Journal of Building Design and Environment

Science Exploration

Open Access Research Article

A multi-agent debate workflow for construction projects: A cross-stage decision framework

Hao Yin*

Department of Computing, Faculty of Computer and Mathematical Sciences, The Hong Kong Polytechnic University, Hong Kong SAR, China.

***Correspondence to:** Hao Yin, Department of Computing, Faculty of Computer and Mathematical Sciences, The Hong Kong Polytechnic University, Hong Kong SAR, China. E-mail: hi-bruce.yin@connect.polyu.hk

Received: October 24, 2025 **Accepted:** November 07, 2025 **Published:** November 14, 2025

Cite this article: Yin H. A multi-agent debate workflow for construction projects: A cross-stage decision framework. J Build Des Environ. 2025;3:202562. https://doi.org/10.70401/jbde.2025.0018

Abstract

The growing complexity of construction projects demands decision processes that can integrate diverse professional perspectives across design, cost control, construction, and acceptance. Traditional sequential management approaches often leave conflicts unresolved until late stages, causing delays, cost overruns, and rework. To address this issue, this study develops a cross-stage workflow for construction projects that embeds artificial intelligence into collaborative decision-making. The workflow employs digital agents representing different roles, which engage in structured debates supported by evidence from building information modeling models, engineering drawings, cost records, and technical standards. Through iterative exchanges, alternative solutions are proposed, challenged, and refined, with the process producing decisions that are transparent, traceable, and adaptable to changing project conditions. Results from four representative scenarios confirm significant improvements over conventional practice: unresolved design clashes were reduced by more than 40%, cost forecast deviations narrowed from around 12% to below 5%, schedule variance under disruption decreased by about 18%, and first-pass acceptance rates increased from 74% to 88% with fewer reworks. A longer-term case study further demonstrated reductions in cost variance and project duration, together with higher stakeholder satisfaction. By coupling AI-enabled debate mechanisms with established digital construction environments, the workflow turns disciplinary conflicts into structured reasoning that enhances decision quality. The findings highlight how intelligent and collaborative approaches can deliver measurable gains in cost, time, and quality, while offering a practical pathway for advancing smart construction practices.

Keywords: Artificial intelligence in construction, multi-agent collaboration, smart construction management, cross-stage decision making, collaborative workflows

1. Introduction

Modern construction projects are increasingly complex, involving long durations, multidisciplinary coordination, and multiple stakeholders. Decisions in early stages, such as design and cost estimation, often propagate downstream, creating conflicts in construction and quality acceptance. Conventional project management methods, while structured, remain largely sequential and fragmented. This separation limits the ability to coordinate objectives across stages in real time, and recurring problems such as design-construction clashes, cost overruns, delays, and quality defects continue to affect project performance. Recent developments in Building Information Modeling (BIM)-based optimization and learning frameworks have improved coordination and scheduling performance in isolated project phases. For instance, optimization-based clash resolution approaches[1] and graph convolutional models for mechanical, electrical, and plumbing (MEP) change prediction[2] have demonstrated the potential of intelligent automation in design coordination and conflict analysis. However, these methods are primarily limited to static or single-disciplinary problem settings and depend on fixed optimization or prediction structures. In contrast, the present study establishes a cross-stage, debate-driven decision framework that integrates design, cost estimation, construction, and acceptance. Each debate process is anchored in verifiable evidence derived from industry foundation classes (IFC) models, bill of quantities (BoQ) data, and regulatory clauses, ensuring transparent reasoning and traceable decision paths. Furthermore, a trust-weighted voting mechanism dynamically calibrates agent influence based on historical reliability, complementing rather than replacing traditional optimization strategies. This formulation extends existing coordination and scheduling paradigms toward a more adaptive, auditable, and explainable decision process across the project lifecycle.

Recent advances in artificial intelligence provide opportunities to address these longstanding challenges[3]. Multi-agent systems (MAS), in particular, offer a way to model diverse roles, each with autonomous reasoning and adaptive learning capabilities. When

combined with debate-oriented coordination, these agents can simulate how professionals from different domains argue, challenge, and reconcile proposals before decisions are finalized. Rather than treating AI as an isolated optimization tool, the framework positions intelligent mechanisms as mediators that help construction engineers and managers evaluate trade-offs with evidence and reach more balanced solutions.

In this study, we propose a cross-stage decision workflow for construction projects that integrates debate among AI-driven agents representing design, cost, construction, and quality perspectives. Each agent uses project data (including BIM models, engineering drawings, cost records, and standards) to generate arguments and counterarguments. Through iterative rounds of structured debate, conflicts are explicitly surfaced, evidence is tested, and consensus is formed on decisions that are both technically sound and practically feasible. Unlike static rule-based systems, the agents continuously refine their strategies using feedback from project outcomes[4], making the workflow adaptive to changing conditions, as shown in Figure 1.

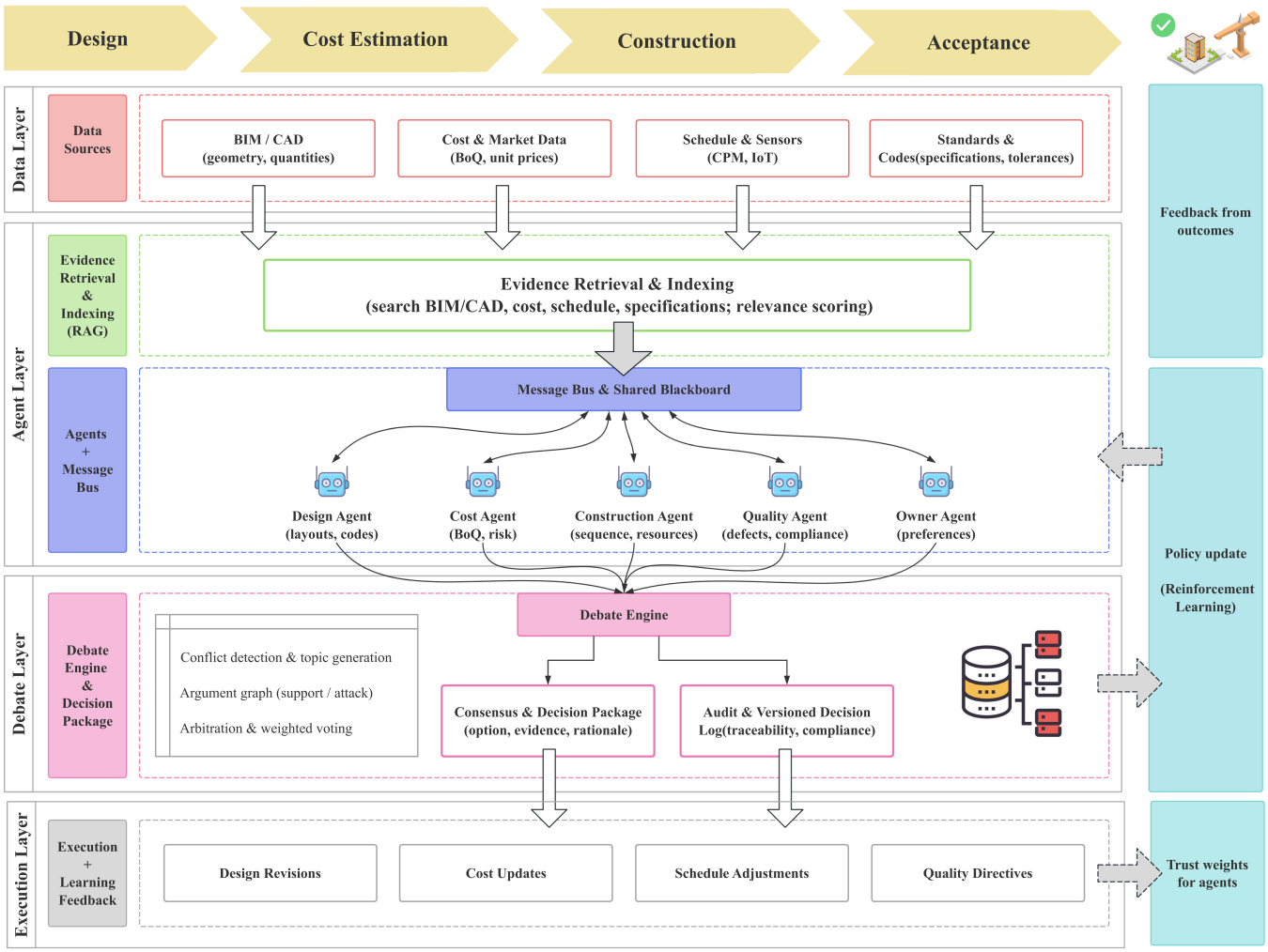

Figure 1. Conceptual framework of the multi-agent debate workflow. Evidence retrieval involves structured searches across BIM models, CAD drawings, cost ledgers, and standards (e.g., querying IFC attributes, Republished with permission from[5], CAD layers, BoQ entries, or tolerance clauses), rather than general text-based retrieval. BIM: building information modeling; IFC: industry foundation classes; CAD: computer-aided design; BoQ: bill of quantities; CPM: critical path method; IoT: Internet of Things; RAG: retrieval-augmented generation.

