2026 European Conference on Computing in Construction
Corfu, Greece
July 12–15, 2026

ENABLING CROSS-STUDY COMPARISON: A FRAMEWORK FOR AUTOMATED BIM-QA EVALUATION

Sylvain Hellin[1,2], Stefan Fuchs[1,2], Stavros Nousias[1,2], André Borrmann[1,2]
[1]Chair of Computing in Civil and Building Engineering, Technical University of Munich, Germany
[2]TUM Georg Nemetschek Institute, Technical University of Munich, Munich, Germany

Abstract

Driven by LLM advances, Question Answering (QA) systems for Building Information Models (BIM) have proliferated, yet no standardized methods exist for cross-study comparison. We present an automated evaluation pipeline employing LLM judges to assess BIM-QA answers across five quality criteria and validate it through inter-rater agreement analysis with two domain experts and two LLM judges on 74 IFC-Bench question-answer pairs. LLM judges achieve high inter-rater reliability (Krippendorff's α = 0.70–1.00), exceeding human experts (α = 0.32–0.57), and show good agreement with expert consensus (α = 0.48–0.85), positioning the pipeline as a suitable solution for automated, reproducible BIM-QA benchmarking.

Introduction

Building Information Models (BIMs) store comprehensive building data, yet accessing this information requires specialized software and domain expertise. This creates a barrier for many project stakeholders, such as facility managers, owners, or contractors, who need timely access to building information but lack the technical skills to navigate complex BIM software. Natural language Question Answering (QA) systems offer a promising solution by enabling stakeholders to query building models using natural language (Zheng and Fischer, 2023; Hellin et al., 2025). As these systems transition from research prototypes toward practical deployment, a critical question arises: how can we systematically evaluate their answer quality?

Traditional exact-match metrics fail when answers vary in phrasing, and automated metrics like BLEU and ROUGE correlate poorly with human preference when multiple valid responses exist (Liu et al., 2023). This limitation reveals a deeper challenge: the BIM-QA field currently lacks standardized evaluation methodology. Research groups typically develop custom benchmarks and evaluate on non-shared datasets, making cross-study comparison difficult and hindering cumulative scientific progress.

While IFC-Bench provides a shared benchmark dataset (Hellin et al., 2025), a consistent evaluation methodology remains missing. A validated, reproducible evaluation method would enable meaningful performance comparison across systems and research groups. For non-deterministic LLM-based systems, such standardized evaluation is essential before practitioners can trust them for professional use.

This paper addresses this gap by introducing an automated LLM-based evaluation pipeline for BIM-QA systems. The pipeline implements a multi-dimensional evaluation framework comprising a four-category question taxonomy and five binary quality criteria (Hellin et al., 2026), detailed in the Evaluation Framework section. While multi-dimensional evaluation offers richer assessment than binary accuracy, its practical utility depends on reliable and automated application across evaluators. Specifically, this validation targets inter-rater *reliability*: whether different evaluators reach the same judgments when assessing the same answers against the same ground truth. Assessing evaluation *validity* for novel questions without established ground truth remains a separate, open direction. To validate the pipeline and the underlying evaluation framework, we investigate the following research questions:

- **RQ1:** What discriminative power and redundancy do the evaluation criteria exhibit across BIM-QA systems of varying quality?
- **RQ2:** How reliably can the proposed framework be applied across different evaluators?
- **RQ3:** How well do LLM judges approximate human expert evaluation?

The experimental results yield three key findings:

- *Criteria differ in discriminative power and redundancy (RQ1).* *Transparency* and *Completeness* best differentiate BIM-QA systems (56–57 percentage points (pp) cross-system range). *Relevance* shows ceiling effects (38pp range) and high correlation with *Completeness* (ρ = 0.76–0.81), suggesting potential consolidation to four criteria.
- *LLM judges achieve higher inter-rater reliability than human experts (RQ2).* Percentage agreement exceeds 77% across all evaluator combinations and criteria. However, chance-corrected reliability (α) reveals a clear split: LLM-LLM agreement is high (α = 0.70–1.00), while human-human agreement remains lower (α = 0.32–0.57).
- *Automated evaluation enables standardized cross-study comparison (RQ3).* Post-hoc consensus analysis confirms that LLMs align with calibrated expert judgment better than uncalibrated experts agreed with each other. This alignment holds across two model families (GLM-

4.7 and Gemini), positioning LLM-as-judge as a practical solution to the field's fragmentation problem.

Related work

BIM question answering systems

The evolution from BIM Natural Language Query (NLQ) systems to generative BIM-QA systems has transformed evaluation requirements. Traditional BIM-NLQ systems translate natural language into structured queries (e.g., BIMQL, SPARQL), enabling evaluation through query execution correctness: the generated query either matches the expected query or it does not (Yin et al., 2023; Guo et al., 2020). In this paradigm, evaluation remains tractable: correct queries produce correct results.

