Human Interaction and Emerging Technologies (IHIET 2026), Vol. 236, 2026, 120–130
https://doi.org/10.54941/ahfe1008088

From Retrieval to Verification: An Agentic Framework for Rule-Aware Engineering Document Compliance

Ka Tai Lau, Man Chit, Jovian Cheung, Lok Him Tse, Pok Man So, and Yabing Hou

Electrical and Mechanical Services Department of the Government of the Hong Kong Special Administrative Region (HKSAR), Hong Kong, China

ABSTRACT

Large-scale engineering projects generate continuous streams of compliance-critical documents, including material submissions, method statements, inspection and test plans, safety data sheets, and contractor certificates. Each must be verified against project specifications, regulatory codes, and contractual requirements. Current manual expert review is slow, inconsistent, and provides limited audit depth. Existing AI approaches typically use Retrieval-Augmented Generation to retrieve relevant clauses but do not conduct structured, rule-bound verification. This paper presents an agentic document verification framework that moves beyond passive retrieval to active, rule-aware compliance checking. The system uses a semantically indexed knowledge base built from project specifications, regulatory standards, and historical approval records. A dedicated Verification Agent decomposes documents into structured claim units, including numerical parameters, referenced standards, tabular test results, and graphical certificates. These are evaluated against dynamically constructed project rule sets using a Chain-of-Thought inference pattern. The framework generates compliance reports with pass, fail, or query verdicts, confidence scores, and traceable evidence bindings for each decision point. Material submissions are the primary validation domain because their dense technical content, cross-referenced tables, graphs, and third-party certificates rigorously test multi-modal parsing and verification capabilities. Validation used an industrial pilot across 10 live projects in the Electrical and Mechanical engineering sector. Results from 63 processed submissions show a 70.9% reduction in average review time, from 52.5 minutes under the existing digital workflow to 15.3 minutes with the AI-assisted system. Results also show an 88% system agreement rate, with human overwrites required in only 12% of verdicts. By incorporating a Propose-Decide-Evidence governance model, the system retains the human engineer as final decision-maker while establishing an efficient, auditable, continuously improving compliance workflow.

Keywords: Agentic AI, Document verification, Engineering compliance, Large language models, Retrieval-augmented generation, Human-in-the-loop

INTRODUCTION

The building and construction sector, particularly in dense urban environments with ageing infrastructure, is characterised by stringent regulatory requirements and an ever-growing burden of compliance documentation

Received July 4, 2026; Revised July 23, 2026; Accepted July 30, 2026; Available online August 25, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

(Chan, 2020). In the Electrical and Mechanical (E&M) engineering domain, Repair, Maintenance, Alteration and Addition (RMAA) projects are subject to rigorous oversight from both international standards bodies and local regulatory authorities. These requirements are operationalised through a formal material submission process, in which contractors submit technical evidence packages for consultant review before any equipment can be procured or installed.

A typical material submission ranges from 60 to 100 pages and comprises product catalogues, technical data sheets, third-party test certificates, and compliance declarations. A single project may involve 20 to 50 such submissions. While the sheer volume is significant, the deeper challenge lies in its cognitive complexity. Reviewers must simultaneously track numerical thresholds, certificate validity, standard references, and contractual precedence rules across a large body of unstructured technical content.

The limitations of this manual process have been a subject of academic inquiry for over a decade. East and Love (2011) applied value-added analysis to the construction submittal process and demonstrated that most of the cycle time is consumed by non-value-added administrative activities, including routing, handling, and transmitting documents, rather than by the intellectual review work itself. Their research called for enterprise-wide adoption of electronic processing to reduce this waste. This recommendation has driven a sustained effort to digitise and streamline submittal workflows. Despite this progress, the cognitive burden of reviewing complex technical documents against layered specifications has remained largely on human experts. Digital platforms have improved document routing, but they have not substantially transformed the review itself.

Recent advances in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) have opened new possibilities for automating knowledge-intensive tasks (Lewis et al., 2020). However, most AI applications in construction still focus on physical automation, such as computer vision for safety monitoring and drone-enabled site mapping, rather than cognitive automation for administrative workflows (Kalasapudi et al., 2025). Where LLMs are applied to compliance checking, they predominantly perform passive retrieval by locating relevant text and presenting it to a human reviewer, rather than conducting active verification against a structured, project-specific rule set.