Initial applications in representative project scenarios show that this approach can improve core project outcomes: design changes are resolved more quickly, cost forecasts are more accurate, and acceptance processes yield higher first-pass success rates. These results highlight the potential of integrating AI debate mechanisms into construction workflows, not as a replacement for human expertise, but as a scalable tool that strengthens collaboration and decision quality.

The main contributions of this study are:

(1) A debate-oriented, AI-driven multi-agent workflow that connects design, cost, construction, and acceptance into a unified decision process.

(2) An evidence-based mechanism that grounds agent interactions in verifiable project data, enhancing transparency, traceability, and explainability.

(3) Experimental validation across multiple project stages, demonstrating measurable improvements in cost control, schedule adherence, and quality outcomes compared with conventional coordination methods.

2. Literature Review

2.1 MAS in construction engineering

The application of MAS in the architecture, engineering, and construction (AEC) industry has gained momentum as projects have grown more complex and collaborative. Early studies focused on agent-based coordination of design tasks and supply chain logistics, highlighting the ability of autonomous entities to represent different project roles and exchange information in real time. MAS have been applied to clash detection in design models, distributed resource scheduling, and monitoring of construction site safety[6]. These approaches demonstrate that multi-agent coordination can reduce human communication overhead and support decentralized decision-making. However, most existing systems are limited to single project phases, such as design optimization or site scheduling, without providing a cross-stage mechanism to ensure consistency across the lifecycle[7]. Moreover, the coordination logic often relies on predefined rules or heuristics, which limits adaptability in dynamic project environments. This gap motivates the integration of advanced reasoning and debate mechanisms into agent workflows, enabling decisions to be tested and refined across multiple stages.

2.2 Computational debate and decision-making models

Beyond the construction domain, research in artificial intelligence has explored debate and argumentation frameworks as mechanisms for structured decision-making. Computational models of argumentation provide ways to represent conflicting claims, supporting and attacking relations, and criteria for determining which arguments are accepted[8]. These models have been applied in domains such as legal reasoning, e-commerce negotiation, and healthcare diagnostics, where transparency and accountability are crucial[9]. In MAS, debate allows agents to challenge one another's proposals, cite evidence, and converge toward acceptable solutions[10,11]. For construction projects, this paradigm is particularly relevant because real-world decisions often involve competing priorities—cost versus quality, speed versus safety, innovation versus compliance. Yet, existing argumentation models are rarely adapted to the engineering context, where evidence must be drawn from highly structured sources such as BIM models, standards, and project records. The integration of argumentation into construction decision-making thus represents both a technical and disciplinary innovation.

2.3 Data-driven methods with BIM and digital twins

BIM has become a cornerstone of digital construction, providing a shared environment for design geometry, material quantities, scheduling, and facility management data. More recently, digital twins have extended this concept by linking BIM models with real-time sensor and operational data, creating living models of projects during construction and operation[12,13]. BIM- and twin-based approaches have been applied to clash detection, energy analysis, progress monitoring, and maintenance planning[2,14]. In the context of decision support, BIM provides a valuable evidence base, enabling agents to retrieve verifiable information to support arguments during debate[15]. However, most existing BIM applications focus on visualization and static information management rather than structured reasoning. There is limited research on using BIM and digital twin data to ground multi-agent discussions in verifiable evidence, despite the clear potential for improving transparency and traceability. Bridging this gap requires methods that integrate data retrieval with argumentation processes, ensuring that each decision is supported by explicit, traceable information.

2.4 Cost estimation, scheduling, and quality control studies

Research in intelligent methods for cost estimation, scheduling, and quality management has produced a wide range of tools. Machine learning models have been used to predict construction costs based on historical projects, improving accuracy compared to traditional regression methods. Reinforcement learning and optimization algorithms have been applied to construction scheduling, particularly in resource allocation and delay mitigation. For quality control, computer vision methods have enabled defect detection from images, and Internet of Things (IoT)-based monitoring systems have improved real-time compliance checking. These studies confirm the feasibility of AI-driven methods for individual project tasks. However, they are typically designed in isolation, without mechanisms to coordinate across domains. A cost estimation model may suggest material substitutions that reduce expenses, but without considering constructability or long-term quality implications. Similarly, scheduling optimizations may fail to capture downstream acceptance requirements. A cross-stage debate workflow that integrates these perspectives can prevent such misalignments by enabling agents to challenge and refine each other's proposals before implementation.

2.5 Summary of gaps

In summary, existing research demonstrates the potential of multi-agent coordination, argumentation models, BIM and digital twins, and intelligent methods for cost, schedule, and quality management. However, current approaches remain fragmented. MAS

applications in AEC are often restricted to single phases and rely heavily on static rule-based coordination. Argumentation models, while powerful in other domains, have not been adapted to construction-specific evidence sources. BIM and digital twins provide rich data environments, but their potential as evidence bases for structured reasoning remains underexplored. Studies in cost, scheduling, and quality show promising results but operate independently, without a unifying mechanism for cross-stage decision-making. This paper addresses these gaps by proposing an AI multi-agent workflow in which debate and evidence integration form the backbone of decision support, connecting design, cost estimation, construction, and acceptance into a coherent and adaptive process.

3. Methodology

3.1 Framework overview across project stages

The proposed workflow is designed to span the major phases of a construction project, including design, cost estimation, construction, and acceptance, where decisions are most interdependent. Instead of viewing these phases as sequential, the workflow organizes them into a network of interacting decision agents. Each agent represents a professional role, such as designer, cost engineer, construction manager, or quality inspector. The central principle is that these agents engage in structured debate whenever conflicts or uncertainties arise, with their exchanges grounded in evidence drawn from digital project data.

The framework consists of four layers. At the data layer, BIM models, computer-aided design (CAD) drawings, cost records, standards, and sensor data provide the evidence base. The agent layer comprises specialized agents, each with domain knowledge and learning capabilities. The debate layer orchestrates structured argumentation, including topic generation, evidence retrieval, argument evaluation, and arbitration. Finally, the execution layer translates consensus decisions into project actions, such as design revisions, cost updates, schedule adjustments, or quality directives. Together, these layers ensure that decisions are both transparent and traceable across the project lifecycle.

3.2 Roles of decision agents

Each agent encapsulates both domain expertise and adaptive reasoning, as shown in Table 1.

Table 1. Roles and information flows of agents in the debate-oriented framework.

Agent	Main Inputs	Core Responsibilities	Outputs to Debate Engine
Design Agent	BIM/CAD models; design codes	Generate layout alternatives; check code compliance; detect clashes and spatial constraints	Proposed design options; constraint/clash flags
Cost Agent	BoQ; market prices and cost indices	Estimate costs and sensitivities; assess budget risks and change impacts	Cost projections; risk alerts; cost-impact justifications
Construction Agent	Schedules; resource availability; site constraints	Monitors and optimizes activity sequences and resource allocation; identify bottlenecks and disruption risks (e.g., logistics, weather)	Feasibility assessments; rescheduling proposals; bottleneck flags
Quality Agent	Standards and specifications; tolerance rules; inspection data/images	Compliance checking; defect detection (rule- and vision-based); recommend corrective actions	Compliance reports; defect tickets with evidence
Owner Agent (optional)	Client requirements; priority settings (budget caps, performance goals, sustainability preferences)	Declare stakeholder preferences; set acceptance thresholds and trade-off weights	Preference weights; accept/reject signals; policy constraints
Arbiter Agent	Arguments and evidence from all agents; historical trust scores	Moderate and structure debates; maintain the argument graph; run arbitration with trust-weighted voting; update agent trust	Consensus decision record; rationale trace; updated trust weights

Each agent module operates within a shared data environment linked to BIM, CAD, and project databases. Outputs are standardized for parsing by the central debate engine, ensuring interoperability and traceability across decision stages. BIM: building information modeling; CAD: computer-aided design.