Recent BIM-QA systems go beyond query translation, responding in natural language directly and introducing evaluation challenges absent in query translation approaches. Prior work presented an LLM-based agentic workflow achieving 79.8% accuracy on IFC-Bench, but acknowledged that binary accuracy fails to capture nuanced answer quality: a system might provide correct values without disclosing calculation methods, leaving engineers unable to verify results (Hellin et al., 2025). This transition from query generation to answer generation requires different evaluation approaches, as binary accuracy cannot capture the nuanced quality dimensions that practitioners need. Multi-dimensional frameworks address this gap by decomposing answer quality into distinct, measurable criteria, motivating their application to BIM-QA.

Multi-dimensional QA evaluation

Beyond BIM, the inadequacy of traditional metrics for evaluating generative QA has driven the development of multi-dimensional frameworks. When evaluating long-form QA answers, domain experts consistently prioritize factuality and completeness (Xu et al., 2023). Traditional reference-based metrics such as BLEU and ROUGE, which rely on surface-level lexical overlap between generated and reference answers, fail to capture these semantic quality dimensions (Liu et al., 2023). The RAGAS framework (Es et al., 2024) operationalized evaluation through faithfulness (grounding in retrieved context), answer relevance, and context precision, while other work distinguished *correctness* from *faithfulness* (Adlakha et al., 2024). Notably, these frameworks lack explicit transparency criteria, specifically disclosure of methods and sources, identified as essential for high-stakes technical decisions where practitioners must verify results (Kell et al., 2024). Despite these advances, no validated and automated evaluation method currently exists for BIM-QA systems or AEC-related QA systems in general.

LLM-as-judge for evaluation

Applying multi-dimensional frameworks at scale faces a practical constraint: human expert evaluation is time-consuming and resource-intensive. The LLM-as-judge paradigm offers a potential solution by enabling consistent, automated evaluation at scale. Zheng et al. (2023) demonstrated that GPT-4 achieves over 80% agreement with human preferences on MT-Bench, a multi-turn chatbot evaluation benchmark, while the G-Eval framework (Liu et al., 2023) incorporated chain-of-thought reasoning for improved alignment with human judgments.

However, systematic biases may limit reliability. Studies have identified position bias, verbosity bias, and self-enhancement bias (Zheng et al., 2023), while other work demonstrated that swapping response order can flip evaluation outcomes, revealing sensitivity to presentation rather than content quality (Wang et al., 2023).

Critically, LLM-as-judge has not been systematically tested for reliability in technical domains like BIM-QA, where answers require verification against structured building data. Its applicability to such domain-specific evaluation remains an open question.

Inter-rater agreement for evaluation validation

Validating evaluation frameworks requires quantifying agreement between evaluators. Krippendorff's α is recommended for computational linguistics tasks: unlike Cohen's κ, it handles missing data and multiple raters without assuming rater-specific marginal distributions (Artstein and Poesio, 2008). Following established guidelines, $\alpha > 0.8$ indicates good reliability and $0.67 < \alpha < 0.8$ allows tentative conclusions (Krippendorff, 2004). Since α can be misleadingly low with imbalanced class distributions, percentage agreement provides a complementary view of rater concordance.

These four research threads converge on three unexplored questions: what evaluation criteria should be selected for granular multi-dimensional evaluation of BIM-QA systems, how reliably can different judges apply these criteria, and to what extent can LLM-as-judge replicate human expert judgments?

Evaluation framework

The evaluation framework comprises two components, both introduced in a prior study (Hellin et al., 2026). First, questions are categorized by data retrieval complexity; then answers are assessed using category-informed criteria. The question's category influences how the *Faithfulness* criterion is assessed, as detailed in the criteria definitions. We briefly describe both components below.

Question taxonomy

Most QA taxonomies categorize questions by syntactic structure or reasoning complexity (Rogers et al., 2021). This approach works well for text-based corpora (e.g., documents or web pages) where answers are stated explicitly, but BIM models contain heterogeneous information: some values are stored as explicit properties, others must be computed through aggregation or geometric operations, and some questions address data not present in the model at all. Prior work in BIM compliance checking proposed categorizing rules by data retrieval complexity (Solihin

and Eastman, 2015). Building on this, Hellin et al. (2026) adapted this classification to BIM-QA, defining four categories based on data retrieval complexity:

Category 1 (Direct Retrieval): Queries answerable through direct access to explicitly stored BIM properties without computation. *Example:* ''What is the fire rating of door D-101?'' The value is stored as an explicit property.

Category 2 (Aggregation): Queries requiring simple, deterministic operations (counting, summing, averaging) on accessible data. *Example:* ''How many windows are on the second floor?'' This requires filtering and counting, not stored directly.