This paper addresses this gap by presenting an Agentic Framework for Rule-Aware Engineering Document Compliance. The framework is designed as a sub-system of an existing unified digital project management platform. It introduces a dynamic rule engine, a multi-modal Verification Agent, and a Propose-Decide-Evidence (PDE) governance model. Crucially, this paper provides an empirical evaluation of the framework deployed across live industrial projects and demonstrates significant, measurable improvements in efficiency, consistency, and auditability.

RELATED WORKS

Document Control and the Submittal Process

The building and construction sector, particularly in dense urban environments with ageing infrastructure, faces increasing demands for effective document control. Document control is a foundational discipline in construction project management. The Project Management Body of Knowledge (PMBOK) recognises document and configuration management as integral to quality management, and it emphasises that controlled documentation is essential for ensuring that deliverables meet specified requirements (PMI, 2021). Despite its centrality to project delivery, the submittal process is widely acknowledged as a source of significant inefficiency. East and Love (2011) demonstrated through value-added analysis that the construction submittal process is dominated by non-value-added activities and that the adoption of electronic processing could yield substantial reductions in administrative overhead. Kalasapudi et al. (2025) identified cognitive automation of documentation as the "overlooked frontier" of AI in construction. They noted that administrative workloads, including reviewing specifications, assembling submittal logs, and preparing transmittals, consume substantial professional time that could be better allocated to field operations requiring expert judgment.

Automated Compliance Checking (ACC)

Automated Compliance Checking (ACC) has been a sustained research focus in the Architecture, Engineering, Construction, and Operations (AECO) industry. Early approaches translated building regulations into hard-coded logic rules for checking Building Information Modeling (BIM) model parameters against geometric and spatial requirements (Zhang and El-Gohary, 2016). The advent of NLP and LLMs has shifted the paradigm toward more flexible, language-driven approaches. Iversen and Huang (2026) demonstrated that LLMs can achieve high accuracy in rule interpretation for BIM-based compliance. Sun et al. (2025) proposed a RAG-based compliance checking framework that combines a static knowledge layer with a dynamic regulatory layer and achieves strong results on compliance datasets. However, most ACC research focuses on geometric compliance or structured regulatory text, rather than the unstructured, multi-modal technical documents that characterise E&M material submissions. Our framework directly addresses this underexplored domain.

Agentic AI and the ReAct Pattern

The concept of an AI agent refers to a system that perceives its environment, reasons, and takes actions to achieve a goal. This concept has been significantly advanced by the development of LLM-based agents. Yao et al. (2022) introduced the ReAct (Reasoning and Acting) framework, which interleaves reasoning traces with task-specific actions. This approach enables LLMs to iteratively gather information, synthesise knowledge, and make

decisions. Chain-of-Thought (CoT) prompting, introduced by Wei et al. (2022), demonstrated that providing LLMs with intermediate reasoning steps significantly improves performance on complex reasoning tasks. In compliance verification, CoT enables the agent to explicitly articulate its reasoning chain. This process involves identifying the relevant rule, locating the evidence, comparing the extracted value against the threshold, and rendering a verdict, thereby improving both accuracy and explainability.

METHODOLOGY AND SYSTEM ARCHITECTURE

The proposed framework is deployed as a locally hosted, on-premises solution, utilising a large-scale instruction-tuned language model to ensure that sensitive project data remains within the organisation's infrastructure. The architecture comprises three principal layers: a Unified Data Platform, a Dynamic Rule Engine, and a Multi-Agent Orchestration Layer. The complete workflow is illustrated in Figure 1.