· Design Agent: evaluates architectural and engineering drawings, checks compliance with design codes, and generates alternative layouts.

· Cost Agent: analyzes BoQ and market prices, estimates cost implications of design changes, and forecasts budget risks.

· Construction Agent: monitors schedules, resources, and site constraints; identifies conflicts such as equipment bottlenecks or

weather delays.

· Quality Agent: inspects compliance with specifications, detects defects using rule-based and vision-based methods, and recommends corrective actions.

· Owner Agent (optional in some scenarios): reflects client priorities such as budget caps, performance goals, or sustainability preferences.

· Arbiter Agent: acts as moderator, maintaining the debate structure, recording arguments and evidence, and updating trust values based on agent performance.

This multi-agent arrangement reflects the professional diversity in real projects while allowing systematic coordination.

3.3 Debate process and evidence integration

At the core of the workflow is the debate process, which transforms potential conflicts into structured reasoning and resolution.

3.3.1 Conflict detection and topic generation

Debates are triggered when inconsistencies or risks are detected. Examples include clashes in BIM models, discrepancies between cost forecasts and budgets, or deviations between planned and actual progress. These events are translated into topics such as "layout feasibility", "cost overrun risk", or "acceptance of defect thresholds". Each topic then becomes the focus of an iterative debate cycle[16].

3.3.2 Evidence retrieval from drawings, standards, and records

For each topic, agents gather supporting evidence from project data[17]. BIM models supply geometric and quantity information; CAD drawings provide layout and detailing; cost databases offer unit rates; and standards specify compliance thresholds.

Evidence is organized into structured packages tagged with source, timestamp, and relevance. This ensures that arguments are not based on opaque reasoning but on verifiable project records. In our implementation, evidence retrieval is based on project data schemas: IFC property fields for BIM models, CAD layer and block attributes for drawings, BoQ line items for cost data, and clause identifiers for standards. Retrieved items are ranked by relevance scores and de-duplicated before being passed into the debate engine[18].

3.3.3 Argument structure and resolution

Agents present arguments in favor or against specific proposals, attaching evidence packages as justification[19]. Arguments are represented as nodes in an argument graph, with edges denoting support or attack relations. The system evaluates these arguments using criteria such as logical consistency, evidence credibility, and alignment with project objectives. Agents may revise their arguments in subsequent rounds, either reinforcing their position with new evidence or conceding when counterarguments prove stronger, as shown in Figure 2.

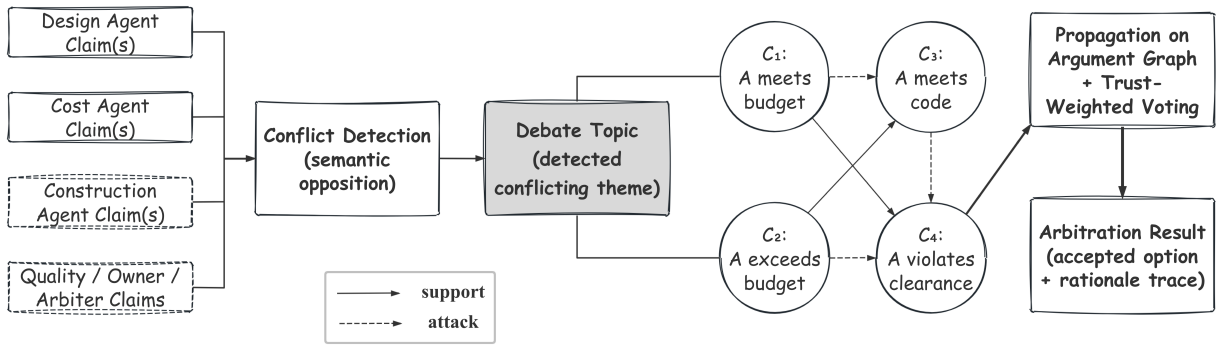

Figure 2. Argument structure and resolution process in the multi-agent debate. The process involves sequential stages of conflict detection, debate topic formulation, argument graph construction (support/attack relations), propagation with trust-weighted voting (see Formula (2)), and arbitration result generation.

3.3.4 Arbitration and consensus building

The Arbiter Agent moderates the process, ensuring that debates progress toward resolution rather than endless cycles. It maintains trust values for each agent based on past accuracy and reliability, which influence the weighting of their votes. Consensus is achieved when the majority of weighted votes converge on a decision, or when no new counterarguments emerge after a set number of rounds. The result is a decision package containing the chosen option, supporting evidence, and a rationale trace[20].

To formalize this process, a trust-weighted, multi-objective consensus scheme is adopted. Each agent evaluates a candidate option o across multiple project objectives $k \in$ {cost, schedule, quality} using normalized scores $s_{i,k}(o) \in [-1, 1]$. Stakeholders specify trade-off weights λ_k (with $\sum_k \lambda_k = 1$) according to the project charter—for example, prioritizing cost over schedule in budget-capped projects. The agent stance is then aggregated as shown in Formula (1):

$$v_i(o) = \text{sign}\left(\sum_k \lambda_k s_{i,k}(o)\right) \in \{-1, 0, +1\} \tag{1}$$

where $v_i(o)$ denotes agent i stance on option o (oppose/neutral/support), and w_i is the dynamically updated trust weight of agent i. The overall consensus score is then computed as shown in Formula (2):

$$S(o) = \frac{\sum_{i=1}^{N} w_i v_i(o)}{\sum_{i=1}^{N} w_i} \tag{2}$$

which balances supportive and opposing evidence with respect to agent reliability. An option o is accepted into the decision package if $S(o) \geqslant \theta$, as shown in Formula (3), with θ typically set between 0.6 and 0.7:

$$S(o) \geq \theta, \quad \theta \in [0.6, 0.7] \tag{3}$$

In practice, lower thresholds (e.g. 0.55) admit more alternatives but risk instability, while higher thresholds (e.g. 0.75) enforce stricter consensus at the cost of longer debate rounds. A sensitivity analysis of θ is provided in Section 5.2.

The trust weights themselves evolve over time to reflect agents' historical reliability. After each debate round t, the weight of agent i is updated as shown in Formula (4):

$$w_i^{(t+1)} = (1 - \alpha) w_i^{(t)} + \alpha \hat{a}_i^{(t)} \tag{4}$$

where $\alpha \in (0, 1)$ controls the adaptation rate, and $\hat{a}_i^{(t)} \in [0, 1]$ is the normalized accuracy of agent i within a fixed evaluation window W (e.g., the last ten accepted decisions). We compute $\hat{a}_i^{(t)}$ from expert audit scores and outcome validation (e.g., whether cost deviations (CDev), schedule delays, or defect rates aligned with the agent's claims), rescaled to [0, 1]. This update mechanism ensures that reliable agents gradually gain greater influence, while less accurate or adversarial agents lose weight. Through this adaptive trust learning, the arbitration process becomes more robust, transparent, and aligned with the objectives of trustworthy AI-enabled engineering decision support.

In practice, this means that a cost engineer's consistent accuracy in budget forecasting increases their influence in later debates, which mirrors the trust-building process in project management teams.

3.3.5 Argument credibility

Each argument submitted by an agent is evaluated through a credibility score that reflects both the quality of its supporting evidence and the agent's historical accuracy. For an argument a_i supported by a set of evidence items $\mathcal{E}_i = \{e_1, ..., e_m\}$, the credibility is defined as shown in Formula (5):

$$C(a_i) = \frac{1}{m} \sum_{j=1}^{m} \left(w_s(e_j) r(e_j) h(e_j)\right) \tag{5}$$

where $w_s(e_j)$ denotes the source-type weight (standards > IFC > CAD > cost logs), $r(e_j)$ is the reliability score of each evidence item derived from its verification frequency and source quality, and $h(e_j)$ represents the historical accuracy of that evidence type in previous debate rounds. The aggregated value $C(a_i)$ is normalized within each debate topic and incorporated into the trust-weighted voting process in Formula (3) as the evidence-credibility term. Arguments supported by diverse and repeatedly validated evidence achieve higher credibility, whereas those relying on single or outdated sources are automatically down-weighted.