Category 3 (Geometric/Spatial): Queries requiring complex computation beyond simple aggregation, where multiple valid approaches (and potentially multiple valid answers) may exist. *Example:* ''What is the gross floor area?'' This requires complex geometric computation, and different standards (ISO 9836, BOMA) may yield different valid results.

Category 4 (Incomplete Information): Queries addressing information unavailable in the BIM model, potentially answerable through explicit assumptions. *Example:* ''What is the energy consumption of the building?'' This cannot be answered deterministically and requires assumptions about occupancy, climate, and systems not modeled to provide a reasonable estimation.

Note on Category 4: This category diverges from (Solihin and Eastman, 2015)'s original Category 4 (complex simulations requiring external software). In BIM compliance checking, a BIM model containing all necessary information is considered a prerequisite, without which the process is meaningless. In contrast, for BIM-QA, users do not know in advance whether the information is available in the BIM model. Even if not definitive, an estimation or partial answer with properly disclosed assumptions may still be valuable.

Since the category depends on information accessibility, the same question might be categorized differently depending on the model being queried. For example, a floor area question could be Category 1 if the value is stored as a space property, or Category 3 if it must be computed geometrically.

This taxonomy informs how evaluation criteria, particularly *Faithfulness*, are applied, as described below.

Evaluation criteria

Answer quality is evaluated using binary judgments (Yes or No) across five criteria defined in Hellin et al. (2026). Binary judgments facilitate clear decision boundaries and reliable inter-rater agreement measurement. The N/A option applies when a criterion cannot meaningfully be assessed, either because the system abstained or because the question type does not require that criterion:

Criterion 0 (Abstention): Whether the system provided an answer or declined. If abstained, Criteria 1–4 are marked N/A since no content exists to evaluate. This criterion captures the coverage-reliability tradeoff: systems that appropriately abstain on uncertain queries avoid errors at the cost of reduced coverage, while systems that never abstain may overconfidently provide unreliable answers.

Criterion 1 (Faithfulness): Whether all claims are grounded in acceptable sources for the question's category. Acceptable grounding depends on question type:

- **Category 1:** BIM model data only: values must be directly retrieved from explicit properties
- **Category 2:** BIM model data plus deterministic computations (counting, summing, averaging)
- **Category 3:** BIM data plus complex geometric computations derived from model
- **Category 4:** BIM data plus computations plus *explicitly stated* assumptions

This category-dependent approach maintains alignment with RAG evaluation literature for Categories 1–2, while extending acceptable grounding to accommodate geometric derivations (Category 3) and incomplete information handling (Category 4). N/A only if system abstained.

Criterion 2 (Completeness): Whether the answer includes all relevant facts. N/A if the question expects only a single value with no ancillary information (e.g., ''What is the building height?''). This criterion is equivalent to recall for questions asking about a list of specific facts or items, but has a broader application.

Criterion 3 (Transparency): Whether the answer explicitly identifies sources or methods. Users must be able to verify information sources for the system to be useful in practice (Kell et al., 2024). *Example:* For ''What is the gross floor area?'', a transparent answer would state ''The gross floor area is 2,450 m², calculated by summing the areas of all IfcSpace entities with PredefinedType='FLOOR'.'' An opaque answer stating only ''2,450 m²'' would fail this criterion. This criterion is also essential for faithfulness assessment in Categories 3 and 4, since the system must explicitly reveal what assumptions were made. Transparency further enables post-hoc verification: users can trace reported values back to their source in the BIM model and confirm intermediate reasoning, complementing answer-level assessment. N/A only if system abstained.

Criterion 4 (Relevance): Whether the answer directly addresses the question asked. *Example of irrelevance:* For ''What is the fire rating of door D-101?'', an answer describing the door's dimensions, material, and manufacturer without mentioning fire rating would fail this criterion: the information provided, while accurate, does not address what was asked. N/A only if system abstained.

Validation methodology

We conducted inter-rater reliability analysis with LLM judges alongside human experts, an approach that enables simultaneous assessment of framework reliability and automated evaluation feasibility.

Dataset and evaluated system

We used 74 question-answer pairs from IFC-Bench (Hellin et al., 2025). Questions were distributed across four categories: Category 1 (13), Category 2 (39), Category 3 (5), and Category 4 (17).

The evaluated BIM-QA system is an LLM-based agentic workflow that processes natural language queries against IFC-encoded building models (Hellin et al., 2025). The system (1) receives a natural language question and the path to a BIM model, (2) dynamically generates and executes Python code to query the BIM model, (3) iteratively refines queries based on intermediate results, and (4) synthesizes a final response in natural language.

Evaluators

Human Evaluators: Two BIM domain experts, both with academic backgrounds in the AEC domain and over five years of practical experience with BIM workflows, independently evaluated all 74 responses. Both followed documented evaluation guidelines.