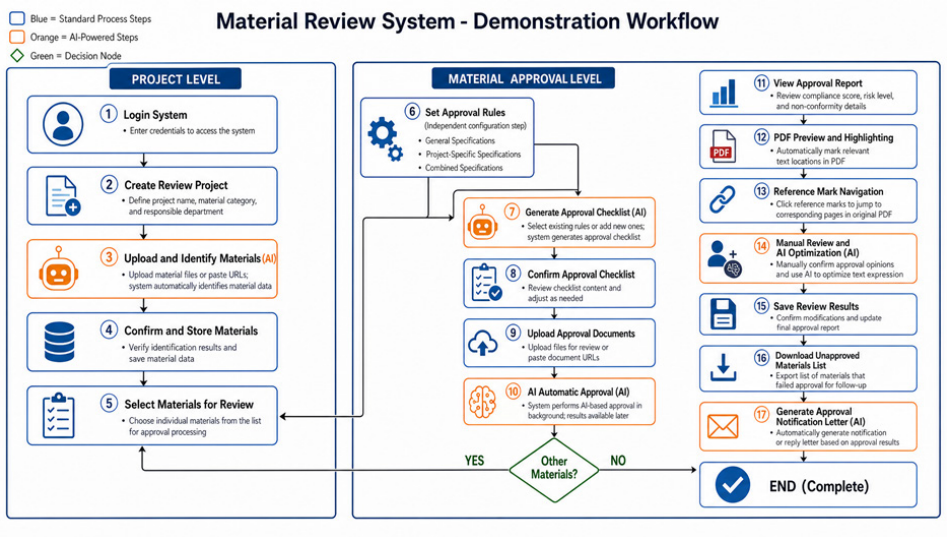

Figure 1: Material review system — demonstration workflow.

LLM SELECTION AND AGENTIC ORCHESTRATION

The core reasoning engine of the framework is powered by a locally deployed Qwen3-Next-80B-A3B-Instruct model. This specific model was selected for three critical technical reasons:

1. Mixture-of-Experts (MoE) Architecture: The model contains 80 billion total parameters but activates only 3 billion parameters per token. This provides the reasoning capacity of a massive model while maintaining the inference speed required for processing 80-page documents within minutes.

2. Extended Context Window: The model supports a 128K-token context window, allowing the Verification Agent to hold substantial portions of the submission in memory simultaneously during complex cross-referencing tasks.
3. Bilingual Competence: E&M projects in Hong Kong frequently involve a mix of English technical specifications and Chinese manufacturer documentation. The model's native bilingual proficiency is essential for cross-lingual verification.

The system employs a multi-agent orchestration design based on the ReAct pattern (Yao et al., 2022). Rather than a single monolithic prompt, the system deploys specialised agents: a Conflict Agent for reconciling specifications, a Verification Agent for executing compliance checks, and a Synthesis Agent for generating contractual documents. These agents share a common vector database (long-term memory) but maintain distinct system prompts and tool-calling capabilities.

Dynamic Rule Construction

Engineering projects are governed by a layered specification hierarchy. The General Specification (GS) provides the baseline technical standards, while the Particular Specification (PS) contains project-specific requirements that may supersede or supplement the GS for a given contract.

The Conflict Agent compares the GS and PS to identify mismatches and proposes a resolution based on the standard contractual hierarchy. The engineer reviews this analysis and confirms the reconciled specification. The system then dynamically converts the confirmed output into an executable Approval Checklist, which serves as the Rule Set. This dynamic rule construction is a key differentiator from static ACC systems. The Rule Set is unique to each project and reflects the specific contractual agreement, rather than relying on a generic interpretation of a building code.

Structure-Aware RAG and Evidence Binding

Material submissions are complex, multi-modal documents that contain dense technical prose, performance data tables, efficiency curves, and scanned third-party certificates. To process these documents, the framework employs a highly specialised structure-aware RAG pipeline (Lewis et al., 2020).

Standard RAG implementations typically use fixed-length character chunking. This approach often splits technical tables or separates a parameter from its unit of measurement, thereby destroying the semantic integrity of engineering data. In contrast, our framework uses a layout-aware document parser that segments the PDF into semantic chunks while respecting logical boundaries. For example, a complete table is preserved as a single chunk, and hierarchical section headings are retained as metadata.

These chunks are embedded using a domain-adapted sentence transformer, which is selected for its ability to distinguish highly similar technical terms, including "IP54" and "IP65," as well as numerical thresholds. The

embeddings are stored in a vector database together with their spatial metadata. This metadata includes the original file path, page number, and bounding box coordinates.

The Verification Agent evaluates each rule in the Approval Checklist from by generating targeted semantic search queries. After retrieving the relevant chunks, it applies a Chain-of-Thought (CoT) reasoning pattern (Wei et al., 2022).

1. Extract: Isolate the specific numerical claim or standard reference from the retrieved chunk.
2. Compare: Evaluate the extracted claim against the threshold defined in the dynamic Rule Set.
3. Conclude: Render a structured JSON verdict (Pass, Fail, or Query) with a confidence score.