3.3.6 Explanation and traceability of decisions

A defining feature of the workflow is its emphasis on explainability. Each decision package is automatically documented with a reasoning chain—from the initial conflict and the evidence considered, through argument exchanges, to the final resolution. This documentation serves multiple purposes: it enhances transparency for stakeholders, supports auditing and compliance, and provides training data for improving agent strategies. By embedding traceability, the system aligns with industry requirements for accountability in project delivery[21].

3.4 Learning and system adaptation

While rule-based logic is sufficient for some tasks, construction projects demand adaptability. The workflow incorporates reinforcement learning for agents to refine their strategies.

The reinforcement learning process underlying the debate engine treats each decision topic as an individual episode. An episode begins with a defined issue—such as a design modification, cost adjustment, or schedule revision—and proceeds through up to six debate rounds. At each round t, the environment state s_t encodes the current argument graph, including agent stances, credibility scores, trust weights, and evidence coverage, while each agent selects an action $a_t \in$ {propose, support, refute, request-evidence, concede} to advance the discussion. The corresponding reward integrates both project-level outcomes and procedural quality, as shown in Formula (6):

$$r_i^{(t)} = \kappa_1(-\Delta\mathrm{CDev}) + \kappa_2(-\Delta\mathrm{SV}) + \kappa_3\ \mathrm{Transp}\ - \kappa_4\ \mathrm{Rounds} \tag{6}$$

where ΔCDev and ΔSV denote the normalized changes in CDev and schedule variance (SV) relative to the previous round, Transp is a transparency score derived from the proportion of verified evidence used in the final decision, and Rounds penalizes redundant exchanges. The coefficients were empirically set to κ_1 = 0.35, κ_2 = 0.35, κ_3 = 0.20, and κ_4 = 0.10, providing balanced sensitivity across objectives. An episode terminates once the consensus score $S(o) \geqslant \theta$ or when the maximum round limit is reached. This configuration encourages agents to reach evidence-based agreements efficiently while maintaining decision transparency and stability.

Agents receive feedback based on project outcomes—for example, whether a design change reduced CDev, or whether a scheduling decision improved on-time performance. Rewards are defined across multiple objectives: cost variance, schedule adherence, quality compliance, and debate efficiency (e.g., fewer rounds to reach consensus)[22].

Learning mechanisms also ensure robustness. Agents are trained with perturbed data, such as fluctuating material prices or weather conditions, to build resilience to uncertainty[23]. Trust values are updated dynamically: agents that consistently provide accurate and useful arguments gain higher influence in future debates, while unreliable agents are down-weighted. This creates a self-correcting system in which expertise and credibility accumulate over time.

3.5 Implementation details

The workflow has been prototyped using widely adopted digital platforms and open-source tools. BIM data are accessed through IFC interfaces, while CAD drawings are parsed via standard exchange formats (drawing (DWG)/drawing exchange format (DXF))[1]. Cost data are linked from structured spreadsheets or enterprise resource planning systems. The agent framework is implemented with a message-passing architecture, allowing asynchronous debate rounds and scalable communication. Debate logs are maintained in a controlled repository to support auditing and reproducibility, similar to common engineering document management practices, ensuring reproducibility and enabling post-project analysis.

Although advanced models such as deep neural networks and natural language processing can be integrated, the implementation emphasizes compatibility with existing construction IT environments[24]. By building on familiar data standards and workflow engines, the system lowers adoption barriers for industry practitioners while providing a platform for incremental integration of AI capabilities.

The implementation conforms to International Organization for Standardization (ISO) 19650[25,26] information management principles, exchanges IFC4.3[27] entities where available, and can export COBie-style decision records for asset handover[28].

To enhance reproducibility without altering the system footprint, the core algorithmic settings are summarized below.

(1) **Evidence ranking**. For each debate topic, retrieved evidence items e_j are scored as shown in Formula (7):

$$R(e_j) = \beta_1 r_s + \beta_2 h(e_j) + \beta_3 v(e_j) - \beta_4 d(e_j) \tag{7}$$

where r_s denotes source reliability (standards > BIM > CAD > cost logs), $h(e_j)$ the historical accuracy of the evidence type observed in prior decisions, $v(e_j)$ the count of cross-source confirmations, and $d(e_j)$ a temporal decay term. Coefficients are set to β_1 = 0.35, β_2 = 0.25, β_3 = 0.25, β_4 = 0.15 (normalized to $\sum \beta_i = 1$); the top-k items (k = 5) are forwarded to the debate module.

(2) **Reinforcement-learning formulation**. Each debate episode corresponds to one decision objective (e.g., clash disposition, cost adjustment, or schedule revision). State s_t: embedding of the current argument graph, including agent stances, trust weights, and evidence scores. Action $a_t \in$ {propose, support, refute, request-evidence, concede}. Reward $r_t = \kappa_1(-\Delta\mathrm{CDev}) + \kappa_2(-\Delta\mathrm{SV}) + \kappa_3\mathrm{Transp} - \kappa_4\mathrm{Rounds}$, with κ_1 = 0.35, κ_2 = 0.35, κ_3 = 0.20, κ_4 = 0.10; Transp is a transparency score derived from evidence coverage and traceability, and Rounds penalizes excessive exchanges. The episode terminates when consensus $S(o) \geqslant \theta$ or after six rounds.

(3) **Proximal policy optimization (PPO) hyperparameters.** Policy optimization follows a standard PPO configuration: learning rate 3×10^{-4}, discount factor $\gamma = 0.99$, GAE $\lambda = 0.95$, clip range 0.20, minibatch size 256, epochs per update 4, entropy coefficient 0.01, and value-loss coefficient 0.50. A maximum of 300 environment steps per episode is used. The PPO implementation adheres to canonical formulations and integrates with the message-passing agent architecture described above.

3.6 Metrics definition and quantification

To ensure transparency and reproducibility, the key performance indicators reported in this study were quantified using standardized engineering definitions and verifiable project data extracted from BIM, CAD, cost, and inspection records. All percentages are reported as mean ± standard deviation across $N = 5$ independent runs (random seeds 1–5), with consistent sampling windows across baseline and proposed workflows.

(1) **Unresolved Design Clashes.** Design clashes were identified through BIM-based clash detection between architectural, structural, and MEP models. A clash is classified as unresolved when it persists beyond two design-review iterations or remains in the model at milestone $M_k + \Delta t$ ($\Delta t = 14$ days). The reduction ratio is expressed as shown in Formula (8):

$$\text{ClashReduction}\ (\%) = 100 \times \frac{N_{\text{baseline}} - N_{\text{workflow}}}{N_{\text{baseline}}} \tag{8}$$

where N_{baseline} and N_{flow} are the average clash counts before and after applying the proposed workflow. For volumetric conflicts, a normalized measure is used as shown in Formula (9):

$$\text{ClashVol}\ = 100 \times \frac{V_{\text{conflict}}}{V_{\text{relevant elements}}} \tag{9}$$

where $V_{\text{conflicts}}$ denotes the intersected volume of conflicted geometry and $V_{\text{relevant elements}}$ represents the total volume of relevant model elements.

(2) **CDev.** Cost accuracy was measured as the absolute percentage deviation between the forecasted cost $\widehat{C}$ and the actual or settled cost C, as shown in Formula (10):

$$\text{CDev} = 100 \times \frac{|\widehat{C} - C|}{C} \tag{10}$$

Forecasts were drawn from the BoQ-based cost model at each decision point, while actual costs were obtained from corresponding progress claims or final accounts. Mean values were aggregated over all decision events within each scenario.

(3) **SV.** Schedule performance was evaluated as the percentage difference between planned and actual progress as shown in Formula (11):

$$SV = 100 \times (\ \text{Planned}\ \% - \ \text{Actual}\ \%) \tag{11}$$

where progress percentages were computed from earned-value data and critical path method (CPM) schedules. "Disruption" scenarios were simulated by introducing material delivery delays or resource shortages; the reported SV corresponds to the deviation averaged over the disruption window.

(4) **First-Pass Acceptance Rate (FPAR).** Quality performance was measured by the proportion of inspection items accepted at the first attempt without rework, as shown in Formula (12):

$$\text{FPAR} = 100 \times \frac{N_{\text{pass}}}{N_{\text{total}}} \tag{12}$$

where N_{pass} is the number of items meeting tolerance and specification criteria in initial inspections, and N_{total} is the total number of inspected items. Data were derived from digital inspection logs linked to BIM object identifiers.

