	**0.910**	0.195	0.115	0.171
GPT-4.1 Nano	0.175	0.390	0.444	0.131	0.350	0.118	0.854	0.415	**0.142**	0.333
Gemma3:12b	0.232	0.263	0.122	0.096	0.114	0.149	0.182	0.146	0.094	0.130
Qwen3-VL:8B	0.077	0.000	0.171	0.020	0.126	0.090	0.000	0.146	0.018	0.108

RESULTS AND DISCUSSION

Table 1 presents the full results under both evaluation conditions. IE results show that GPT-5.4 and Gemini 2.5 Flash achieve comparable performance, with F1-BoW scores of 0.563 and 0.553 respectively under the distractor condition, rising marginally to 0.595 and 0.610 under guaranteed context. The modest improvement when distractors are removed suggests that the primary IE challenge for frontier models is precision in text localization and extraction rather than sheet identification—the dense, cluttered layout of industrial drawing sheets impedes exact value retrieval even when the correct sheet is provided. Smaller models cluster in the range of 0.077–0.232 (distractor) and 0.090–0.149 (guaranteed context), indicating that they often return reformatted or partially correct values rather than exact text. Importantly, no model exceeds F1-BoW of 0.63, highlighting that accurate text extraction from construction drawings remains a substantial open challenge.

SR results reveal the most dramatic performance gap between model tiers. Under the guaranteed-context condition, GPT-5.4 and Gemini 2.5 Flash achieve F1-Sheets of 0.902 and 0.910, respectively, demonstrating strong ability to resolve sheet references when distractor pressure is removed. Under the distractor condition, however, Gemini 2.5 Flash declines to 0.801 while GPT-5.4 holds at 0.868, indicating that Gemini 2.5 Flash is more sensitive to distractor-induced ambiguity than GPT-5.4. Most strikingly, Qwen3-VL:8B completely fails to interpret the alphanumeric sheet numbering conventions used in this drawing set. GPT-4.1 Nano achieves F1 = 0.854 under guaranteed context but drops to 0.390 with distractors, confirming that lightweight proprietary models possess navigation capability that is severely degraded by context overload.

Unlike IE and SR, which primarily test extraction and navigation capabilities, CDC task type aims to evaluate Cross-discipline coordination. The results reveal complex interaction effects. GPT-5.4 achieves the highest YN Accuracy under the distractor condition (0.756) and the highest combined CDC score in both conditions (0.563 distractor; 0.461 guaranteed context). Gemini 2.5 Flash shows a pronounced anomaly: its YN Accuracy was notably poor and collapses from 0.293 (distractor) to 0.195 (guaranteed context), indicating unstable decision behavior that does not improve with better context, suggesting that improved retrieval does not directly translate to

stronger cross-drawing reasoning. This behavioural pattern may reflect over-reliance on surface-level agreement when context is sparse. ROUGE-L scores are uniformly low across all models (0.018–0.142). These results suggest that improving retrieval alone is insufficient, as accurate cross-drawing reasoning remains a separate and unresolved challenge for current MLLMs.

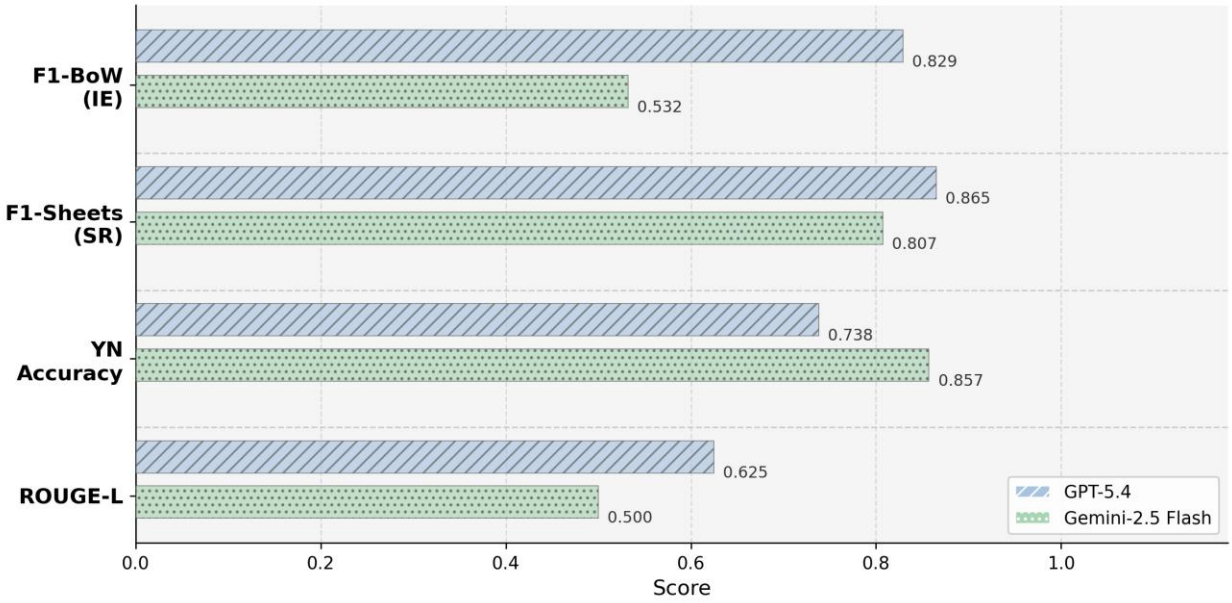


Figure 2. Native-PDF evaluation results comparing GPT-5.4 and Gemini 2.5 Flash across four metrics: Information Extraction (F1-BoW), Sheet Reference Retrieval (F1-Sheets), Cross-Drawing Compliance (YN Accuracy), and justification quality (ROUGE-L). Results highlight performance differences when models directly process original PDF documents without standardized rasterization.