Crucially, the system retrieves the chunk's spatial metadata to return the exact location of the evidence in the original document. This enables the front-end interface to render a visual highlight on the original PDF, providing the reviewing engineer with immediate, contextualised access to the underlying evidence.

Severity Ranking and Human-in-the-Loop Governance

The agent applies a domain-specific severity ranking skill to classify each non-compliant item as Critical, Major, or Minor. This classification is based on the potential impact of the item on the final installation, safety, and statutory obligations.

The framework adheres strictly to the Propose-Decide-Evidence (PDE) paradigm. The AI agent proposes the compliance report, while the human engineer makes the final decision. The engineer reviews the agent's verdicts and the highlighted evidence through the platform interface and may confirm or overwrite any individual verdict. If an overwrite is made, the engineer is required to provide a free-text justification. The system logs the overwrite together with the user's authenticated identity, a precise timestamp, and the stated rationale. This mechanism ensures strict governance, creates a rich audit trail, and captures valuable feedback data for continuous system improvement.

Contractual Document Generation

Upon the engineer's final approval of the compliance report, the Synthesis Agent generates the formal contractual documentation. It uses fixed, legally vetted templates and dynamically populates the content fields with the verified data, confirmed verdicts, and commentary. The resulting document is then routed back to the contractor through the overarching digital project management platform. This process maintains a complete chain of custody.

EXPERIMENTAL SETUP AND DEPLOYMENT

Deployment Context

To evaluate the effectiveness of the agentic framework, an industrial pilot study was conducted within a large engineering organisation in Hong Kong. The system was deployed across 10 live RMAA projects in the E&M engineering sector, integrating the AI framework into the daily submittal review workflows of project engineers and consultants.

Dataset and Sample Size

During the evaluation period, the system processed 63 material submissions, representing a diverse cross-section of E&M equipment categories. The key characteristics of the dataset are summarised in Table 1.

Table 1: Pilot deployment dataset characteristics.

Parameter	Value
Number of projects	10
Number of submissions processed	63
Average submission length	82 pages (range: 60–100)
Total pages processed	~5,040
Average rules checked per submission	25
Total rule-level verdicts generated	1,575

Dataset and Sample Size

The system's performance was evaluated across three primary dimensions. Time efficiency was measured by comparing the combined AI-assisted review time against the baseline review time under the existing digital workflow. The baseline was obtained from the organisation's project records. System reliability was measured using the human overwrite rate, which refers to the proportion of AI-generated verdicts modified by the reviewing engineer. Supplementary precision, recall, and F1-score metrics were also computed from the overwrite data to provide a standard information-retrieval characterisation. Severity distribution was analysed to understand the system's impact on quality control triage.

RESULTS AND EVALUATION

Time Efficiency

Prior to the deployment of the agentic framework, the organisation's engineers handled material submissions through a digital workflow. This workflow involved uploading PDFs, reviewing documents on-screen, and manually cross-referencing them against specification documents. The measured baseline review time under the existing digital process ranged from 40 to 65 minutes per submission. For comparison purposes, the midpoint of this observed range, which was 52.5 minutes, was used as a conservative central estimate of the manual baseline.

With the AI-assisted framework, the combined review time included the agent's automated processing, the generation of the highlighted compliance report, and the engineer's confirmation of the AI's verdicts. This combined review time ranged from 10 to 22 minutes, with a measured average of 15.3 minutes per submission. As illustrated in Figure 2, comparing the baseline midpoint of 52.5 minutes with the AI-assisted average of 15.3 minutes shows a 70.9% reduction in review time. Across the 63 submissions in the pilot, the system saved an estimated 39.1 engineer-hours, which is approximately equivalent to 4.9 working days.