All metrics were computed using consistent sampling windows and verified against project documentation. The same definitions were applied to both baseline and experimental workflows to ensure comparability across scenarios.

4. Experiments

4.1 Case scenarios and datasets

To evaluate the proposed workflow, four representative scenarios were designed, reflecting typical points of conflict in construction projects. Each scenario was constructed using real project data where possible, supplemented by anonymized datasets and synthetic

variations to test robustness.

All performance indicators and quantification procedures used in the experiments have been defined in Section 3.6. This section presents the datasets and experimental setup under those standardized evaluation criteria.

4.1.1 Design optimization example

This scenario was based on the schematic design of a mixed-use commercial complex[29]. BIM and CAD data included structural layouts, mechanical and electrical routing, and architectural plans. Conflicts emerged in space allocation where heating, ventilation, and air conditioning ducts intersected with structural beams[30]. Agents debated multiple layout alternatives, with the design agent prioritizing architectural coherence, the construction agent emphasizing buildability, and the cost agent estimating financial implications.

4.1.2 Costing and budget change scenario

The second scenario focused on cost estimation for a high-rise office building. BoQ data, historical unit prices, and market cost indices formed the dataset. During the process, the design agent introduced a material substitution to improve façade aesthetics, while the cost agent raised concerns about exceeding the budget cap[31,32]. The debate mechanism allowed agents to quantify CDev, compare alternatives, and reach consensus on a compromise solution.

4.1.3 Construction scheduling case

In the third scenario, drawn from a large infrastructure project, the construction schedule included critical path activities and resource assignments. A sudden disruption, simulated as delayed steel delivery due to supply chain issues, triggered conflicts[33]. The scheduling agent proposed re-sequencing, the construction agent suggested overtime work, and the cost agent flagged budget risks[34]. Debates enabled agents to weigh trade-offs and recommend a balanced adjustment.

4.1.4 Quality acceptance case

The final scenario addressed quality inspection at the acceptance stage of a residential development. BIM models were linked with defect checklists, site inspection images, and compliance standards. Quality agents identified deviations such as misaligned precast panels and incomplete fireproofing. While the quality agent pushed for corrective rework, the cost and schedule agents argued for tolerances in non-critical areas. The debate process documented evidence and rationales, resulting in transparent decisions on defect closure.

4.1.5 Cross-scenario quantitative summary

Table 2 summarizes the outcomes across the four scenarios. The proposed workflow reduces clashes in Design by 42%, lowers cost forecast error from 12% to 5%, improves SV from 10% to 8.2%, and raises quality acceptance from 74% to 88%, with the rework ratio reduced by 30%.

Table 2. Performance comparison across representative scenarios.

Scenario	Baseline (%)	Proposed Workflow (%)	Acceptance Rate (%)	Notes
Design	Clash count: 100 ± 0; clash volume: 100 ± 0; revisions: 100 ± 0	Clash count reduced by 42 ± 3; clash volume reduced by 57.1 ± 3.2; revisions reduced by 62 ± 4	N/A	All values expressed as reduction relative to baseline.
Cost	CDev: 12.0 ± 0.8	-58.3 ± 2.1	N/A	Avg. debate rounds: 4.0 ± 0.3.
Schedule	SV: 10.0 ± 0.6	-18.0 ± 1.5	N/A	Improvement in schedule adherence.
Quality	First-pass acceptance: 74 ± 2	+18.9 ± 2.0	88 ± 1	Rework reduced by 30 ± 3; transparency score: 4.6 ± 0.25/5.

All results are reported as percentage improvements relative to the baseline, expressed as mean ± standard deviation over $N = 5$ independent runs. Acceptance Rate applies only to the Quality scenario (marked N/A for others). CDev: cost deviation; SV: schedule variance; N/A: not applicable; Avg.: average.

Synthetic variations were generated by perturbing (1) unit prices with a log-normal factor ($\mu = 0$, $\sigma = 0.12$), (2) delivery delays with a triangular distribution (min = 0, mode = 3, max = 7 days), and (3) defect occurrence with Bernoulli draws calibrated to historical rates ($p = 0.08 - 0.15$). Random seeds {1-5} were fixed across scenarios for comparability.

4.2 Experimental setup

All scenarios were implemented using BIM models in IFC format, CAD drawings in DWG/DXF, and structured spreadsheets for cost data. The multi-agent framework was deployed on a message-passing architecture, allowing asynchronous debates. For machine learning components, agents employed reinforcement learning with PPO and used lightweight neural networks for evidence scoring. Simulations were run on a workstation with 64 GB RAM and an NVIDIA RTX 6000 GPU, although computational demands remained moderate given the scale of tasks.

Baselines were established for comparison:

(1) Rule-based coordination (fixed decision rules without debate).

(2) Traditional BIM clash detection (for design conflicts).

(3) Single-agent optimization (each task optimized independently).

This experimental design, datasets, and configurations establish a consistent foundation for evaluation across representative project scenarios. The following section presents quantitative results, ablation studies, and a case demonstration derived from these experimental settings.

5. Results

5.1 Results and analysis

Results across the four scenarios consistently showed improvements over baselines.

Design Optimization: The debate workflow reduced the number of unresolved clashes by 42% (ClashCnt from 100 to 58), while the conflicted volume relative to element volume (ClashVol) decreased from 7% to 3% (a 57.1% reduction). The final design also required 25% fewer revisions during later phases, as summarized in Table 2. Reported values in Table 2 represent the mean ± standard deviation across five independent simulation runs, reflecting robustness to data variability. This reduction directly lowers redesign workload for architects and engineers, shortening the design-construction information flow and reducing downstream delays.

Costing Scenario: The workflow improved cost forecast accuracy, lowering average deviation from 12% (rule-based method) to 5%. Debate convergence required an average of four rounds, balancing precision with efficiency.

Construction Scheduling: In disruption simulations, the proposed method reduced SV by 18% compared to rule-based re-sequencing, while keeping cost overrun within 2%.

Quality Acceptance: The FPAR improved from 74% (traditional inspection) to 88% with debate, while the rework ratio dropped by 30%. Transparency scores by independent experts averaged 4.6/5, confirming the value of traceable reasoning, as also reported in Table 2. Decision transparency was assessed by three independent experts using a five-point Likert scale, following the procedure defined in Section 3.6. Higher scores indicate greater explainability and traceability of the debate records. This improvement reflects fewer disputes during handover, reducing warranty claims and enhancing owner satisfaction.

These results indicate that the workflow not only improves engineering performance but also enhances decision quality and stakeholder trust. As shown in Figure 3, CDev decreases from 12% (baseline) to about 5% within four debate rounds, while Figure 4 illustrates that SV is reduced from 10% to approximately 8.2% by round 4, confirming consistent and robust convergence.

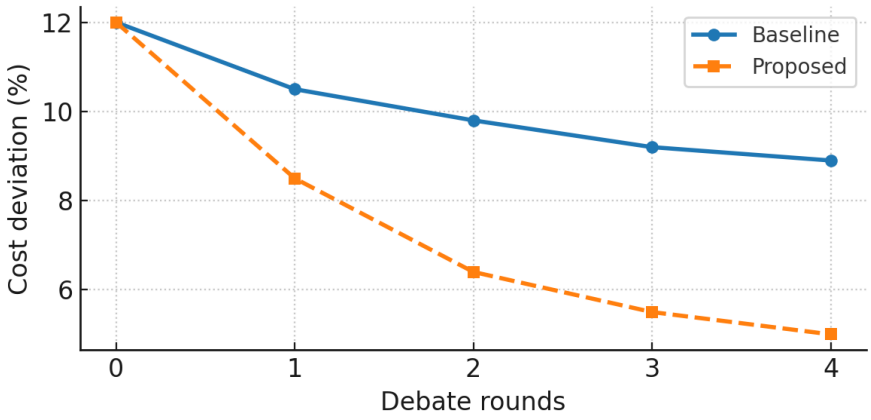

Figure 3. CDev across debate rounds. Curves show mean ± std across five runs (random seeds 1-5). The dashed horizontal line denotes the rule-based baseline (12%), and the solid curve shows the proposed workflow converging by round 4 to approximately 5%. Metrics correspond to Table 2.