LLM Judges: To assess whether LLM-as-judge findings generalize across models, we employed two LLM judges from different providers:

- GLM-4.7, an open-weight ~400B-parameter model by Zhipu AI and successor to GLM-4.5 (GLM-4.5 Team et al., 2025), accessed via the Zhipu AI platform (z.ai)
- Gemini 3 Pro[1], accessed via Google AI Studio API

Both LLMs received identical structured prompts containing: (1) the question, question category, ground truth answer, and system-generated answer; (2) criterion definitions with category-dependent *Faithfulness* rules specifying acceptable grounding sources; (3) output format requiring Yes, No or N/A ratings with mandatory justification. The justification field is intended as a human-inspectable audit trail of the judge's decision rather than as a chain-of-thought elicitation mechanism. The *Faithfulness* criterion is operationalized with category-specific numerical accuracy rules: for Categories 1–2 (deterministic retrieval and aggregation), numerical values must match the ground truth within 2% for continuous measurements and exactly for discrete counts; for Categories 3–4, where multiple valid approaches may exist, the focus shifts to whether the methodology is explicitly disclosed and values are plausible. Full evaluation prompts, scripts, and rating data are available in the project repository (see below).

Evaluation Inputs: Both human and LLM judges received identical inputs for each question: the question text, question category, ground truth answer, and system-generated answer, along with criterion definitions and evaluation guidelines. This design isolates evaluation consistency from verification capability: all evaluators assessed system answers against the same reference standard. Ground truth answers were validated against the BIM models prior to the study to ensure dataset reliability. Evaluation prompts, scripts, and rating data are available at `https://github.com/sylvainHellin/ec3-2026-hellin-fuchs-nousias-borrmann`.

Human Evaluation Protocol: Both human evaluators worked independently and were blind to each other's ratings throughout the study. Evaluators reviewed criterion definitions and evaluation guidelines before beginning but did not complete a formal calibration session. This absence was intentional: in practice, different research groups cannot calibrate with each other before evaluating BIM-QA systems, so the framework must be self-contained enough to produce consistent results without joint training. Evaluators had access to BIM models for reference if needed, though evaluation primarily involved comparing system answers against provided ground truth. Human evaluators averaged approximately three minutes per question-answer pair.

Statistical methods

We measure inter-rater agreement using Krippendorff's α, which handles the N/A responses that arise when criteria do not apply to certain question types. We report both α and percentage agreement, as high percentage agreement can occur with low α when class distributions are highly imbalanced (ceiling effects).

Correlation between criteria is measured using Spearman's ρ to identify redundancy and independence patterns.

Consensus evaluation

After completing the blind evaluation, both human evaluators jointly reviewed all 20 questions where their ratings disagreed on at least one criterion. For each disagreement, they discussed their reasoning against the evaluation guidelines and independently decided whether to revise their original rating. This process resolved 18 of 20 disagreements. The two remaining disagreements, Q316 (whether omitting a window name makes an answer incomplete) and Q472 (how to classify a response where the system provided partial data but stated it could not answer, a ''soft abstention'' case), reflect genuine definitional boundaries not fully specified by the current guidelines. The consolidated expert opinion serves as a calibrated human baseline for assessing LLM surrogate suitability, while all original blind evaluation metrics remain unchanged.

Cross-system discriminative analysis

In addition to the answers from the agentic QA system described above, which were evaluated by both human and LLM judges, we generated two additional answer sets using different QA systems. These additional answers were evaluated by GLM-4.7 only. Comparing criterion scores across these three BIM-QA systems reveals which criteria best differentiate systems of varying capability.

- **Agentic QA**: The LLM-based agentic workflow from prior work (Hellin et al., 2025) using GLM-4.7, which dynamically explores BIM models via Python

[1] `https://deepmind.google/technologies/gemini/pro/`

code execution.

- **Static Baseline**: Same LLM with static context: a fixed summary of model elements provided without dynamic exploration capability.
- **Lightweight Agentic**: Same agentic architecture using Gemini 2.5 Flash Lite, a smaller model.

For each criterion, we computed the cross-system range (maximum minus minimum ''Yes'' rate) as a measure of discriminative power.

Results

Cross-system discriminative power

Table 1 presents criterion scores across three systems. *Transparency* (57pp range), *Completeness* (56pp), and *Faithfulness* (50pp) best discriminate between systems; *Abstention* (41pp) and *Relevance* (38pp) show weakest discrimination.

Table 1: Criterion scores (% Yes) across BIM-QA systems

Criterion	Agentic	Static	Light	Range
Abstention†	15%	52%	56%	41pp
Faithfulness	72%	28%	22%	50pp
Completeness	73%	17%	24%	56pp
Transparency	75%	31%	18%	57pp
Relevance	76%	42%	38%	38pp

†*For Abstention, lower is better (fewer abstentions = more answers attempted).*

Figure 1 shows the full rating distributions across systems, revealing that the agentic system's advantage holds across all criteria.