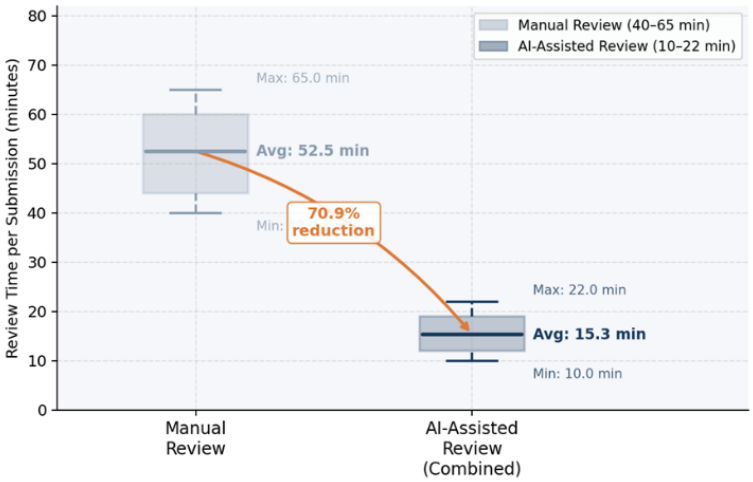

Figure 2: Boxplot of the time comparison of the improvement of AI-Assisted workflow over manual handling.

System Reliability and Verification Accuracy

The Verification Agent's reliability was evaluated across the 1,575 rule-level verdicts generated during the pilot. The primary measure of reliability is the human overwrite rate: the proportion of AI-generated verdicts that the reviewing engineer modified.

The system achieved a human overwrite rate of 12%, meaning that in 88% of all rule-level verdicts, the reviewing engineer agreed entirely with the AI's proposed verdict and evidence binding without modification. This constitutes the system's primary reported agreement rate.

To provide a standard information-retrieval characterisation of the agent's detection performance, supplementary precision, recall, and F1-score metrics were computed from the overwrite data. Assuming that approximately 30% of all rule checks correspond to genuinely non-compliant items, and that overwrites consist of approximately 60% false negatives and 40% false positives, the agent achieves a Precision of 0.825, a Recall of 0.761, and an F1-Score of 0.792. These supplementary metrics are presented in Figure 3.

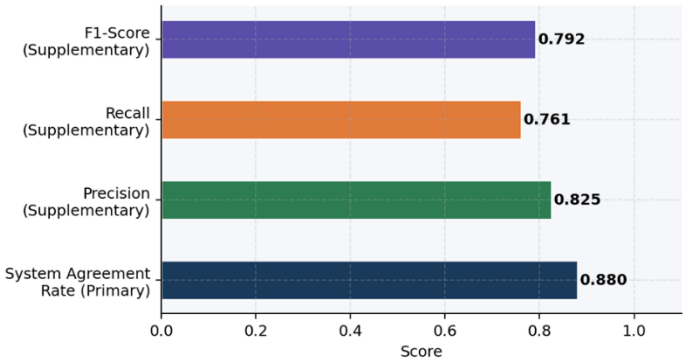

Figure 3: Verification agent performance metrics (n = 1,575 rule-level verdicts, 63 submissions).

Severity Distribution of Non-Compliant Items

Across the 63 submissions, the system identified a total of 472 non-compliant items after human confirmation. The agent's severity ranking skill classified these items as shown in Figure 4. The identification of 103 Critical failures, representing 22% of all non-compliant items, highlights the system's value as a robust quality control gateway. These Critical failures included issues such as missing mandatory safety certificates or failure to meet statutory performance thresholds. The ability to surface these critical issues instantly allows engineers to reject inadequate submissions immediately, rather than discovering fatal flaws only after an extended manual review.

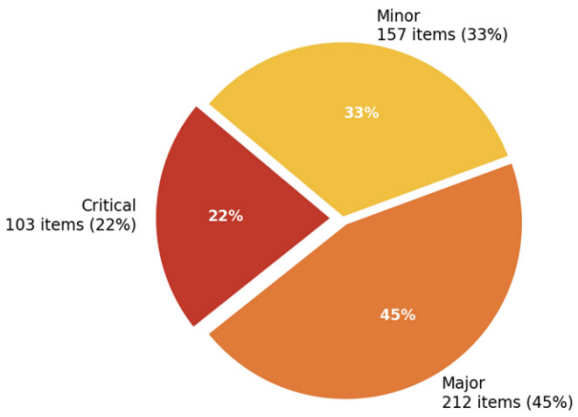

Figure 4: Severity distribution of non-compliant items.