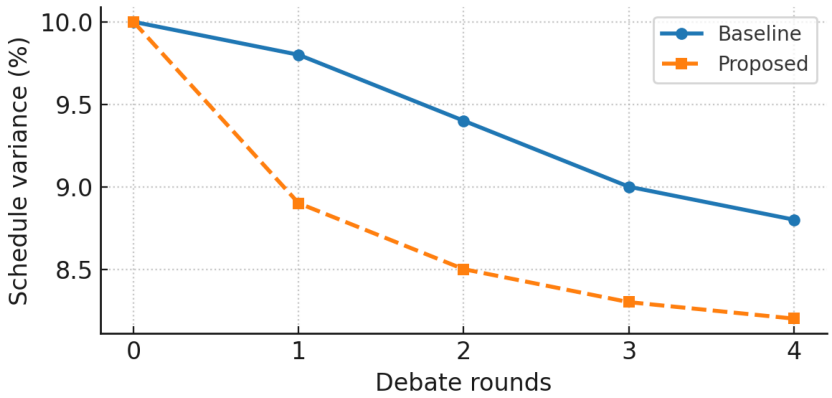

Figure 4. SV across debate rounds. Curves show mean ± std across five runs (random seeds 1-5). The dashed horizontal line denotes the rule-based baseline (10%), and the solid curve shows the proposed workflow reaching approximately 8.2% by round 4. Metrics correspond to Table 2. SV: schedule variance.

System efficiency metrics are summarized in Table 3, showing that runtime per round remained under practical limits, and scalability was sublinear when adding agents.

Table 3. System efficiency metrics across different agent configurations.

Configuration	Median time/round (s)	Scaling effect (%)
4 agents, 3 objectives	7.8 (IQR 6.9-8.6)	baseline
95th percentile time	10.2	N/A
6 agents, 3 objectives	9.6	+23% (sublinear)

Measurements were obtained on a workstation equipped with 64 GB RAM and an NVIDIA RTX 6000 GPU. Median time per debate round is reported in seconds. Scaling effect (%) represents the relative change in median time compared with the 4-agent baseline. IQR: interquartile range; N/A: not applicable.

5.2 Ablation and sensitivity studies

To examine the contribution of each component, ablation experiments were conducted[35].

· Removing evidence retrieval reduced transparency scores by 40% and led to higher cost variance.

· Excluding the arbiter agent resulted in longer debates (average 8 rounds instead of 4) and more unresolved conflicts.

· Eliminating trust weighting decreased decision stability, as less reliable agents exerted equal influence.

As shown in Table 4, removing evidence retrieval significantly increased cost variance and reduced transparency, while excluding the arbiter agent nearly doubled the number of rounds required for convergence. Trust weighting also contributed to decision stability, preventing unreliable agents from exerting excessive influence.

Table 4. Ablation study results under different component configurations.

Configuration	Final CDev (%)	Final SV (%)	Transparency (/5)	Avg. Debate Rounds
Full workflow (all components)	5.0 ± 0.4	8.2 ± 0.3	4.6 ± 0.2	4.0 ± 0.3
No evidence retrieval	7.8 ± 0.6	9.0 ± 0.4	2.7 ± 0.5	4.2 ± 0.4
No arbiter agent	5.5 ± 0.5	8.6 ± 0.4	4.2 ± 0.3	8.1 ± 0.6
No trust weighting	5.6 ± 0.4	8.7 ± 0.5	4.0 ± 0.4	4.5 ± 0.4

Results are reported as mean ± standard deviation across five independent runs. CDev and SV are expressed as percentages. Transparency is evaluated on a five-point scale, and debate rounds indicate the average number of iterations per decision episode. CDev: cost deviation; SV: schedule variance; Avg.: average.

Sensitivity analyses showed that the system remained robust under variations in debate thresholds and cost weighting factors, with performance degrading gracefully rather than collapsing[36]. To further examine robustness, we varied the decision threshold θ from 0.55 to 0.75. As summarized in Table 5, lower thresholds accelerated convergence but slightly reduced decision stability, whereas higher thresholds enforced stricter consensus with marginally more rounds required. Engineering performance metrics (CDev, SV) remained nearly unchanged, confirming graceful degradation.

Table 5. Sensitivity analysis of the consensus threshold θ.

θ	Avg. debate rounds	Final CDev (%)	Final SV (%)
0.55	3.6	5.2	8.5
0.60	3.9	5.1	8.3
0.65	4.2	5.0	8.2
0.70	4.5	5.1	8.2
0.75	4.8	5.3	8.4

Avg. debate rounds denote the mean number of iterations required to reach consensus. CDev and SV are expressed as percentages. Results indicate that system performance remains stable when θ ranges between 0.60 and 0.70. Avg.: average; CDev: cost deviation; SV: schedule variance.

To further examine the stability of the trust-weighted consensus mechanism (Formulas (1-4)), we conducted a concise sensitivity study on the trust update step α and the multi-objective weights $\{\lambda_k\}$. Five values of α (0.1-0.5) and three representative weight profiles for cost (C), schedule (S), and quality (Q) were tested: cost-prioritized (0.5, 0.25, 0.25), balanced $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, and schedule-prioritized (0.25, 0.5, 0.25). Each configuration was evaluated across N = 20 debate topics with five random seeds. The key indicators, average rounds to consensus (R), consensus-score volatility (σ_S), and decision stability under evidence perturbation, are summarized in Table 6.

Table 6. Sensitivity analysis of trust update rate α and objective weights $\{\lambda_k\}$.

Setting	Rounds to consensus $R \downarrow$	Score volatility $\sigma_S \downarrow$	Stability (%) $\uparrow$
α = 0.1, (0.5, 0.25, 0.25)	5.1 ± 0.6	0.061 ± 0.008	92.1
α = 0.2, (0.5, 0.25, 0.25)	4.4 ± 0.4	0.073 ± 0.010	90.4
α = 0.3, (1/3, 1/3, 1/3)	4.0 ± 0.3	0.068 ± 0.007	94.7
α = 0.4, (1/3, 1/3, 1/3)	3.7 ± 0.4	0.085 ± 0.011	91.5
α = 0.5, (0.25, 0.5, 0.25)	3.5 ± 0.5	0.112 ± 0.013	87.6

Results are reported as mean ± standard deviation over five random seeds for N = 20 topics. α controls the trust update rate between agents, while $\{\lambda_k\}$ denotes the objective weighting set for design, cost, and schedule goals. R indicates the average number of debate rounds to reach consensus; σ_S measures the volatility of consensus scores, and Stability (%) represents the fraction of stable outcomes across runs. $\downarrow$: decrease; $\uparrow$: increase.

The results indicate a clear trade-off between adaptation speed and stability. Smaller α values yield smoother convergence but require more debate rounds, while larger α values accelerate consensus at the cost of higher volatility and occasional decision flips. Across all settings, the balanced weight profile $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ achieves the most consistent outcomes, with decision stability above 94%. When cost or schedule is over-weighted, the system converges slightly faster (about one round fewer) but exhibits 3-5% lower stability. These observations suggest that α between 0.2-0.3 and a balanced λ configuration provide the best overall performance.

5.3 Case study demonstration

A full project case study was conducted on a commercial high-rise design. The workflow was applied from schematic design through acceptance. Over a six-month period, the system facilitated 53 debates covering design changes, cost adjustments, scheduling conflicts, and defect inspections. Compared to a matched control project managed with conventional processes, the case study achieved the following. The control project was a comparable high-rise office building developed by the same client within the same urban district. Both projects had similar scale (approximately 60,000 m^2), duration (18-20 months), and procurement model (design-bid-build). This ensured comparability while isolating the impact of the proposed workflow:

· 9.4% lower final cost variance,

· 11.6% reduction in total project duration,

· 17% fewer requests for information,

· 14% higher stakeholder satisfaction scores in post-project surveys.

The case study confirmed that the workflow could scale beyond isolated scenarios to support continuous coordination in real projects, as shown in Figure 5.