Figure 2 shows GPT-5.4 achieves the highest Information Extraction performance (F1-BoW = 0.829) and strong Sheet Reference Retrieval (F1-Sheets = 0.865), indicating effective parsing and navigation of full-document inputs. It also demonstrates improved justification quality (ROUGE-L = 0.625), suggesting improved justification quality under native-PDF inputs. Gemini 2.5 Flash achieves the highest Yes/No Accuracy (0.857), outperforming GPT-5.4 on binary compliance decisions. However, its lower IE (0.532) and SR (0.807) scores indicate weaker precision in extraction and navigation despite benefiting from full-document context. Despite the improved results, performance across all metrics remains below levels required for reliable deployment in high-risk construction workflows, where errors can lead to costly rework or safety issues. These results reinforce that, despite promising capabilities, current MLLMs are not yet reliable for autonomous use in construction document review.

Implications for AI-Assisted Plan Review

The results highlight both the promise and limitations of MLLMs for construction document interpretation. Frontier models demonstrate strong capabilities in structured navigation tasks, particularly Sheet Reference Retrieval (SR), where performance approaches near-perfect levels under controlled conditions. This suggests that MLLMs can effectively assist practitioners in locating relevant information within large drawing sets, reducing the time required for manual

navigation. However, performance in Information Extraction (IE) and Cross-Drawing Compliance (CDC) reveals critical limitations. Even under ideal conditions, models struggle with precise value extraction from dense, cluttered drawings, and exhibit inconsistencies in multi-sheet reasoning. The degradation observed under distractor conditions further indicates that retrieval remains a primary bottleneck, directly impacting downstream reasoning performance.

From a deployment perspective, these findings suggest that MLLMs are best suited as decision-support tools rather than autonomous systems. Practical applications may include assisting with sheet navigation, surfacing relevant details, and providing preliminary coordination checks. However, due to the high sensitivity of construction workflows, human oversight remains essential. Errors in interpretation can propagate into costly rework or safety risks, limiting the feasibility of fully automated plan review at the current stage. Effective deployment will likely require hybrid systems that combine MLLMs with structured document indexing, retrieval pipelines, and domain-specific validation mechanisms.

CONCLUSION AND FUTURE WORK

This paper introduced ConstructionQA, a benchmark designed to evaluate MLLM comprehension of construction drawing sets across three hierarchical tasks: Information Extraction (IE), Sheet Reference Retrieval (SR), and Cross-Drawing Compliance (CDC). Results demonstrate that while frontier models significantly outperform smaller models, performance remains far from the level required for reliable use in construction workflows. A key finding is that retrieval is a primary bottleneck, with substantial performance gains observed under guaranteed-context conditions. Additionally, even the strongest models struggle with precise text extraction and consistent cross-drawing reasoning. Evaluation under full-PDF inputs further shows that while performance can improve with continuous document context, it remains insufficient for construction-based applications. Future work should focus on improving both benchmark design and model capabilities. On the benchmark side, expanding to additional drawing sets and incorporating richer structural annotations will improve generalizability and diagnostic power. On the modelling side, practitioners should consider integrating retrieval-augmented pipelines, improving visual-text alignment for small annotations, and incorporating graph-based representations of drawing relationships. Finally, advancing and standardizing semantic evaluation metrics beyond lexical overlap will enable more robust assessment of reasoning quality particularly in construction.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2222881. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., & Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems (NeurIPS).

American Institute of Architects. (2017). AIA Document A201–2017: General conditions of the contract for construction. AIA Contract Documents.

Doris, A. C., Grandi, D., Tomich, R., Alam, M. F., Ataei, M., Cheong, H., & Ahmed, F. (2024). DesignQA: A multimodal benchmark for evaluating large language models' understandingof engineering documentation (arXiv:2404.07917v2). arXiv. https://arxiv.org/abs/2404.07917

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., & He, R. (2023). MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394.

He, C., Yu, B., Liu, M., Guo, L., Tian, L., & Huang, J. (2024). Utilizing large language models to illustrate constraints for construction planning. Buildings, 14(8), 2511. https://doi.org/10.3390/buildings14082511

Kondratenko, A., Birhane, M., Hsain, H. E., & Maciocci, G. (2026). AECV-Bench: Benchmarking multimodal models on architectural and engineering drawings understanding. arXiv preprint arXiv:2601.04819.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., & Lin, D. (2023). MMBench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281.

Pan, Y., & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. Automation in Construction, 122, 103517.

PlanGrid and FMI Corporation. (2018). Construction disconnected: The high cost of poor data and communication. Industry Report. PlanGrid.

Sulbaran, T. (2024). Comparing two artificial intelligence language modeling approaches to evaluate construction schedule understanding. Proceedings of the 60th Annual Associated Schools of Construction International Conference.

Wu, Y., Wang, L., Liu, R. (2025). CEQuest: Benchmarking large language models for construction estimation. arXiv preprint arXiv:2508.16081.

Zhang, J., & El-Gohary, N. M. (2017). Integrating semantic NLP and logic reasoning into a unified system for fully automated code checking in construction. Automation in Construction, 73, 45–57.