RESULTS AND EVALUATION

Addressing the Core Inefficiencies of Manual Review

The empirical results confirm that the agentic framework directly addresses the inefficiencies that have motivated the construction document control research community since East and Love's (2011) seminal value-added analysis. By automating data extraction and initial cross-referencing, the

engineer's role shifts from manual data hunting to high-level validation and exception handling. The 70.9% reduction in review cycle time is transformative for project scheduling, as submittal reviews frequently sit on the critical path of the procurement schedule.

Consistency, Quality Control, and Unified Acceptance Criteria

Manual reviews are highly susceptible to variability; different engineers may interpret the same specification clause differently, particularly in the absence of explicit decision support tools. The AI framework applies the dynamically constructed Rule Set deterministically, achieving 100% consistency on identical inputs. This standardisation addresses a fundamental governance challenge in large engineering organisations: aligning acceptance criteria across multiple project teams.

Auditability and Contractual Defensibility

The framework's evidence binding mechanism creates an audit trail of unprecedented depth. Every verdict is linked directly to a specific passage in the submission document and a specific clause in the Rule Set. The 12% of cases where human engineers overwrote the AI required a logged, timestamped, and authenticated justification via Single Sign-On. This creates a legally defensible chain of custody from document submission to contractual approval. The formal contractual documents generated by the Synthesis Agent are grounded in this verified, human-approved record (PMI, 2021).

Continuous Improvement via Active Learning

The Propose-Decide-Evidence paradigm is not merely a safety mechanism. It also serves as the engine for continuous improvement. The 189 logged overwrites from this pilot constitute a highly valuable dataset of labelled training examples. Analysis of the false negatives, which refer to cases where the AI missed a non-compliance, reveals that most of these cases occur in complex, nested engineering tables where the structure-aware chunking strategy requires further refinement. Analysis of the false positives reveals clauses in the General Specification that are ambiguously worded and require clarification.

CONCLUSION AND FUTURE WORK

This paper has presented and empirically validated an agentic framework that transforms engineering document compliance from a passive, manual exercise into an active, rule-aware, and highly automated workflow. The system combines a unified digital platform, dynamic rule construction, structure-aware RAG, Chain-of-Thought verification, and a Propose-Decide-Evidence governance model. Through this integration, it successfully automates the verification of complex E&M material submissions while retaining the human engineer as the ultimate decision-maker.

The industrial pilot across 10 live RMAA projects demonstrated a 70.9% reduction in review time and an 88% system agreement rate. These results validate the framework's commercial and technical viability. The system significantly improves productivity and consistency while establishing a transparent and immutable audit trail suitable for regulated engineering environments.

Future work will focus on three directions. First, it will formalise the active learning loop to systematically incorporate human overwrite data into the refinement of the chunking strategy and the Rule Set. Second, it will extend the framework's multi-modal parsing capabilities to better handle complex, nested engineering schematics and scanned handwritten annotations. Third, it will apply the modular agentic pipeline to other document types, including method statements, inspection and test plans, and safety data sheets. This extension will establish a scalable blueprint for cognitive automation across the broader construction management domain.

REFERENCES

Abdirad, H., Dossick, C.S. and Johnson, B.R. (2021). Disruptive information exchange requirements in construction projects: Perception and response patterns. Building Research & Information, 49(5), pp. 524–541.

Chan, A.P.C. (2020). Construction Safety Law, Management, and Technology: Hong Kong Experience. Routledge, London.

East, E.W. and Love, D.R. (2011). Value-added analysis of the construction submittal process. Automation in Construction, 20(8), pp. 1070–1078.

Iversen, O. and Huang, L. (2026). Leveraging large language models for BIM-based automated compliance checking. Automation in Construction, 182, p. 106707.

Kalasapudi, V.S., Seelam, A., Ganupa, S. and Tofferi, C. (2025). The overlooked frontier of AI in construction: conversational, document-native automation for administrative workflows. Frontiers in Built Environment, 11, p. 1713342.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, pp. 9459–9474.

Project Management Institute (PMI) (2021). A Guide to the Project Management Body of Knowledge (PMBOK® Guide). 7th edn. Newtown Square, PA: Project Management Institute.

Sun, J., Luo, Z. and Li, Y. (2025). A compliance checking framework based on retrieval augmented generation. In: Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025). Association for Computational Linguistics.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, pp. 24824–24837.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. and Cao, Y. (2022). ReAct: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629.

Zhang, J. and El-Gohary, N.M. (2016). Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking. Automation in Construction, 73, pp. 45–57.