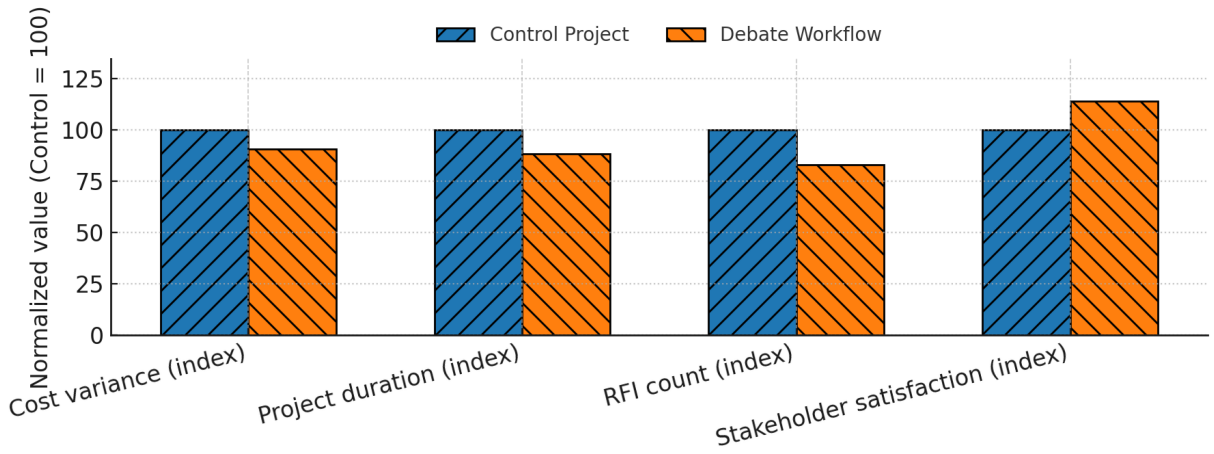

Figure 5. Case study comparison of key performance indicators. Values are normalized with the control project set to 100. The debate-oriented workflow reduces cost variance to 90.6, shortens project duration to 88.4, lowers the number of RFIs to 83.0, and increases stakeholder satisfaction to 114.0 relative to the control project. RFI: requests for information.

5.4 Reproducibility

All datasets, discussion logs, and experimental scripts have been anonymized and archived. BIM and CAD files are standardized using IFC and DWG formats. Cost and schedule data are de-identified before release, and debate records and decision chains are version controlled. To respect project confidentiality, only derived metrics and anonymized models will be made public, while sensitive project details will remain protected.

6. Discussion

6.1 Adoption pathways for engineering practice

Embedding AI-driven debate mechanisms into construction workflows offers a structured way to reconcile the long-standing fragmentation among design, cost, construction, and quality management. Rather than replacing existing digital systems, the framework can be incrementally integrated into current project delivery environments. At the design stage, debate logs and argument graphs can be linked to BIM issue-tracking tools (e.g., Navisworks, BIMCollab) and stored alongside clash reports and revision histories, aligning with collaborative principles in integrated project delivery (IPD) and Lean Construction[37]. At later stages, integration with cost and schedule management platforms enables automated generation of debate packages from BoQ and progress data, creating traceable connections between design reasoning, cost adjustments, and quality outcomes. The workflow also complements Construction 4.0 and Smart Construction initiatives and can be aligned with the process groups of Project Management Body of Knowledge (PMBOK)[38], providing a bridge to established management standards. In practice, this creates a continuous feedback loop: debate records and decision packages can inform digital twins during construction and operation, linking site feedback to future design baselines to enhance building performance[39-41]. By embedding the debate process within routine coordination meetings or 4D planning sessions, organizations can gradually build a repository of transparent, auditable decisions that strengthen accountability and collective learning over time.

6.2 Implementation challenges and limitations

Despite the promising results, several challenges constrain large-scale adoption[42,43]. A primary technical issue is data heterogeneity: BIM, CAD, cost, and inspection data often differ in schema and quality, requiring normalization and ontology alignment to ensure interoperability. User acceptance also represents a cultural challenge. Although the workflow increases transparency, professionals may initially perceive AI-assisted debate as an additional layer of scrutiny, underscoring the need for clear rationale tracing, threshold tuning, and user training. Scalability is another concern—iterative debate rounds and graph propagation become computationally intensive for megaprojects with thousands of design components[36]. Practical deployment therefore requires careful orchestration strategies such as batching or pruning of argument graphs for real-time use. Furthermore, the experimental evaluation focused on representative scenarios and a single extended case study. While these provide proof of concept, broader validation across project types and regions is necessary. The current framework models cost, schedule, and quality objectives, yet other critical aspects—such as sustainability, safety culture, and long-term operational performance—remain to be incorporated. Future studies should explore these dimensions to achieve more comprehensive decision support and enhance generalizability.

In addition, potential bias in agent weighting may influence decision outcomes if trust scores are disproportionately affected by early debate performance or limited historical data. This emphasizes the need for calibration procedures and periodic rebalancing to prevent dominance by specific agents. The framework also relies on the accuracy of IFC-based representations; errors or omissions

in model attributes can propagate through the debate process, affecting evidence retrieval and reasoning quality. Future implementations should therefore include model validation and uncertainty assessment steps to mitigate such data dependencies.

6.3 Ethical and safety considerations

The increasing role of AI in construction decision-making raises ethical and safety concerns that require proactive management[44]. The proposed workflow mitigates these risks through a strict human-in-the-loop design: agents generate structured recommendations and ranked options, but final validation and approval remain with licensed engineers and project managers. At each debate round, human reviewers can override or reject AI-derived arguments that conflict with regulatory constraints or professional judgment, ensuring that responsibility is never displaced onto autonomous systems. All decisions are logged with traceable authorship, timestamps, and rationale paths, forming a digital audit trail that supports post-decision accountability.

Furthermore, liability allocation must be explicitly defined when AI-driven debates influence safety-critical or design decisions. Under the proposed governance model, AI developers and system integrators are accountable for algorithmic transparency and documentation, while project owners and supervising engineers retain legal responsibility for implementation outcomes. This delineation follows established engineering liability principles, ensuring that the AI system functions as an assistive decision-support tool rather than an autonomous authority. To safeguard human oversight, multi-level approval checkpoints, at design review, construction execution, and final acceptance, should remain mandatory. Such safeguards balance innovation with safety, aligning the debate framework with international guidelines for trustworthy and explainable AI in engineering practice.

7. Conclusion

This study proposed a debate-oriented multi-agent workflow for cross-stage decision-making in construction projects. By integrating design, cost, construction, and quality agents into a unified framework, the approach enables early identification and resolution of conflicts through structured debate supported by verifiable evidence from BIM models, CAD drawings, and project records. The mechanism transforms fragmented decision processes into transparent, auditable reasoning cycles, generating outcomes that are technically consistent, economically viable, and operationally feasible.

The framework can be incrementally adopted within existing industry practices, aligning with the process groups of PMBOK[38], Lean Construction principles, and IPD. Its conformity with ISO 19650[25,26] information management standards and compatibility with IFC/COBie-based data environments ensures integration with ongoing digital transformation initiatives such as BIM-based delivery and digital twins[45]. In practice, the system enhances decision transparency and accountability, providing managers with a verifiable digital record of project reasoning and outcomes.

Future research will extend the framework to encompass sustainability, safety culture, and long-term operational performance, as well as to validate its scalability in large and international projects. Rather than replacing established management procedures, the debate-driven workflow is intended to strengthen evidence-based collaboration and adaptive learning across project stages. Ultimately, by embedding AI-driven debate into the fabric of construction workflows, this research lays a foundation for more intelligent, transparent, and collaborative project delivery in the era of smart construction.

Declarations

Authors contribution

The author contributed solely to the article.

Conflicts of interest

The author declares no conflicts of interest.

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

Funding

None.