Correlation analysis

Correlation analysis reveals that *Completeness* and *Relevance* are highly correlated for both LLM judges (GLM: ρ = 0.81, Gemini: ρ = 0.76) and moderately correlated for human evaluators (H1: ρ = 0.40, H2: ρ = 0.65). This suggests potential criterion redundancy: a complete answer is almost always relevant. *Faithfulness* shows the lowest median correlations with other criteria (ρ = 0.26–0.65 across raters), though with considerable variability between evaluators, suggesting it captures a largely distinct quality dimension.

Both LLM judges exhibit stronger inter-criteria correlations than human evaluators, with positive ratings on one criterion predicting positive ratings on others. Full inter-criteria correlation matrices for each evaluator are available in the project repository.

Inter-rater agreement

Krippendorff's α ranges from 0.51 to 0.76 across criteria for the four-rater analysis (both human experts and both LLM judges), indicating moderate to substantial agreement (Figure 2, top). LLM-LLM agreement (α = 0.70–1.00) consistently exceeds human-human agreement (α = 0.32–0.57), with perfect agreement on *Abstention* and *Relevance*.

Since α can be misleadingly low when class distributions are highly imbalanced, we complement it with raw percentage agreement (Figure 2, bottom). Percentage agreement across all four evaluators exceeds 77% for all criteria, with *Abstention* (94.6%) and *Transparency* (93.2%) achieving highest concordance. LLM-LLM agreement reaches 100% for *Abstention* and *Relevance*. The combination of high percentage agreement with moderate α values reflects ceiling effects: when most responses receive positive ratings, expected chance agreement is high, limiting α's discriminative power.

Surrogate validation

The post-hoc consensus process (see Consensus Evaluation) yielded near-perfect human agreement (α = 0.88–1.00), confirming that the original blind disagreements were interpretation-driven rather than reflecting fundamental framework deficiencies. With this calibrated baseline established, we assess LLM surrogate suitability by comparing consensus-LLM agreement against original human-human agreement.

Both LLM judges align with expert consensus at α = 0.48–0.85, though their per-criterion profiles differ: GLM shows stronger *Faithfulness* alignment (H_{cons}-LLM_1: *Abstention* 0.85, *Faithfulness* 0.69, *Completeness* 0.48, *Transparency* 0.63, *Relevance* 0.78), while Gemini aligns better on *Transparency* (H_{cons}-LLM_2: *Abstention* 0.85, *Faithfulness* 0.48, *Completeness* 0.56, *Transparency* 0.73, *Relevance* 0.78). Both exceed the original uncalibrated human-human agreement (α = 0.32–0.57) for every criterion, as visible in Figure 2 (rightmost columns, top; rightmost bars, bottom).

This pattern supports the surrogate argument: LLMs not only agree with each other, but align with calibrated expert consensus at levels exceeding uncalibrated human-human agreement. For practical benchmarking, where individual calibration across research groups is infeasible, LLM judges offer a more reliable approximation of expert evaluation.

Discussion

Our results reveal a notable pattern: LLM judges achieve higher reliability than human experts, and generally align with each individual human at levels comparable to or exceeding human-human agreement.

The four-rater agreement (α = 0.51–0.76) falls within the ''tentative conclusions'' range, below the $\alpha > 0.8$ threshold for ''good reliability'' (Krippendorff, 2004). Our results therefore provide preliminary evidence rather than full validation. A full validation would require larger samples, particularly for Category 3 (n=5), and consistently higher agreement. The findings nonetheless establish a foundation: the framework produces measurable inter-rater agreement, and category-dependent patterns offer actionable guidance.

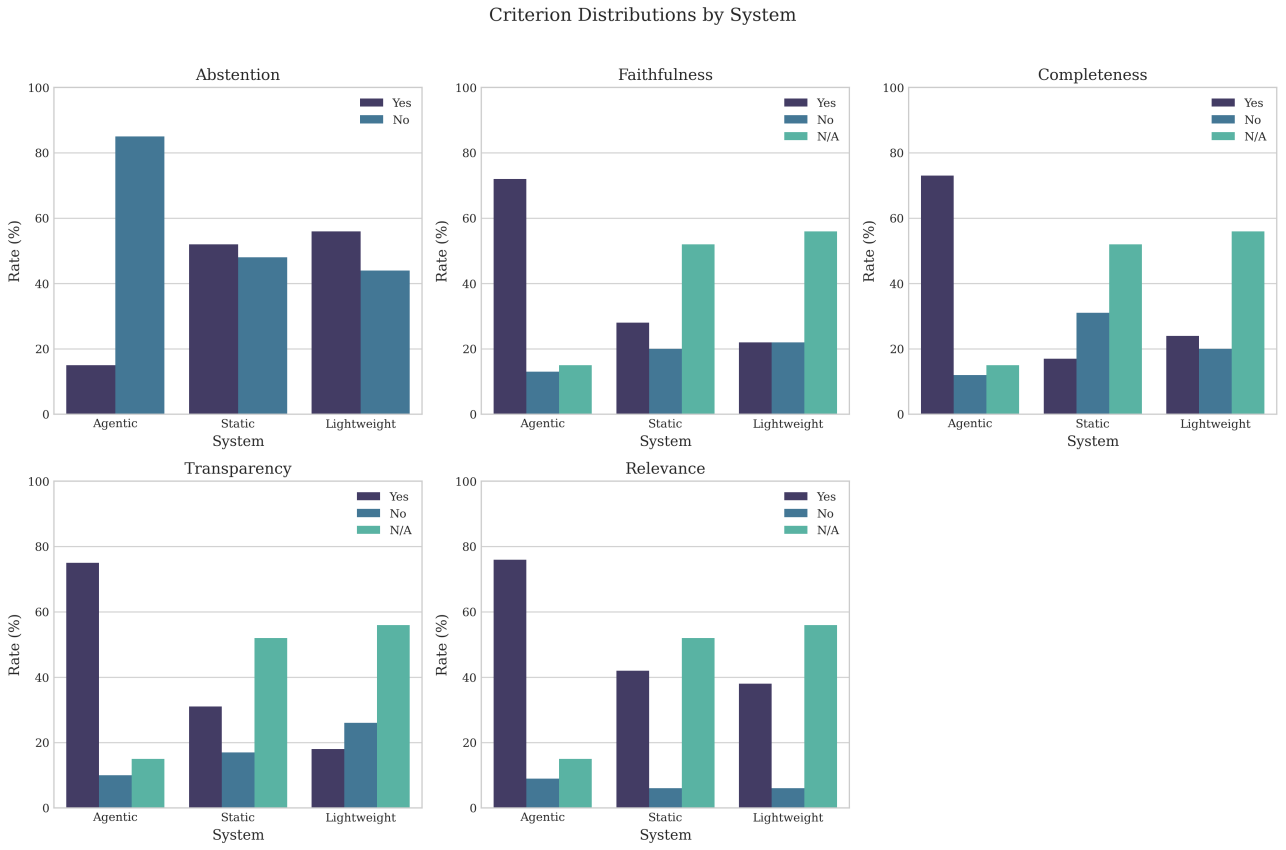

Figure 1: Criterion rating distributions across three BIM-QA systems. The agentic system achieves highest positive rates across all criteria. Higher abstention rates in Static and Lightweight systems produce corresponding N/A rates for other criteria.

Understanding evaluator differences

Why do LLM judges achieve such high reliability (α = 0.70–1.00) while human experts show lower agreement (α = 0.32–0.57)? We hypothesize three contributing factors: (1) *Criterion application consistency*: LLMs apply criteria exactly as specified, while human evaluators may interpret criteria through tacit domain expectations not captured in written guidelines; (2) *Engagement and attention*: evaluating 74 pairs against detailed criteria is repetitive work where human attention may waver, while LLMs process each instance uniformly; (3) *Criterion interpretation*: for *Completeness*, evaluators differed on whether supplementary context was required; for *Faithfulness*, they applied different verification strategies. The absence of a formal calibration session likely amplified these effects. While the criteria themselves, or the briefing provided to evaluators, may carry some inherent ambiguity, this is unlikely to be the whole story: as the post-hoc consensus analysis confirms, brief interpretive alignment is sufficient for the same raters to reach near-perfect agreement on the same criteria.

For example, Human 1's *Faithfulness* ratings exhibited higher correlation with LLM judges (α = 0.73 for H1-GLM vs. α = 0.42 for H1-H2), suggesting systematic differences in evaluation approach. Consider Question 153: ''What is the total length of exterior walls?'' The system answered ''877.47 meters,'' citing its methodology (summing the Length property for walls where IsExternal=True); the ground truth was 267.45 meters. Human 1 rated *Faithfulness* as No, recognizing the numerical discrepancy. Human 2 rated Yes, focusing on whether the methodology was grounded in BIM data. Both LLM judges also rated Yes. This illustrates a fundamental interpretation difference: does *Faithfulness* assess factual correctness, or grounded methodology? The higher LLM reliability on *Faithfulness* may reflect more uniform application of the category-specific numerical accuracy rules (see Validation Methodology).

Practical implications for BIM-QA evaluation

Our results provide preliminary evidence for LLM-as-judge suitability in domain-specific technical evaluation, extending prior findings from general-domain tasks (Zheng et al., 2023). The high LLM-LLM agreement (α = 1.00 for *Abstention* and *Relevance*, 0.89 for *Transparency*, 0.77 for *Completeness*) suggests that automated evaluation offers strong reproducibility for most criteria, with *Faithfulness* (α = 0.70), the most semantically nuanced criterion, showing the lowest but still acceptable reliability.