References

1. Liu X, Zhao J, Yu Y, Ji Y. BIM-based multi-objective optimization of clash resolution: A NSGA-II approach. J Build Eng. 2024;89:109228. [DOI]
2. Hu Y, Xia C, Chen J, Gao X. Clash context representation and change component prediction based on graph convolutional network in MEP disciplines. Adv Eng Inform. 2023;55:101896. [DOI]
3. Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I. Improving factuality and reasoning in language models through multiagent debate. arXiv:2305.14325 [Preprint]. 2023. [DOI]
4. Yao D, de Soto BG. Enhancing cyber risk identification in the construction industry using language models. Autom Constr. 2024;165:105565. [DOI]
5. Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries—Part 1: Data schema [Internet]. Geneva: ISO; 2024. Available from: https://www.iso.org/standard/84123.html
6. Mirindi D, Mirindi F, Bezabih T, Sinkhonde D, Kiarie W. The role of artificial intelligence in building information modeling. In: Riemenschneider C, Sullivan Y, Dinger M, Bantan M, Roberts N, editors. Proceedings of the 2025 Computers and People Research Conference; 2025 May 28-30; Texas, USA. New York: Association for Computing Machinery; 2025. p. 1-11. [DOI]
7. Gao Y, Gan Y, Chen Y, Chen Y. Application of large language models to intelligently analyze long construction contract texts. Constr Manag Econ. 2024;43(3):226-242. [DOI]
8. Islam T, Bappy FH, Zaman TS, Sajid MS, Pritom MM. MRL-PoS: A multi-agent reinforcement learning based proof of stake consensus algorithm for blockchain. In: 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC); 2024 Jan 8; Las Vegas, USA. Piscataway: IEEE; 2024. p. 0409-0413. [DOI]
9. Bentahar J, Baharloo N, Drawel N, Pedrycz W. Model checking combined trust and commitments in multi-agent systems. Expert Syst Appl. 2024;243:122856. [DOI]
10. Akintunde M, Yazdanpanah V, Fathabadi AS, Cirstea C, Dastani M, Moreau L. Actual trust in multiagent systems. In: Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems; 2024 May 6-10; Auckland, New Zealand. Richland: International Foundation for Autonomous Agents and Multiagent Systems; 2024. p. 2114-2116.
11. Akbari B, Yuan M, Wang H, Zhu H, Shan J. A factor graph model of trust for a collaborative multi-agent system. arXiv: 2402.07049 [Preprint]. 2024. [DOI]
12. Milat M, Knezić S, Sedlar J. Resilient scheduling as a response to uncertainty in construction projects. Appl Sci. 2021;11(14):6493. [DOI]
13. Wang X, Yu H, McGee W, Menassa CC, Kamat VR. Enabling Building Information Model-driven human-robot collaborative construction workflows with closed-loop digital twins. Comput Ind. 2024;161:104112. [DOI]
14. Ko J, Ajibefun J, Yan W. Experiments on Generative AI-powered parametric modeling and BIM for architectural design. arXiv: 2308.00227 [Preprint]. 2023. [DOI]
15. Pan Z, Guo F. Digital twin (DT) development strategies in the architecture, engineering, and construction (AEC) industry based on SWOT-AHP analysis. J Build Des Environ. 2025;3:202441. [DOI]
16. Liang T, He Z, Jiao W, Wang X, Wang Y, Wang R, Yang Y, Shi S, Tu Z. Encouraging divergent thinking in large language models through multi-agent debate. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024; Florida, USA. Stroudsburg: Association for Computational Linguistics; 2024. p. 17889-17904. [DOI]
17. Chaudhary N, Uddin SM, Chandra SS, Ovid A, Albert A. Prompt to protection: A comparative study of multimodal LLMs in construction hazard recognition. arXiv: 2506.07436v1 [Preprint]. 2025. [DOI]
18. Wu Q, Bansal G, Zhang J, Wu Y, Li B, Zhu E, et al. Autogen: Enabling next-gen LLM applications via multi-agent conversations. arXiv: 2308.08155 [Preprint]. 2023. [DOI]
19. Li G, Hammoud HAAK, Itani H, Khizbullin D, Ghanem B. Camel: Communicative agents for "mind" exploration of large language model society. arXiv: 2303.17760v2 [Preprint]. 2023. [DOI]
20. Hong S, Zhuge M, Chen J, Zheng X, Cheng Y, Wang J, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. arXiv: 2308.00352 [Preprint]. 2023. [DOI]
21. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: Harmlessness from AI feedback. arXiv: 2212.08073 [Preprint]. 2022. [DOI]
22. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. arXiv: 2203.11171 [Preprint]. 2022. [DOI]
23. Li J, Yu H, Deng X. A systematic review of the evolution of the concept of resilience in the construction industry. Buildings. 2024;14(9):2643. [DOI]
24. Wong S, Zheng C, Su X, Tang Y. Construction contract risk identification based on knowledge-augmented language models. Comput Ind.

2024;157-158:104082. [DOI]
25. Organization and digitization of information about buildings and civil engineering works, including building information modelling (BIM)—information management using building information modelling—Part 4: Information exchange [Internet]. Geneva: ISO; 2022. Available from: https://www.iso.org/standard/78246.html
26. Organization and digitization of information about buildings and civil engineering works, including building information modelling (BIM)—information management using building information modelling—Part 5: Security-minded approach to information management [Internet]. Geneva: ISO; 2020. Available from: https://www.iso.org/standard/74206.html
27. IFC 4.3.2.0 official documentation [Internet]. Kings Langley: buildingSMART International Limited; 2024. Available from: https://standards.buildingsmart.org/IFC/RELEASE/IFC4_3
28. Zhao W, Wang K, Guo F, Hao JL, Zhang Q, Qi J. Unraveling barriers to digital twin adoption under Construction 4.0: A DEMATEL-ISM-MICMAC approach. J Build Des Environ. 2025;3:202438. [DOI]
29. Zheng C, Wong S, Su X, Tang Y, Nawaz A, Kassem M. Automating construction contract review using knowledge graph-enhanced large language models. Autom Constr. 2025;175:106179. [DOI]
30. Bitaraf I, Salimpour A, Elmi P, Shirzadi Javid AA. Improved building information modeling based method for prioritizing clash detection in the building construction design phase. Buildings. 2024;14(11):3611. [DOI]
31. Tran SVT, Yang J, Hussain R, Khan N, Kimito EC, Pedro A, et al. Leveraging large language models for enhanced construction safety regulation extraction. J Inf Technol Constr. 2024;29:1026-1038. [DOI]
32. Saeidlou S, Ghadiminia N. A construction cost estimation framework using DNN and validation unit. Build Res Inf. 2024;52(1-2):38-48. [DOI]
33. Ahmadi E, Wang C. Transforming construction practices with large language models. In: Leathem T, Collins W, Perrenoud A, editors. Proceedings of 60th Annual Associated Schools of Construction International Conference; 2024 April 3-5; Alabama, USA. Manchester: EasyChair; 2024. p. 414-422. [DOI]
34. Zeng N, Liu Y, König M. 4D BIM-enabled look-ahead scheduling for early warning of off-site supply chain disruptions. J Constr Eng Manag. 2023;149(1):04022154. [DOI]
35. Yao S, Yu D, Zhao J, Shafran I, Griffiths T, Cao Y, et al. Tree of thoughts: Deliberate problem solving with large language models. Adv Neural Inf Process Syst. 2023;36:11809-11822. [DOI]
36. Kenton Z, Siegel N, Kramár J, Brown-Cohen J, Albanie S, Bulian J, et al. On scalable oversight with weak LLMs judging strong LLMs. Adv Neural Inf Process Syst. 2024;37:75229-75276. [DOI]
37. Yang L, Dan I, Xu Y, Shuohang W, Xu R, Chenguang Z. G-eval: Nlg evaluation using GPT-4 with better human alignment. arXiv: 2303.16634 [Preprint]. 2023. [DOI]
38. Project Management Institute. A guide to the project management body of knowledge (PMBOK® Guide). 8th ed. PA: Project Management Institute; 2021. Available from: https://www.pmi.org/standards/pmbok
39. 8.Kampelopoulos D, Tsanousa A, Vrochidis S, Kompatsiaris I. A review of LLMs and their applications in the architecture, engineering and construction industry. Artif Intell Rev. 2025;58(8):250. [DOI]
40. Amer F, Koh HY, Golparvar-Fard M. Automated methods and systems for construction planning and scheduling: Critical review of three decades of research. J Constr Eng Manag. 2021;147(7):03121002. [DOI]
41. Wang Y, Chen J, Xiao B, Mueller ST, Guo J. Causation analysis of crane-related accident reports by utilizing ChatGPTand complex networks. J Build Des Environ. 2025;3(2):202535. [DOI]
42. Alathamneh S, Collins W, Azhar S. BIM-based quantity takeoff: Current state and future opportunities. Autom Constr. 2024;165:105549. [DOI]
43. Koo HJ, Guerra BC. Clash relevance prediction in BIM model coordination using artificial neural network. In: Shane JS, Madson KM, Mo Y, Poleacovschi C, Sturgill RE, editors. Construction Research Congress 2024; 2024 March 20-23; Iowa, USA. Reston: ASCE Press; 2024. p. 127-136. [DOI]
44. Hong Y, Guo F. A framework of BIM-IoT application in construction projects through multiple case study approach. J Build Des Environ. 2025;3(1):202438. [DOI]
45. Wang L, Li J, Ye Q, Li Y, Feng A. Automatic planning method of construction schedule under multi-dimensional spatial resource constraints. Buildings. 2024;14(10):3231. [DOI]

Publisher's Note