Two complementary findings position LLM-as-judge as a practical solution for BIM-QA evaluation. First, LLMs achieve high inter-rater reliability across model families. Second, LLMs align with calibrated expert consensus better than uncalibrated humans agreed with each other. To-

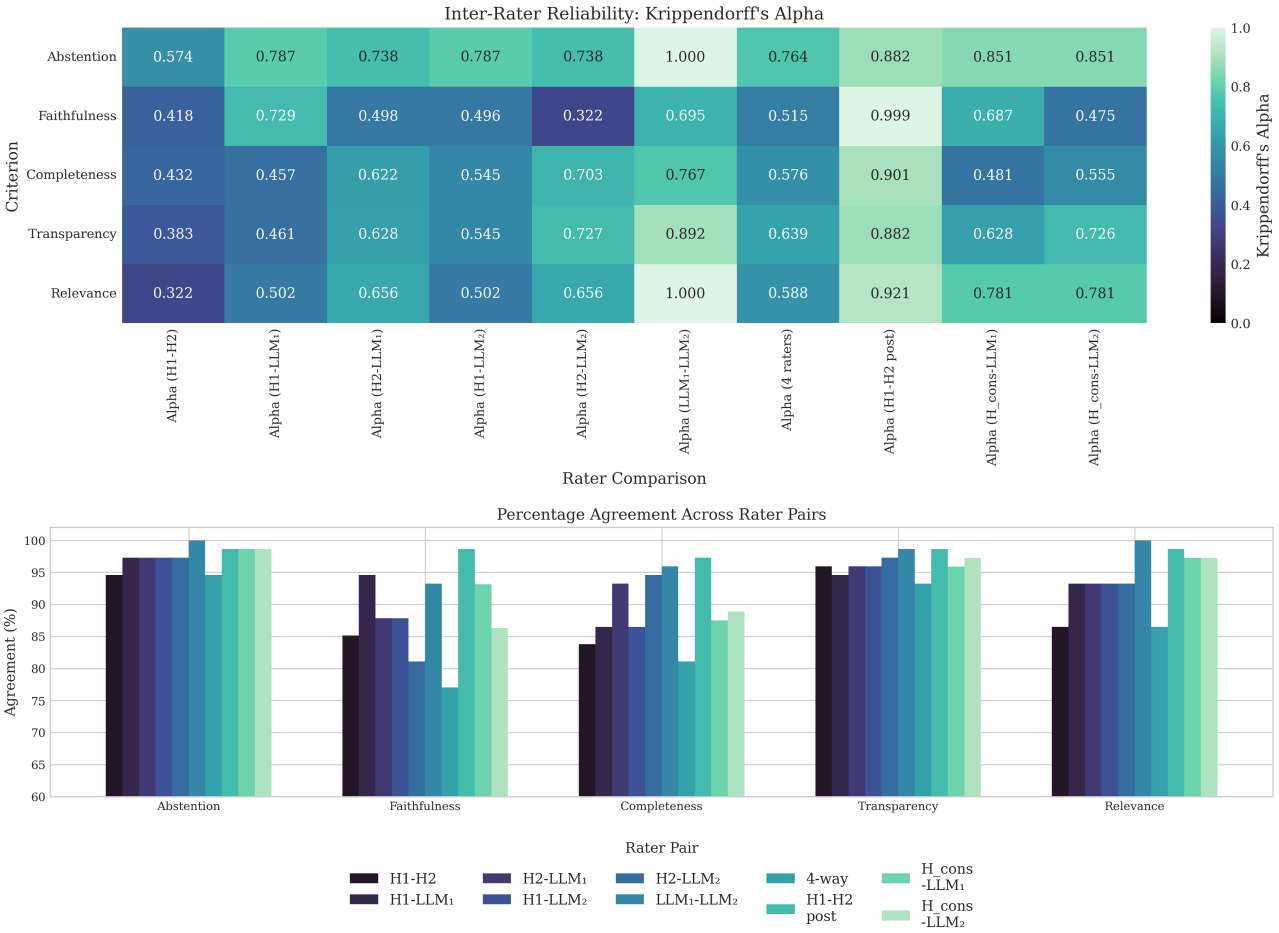

Figure 2: Inter-rater agreement analysis. ***Top:*** *Krippendorff's α heatmap showing agreement between rater pairs (H1, H2: human experts; LLM_1: GLM-4.7; LLM_2: Gemini 3 Pro) across criteria. Higher values (darker) indicate stronger agreement. Human-LLM agreement generally equals or exceeds human-human agreement. Rightmost columns show post-hoc consensus results (see Surrogate Validation).* ***Bottom:*** *Percentage agreement across rater pairs by criterion. LLM-LLM agreement consistently exceeds or matches human-human agreement, with perfect agreement on Abstention and Relevance. Rightmost bars show post-hoc consensus results.*

gether, these findings address the fragmentation problem: different research groups can compare BIM-QA systems using automated evaluation that not only produces reproducible results, but approximates the judgment of calibrated domain experts.

The correlation analysis also suggests framework simplification: given *Relevance*'s high correlation with *Completeness* (ρ = 0.76–0.81), ceiling effects (100% LLM-LLM agreement), and weak discriminative power (38pp range), we recommend consolidating *Relevance* with *Completeness* in future iterations, reducing the framework from five criteria to four.

Limitations and framework refinements

Several limitations constrain interpretation. First, both LLM judges are frontier models; findings may not generalize to smaller models. Second, all questions target IFC-based building models, limiting generalization to other BIM formats or structured-data domains. We hypothesize that LLM judges may benefit from IFC's well-documented presence in pretraining data; generalization to proprietary schemas remains to be tested. Third, the absence of calibration was intentional (to test framework self-sufficiency); the post-hoc consensus analysis revealed two definitional gaps (soft abstention handling, completeness boundaries) that enhanced guidelines should address. Fourth, with only two human evaluators, we cannot estimate population-level variance in human agreement. A replication study with balanced category sampling, the four retained criteria following *Relevance* consolidation, and a larger calibrated rater pool is needed to move from preliminary evidence to full validation. Fifth, while three systems were evaluated for discriminative analysis, inter-rater analysis used only one system's outputs. The agreement claim concerns the evaluation protocol rather than any specific system, and should hold across BIM-QA systems given consistent criteria operationalization and equal information access. Sixth, both human evaluators are co-authors of the study, which may introduce interpretation bias despite independent evaluation protocols and blinding to each other's ratings. Finally, all judges evaluated the system's reported answer against the ground truth rather than independently re-deriving it from the BIM model; intermediate reasoning verification is a property of the

system under test (supported by *Transparency* at the user level) and lies outside the protocol's scope. The study therefore assessed evaluation *reliability* (agreement across raters given the same ground truth) rather than evaluation *validity* (correctness of judgments for novel questions without ground truth); the latter remains an important direction for future work.

Conclusions

This paper introduced an automated LLM-based evaluation pipeline for BIM-QA systems and validated the underlying multi-dimensional evaluation framework through inter-rater agreement analysis with two human experts and two LLM judges. The validation reveals two complementary findings. First, LLM judges achieve high inter-rater reliability (α = 0.70–1.00), exceeding uncalibrated human agreement (α = 0.32–0.57) across all criteria. Second, a post-hoc consensus analysis demonstrates that LLMs align with calibrated expert consensus (α = 0.48–0.85) at levels exceeding the original uncalibrated human-human agreement, confirming that LLMs approximate expert judgment, not merely agree with each other. Together, these findings position LLM-as-judge as a suitable surrogate for scalable, reproducible BIM-QA benchmarking, addressing the field's fragmentation problem.

Cross-system analysis confirms discriminative validity, while correlation and ceiling effects in *Relevance* suggest consolidation with *Completeness*. Two disagreements that survived consensus discussion highlight definitional boundaries (soft abstention handling and completeness thresholds) that future guideline refinements should address.

Acknowledgments

The authors gratefully acknowledge the support of the Leonhard Obermeyer Center and the TUM Georg Nemetschek Institute.

References

Adlakha, V., BehnamGhader, P., Lu, X. H., Meade, N., and Reddy, S. (2024). Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. Transactions of the Association for Computational Linguistics, 12:681–699.

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. Computational Linguistics, 34(4):555–596.

Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics .

GLM-4.5 Team, Zeng, A., Lv, X., Zheng, Q., Hou, Z., Chen, B. et al. (2025). GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models.

Guo, D., Onstein, E., and Rosa, A. D. L. (2020). An Approach of Automatic SPARQL Generation for BIM Data Extraction. Applied Sciences, 10(24):8794.

Hellin, S., Nousias, S., and Borrmann, A. (2025). Natural Language Information Retrieval from BIM Models: An LLM-Based Agentic Workflow Approach. In Proceedings of the 2025 European Conference on Computing in Construction.

Hellin, S., Nousias, S., and Borrmann, A. (2026). A systematic evaluation framework for ai-driven bim question answering systems. In Proc. of the TUM GNI International Symposium on Artificial Intelligence for the Built World.

Kell, G., Roberts, A., Umansky, S., Qian, L., Ferrari, D., Soboczenski, F. et al. (2024). Question answering systems for health professionals at the point of care—a systematic review. Journal of the American Medical Informatics Association, 31(4):1009–1024.

Krippendorff, K. (2004). Reliability in Content Analysis.: Some Common Misconceptions and Recommendations. Human Communication Research, 30(3):411–433.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment.

Rogers, A., Gardner, M., and Augenstein, I. (2021). QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension.

Solihin, W. and Eastman, C. (2015). Classification of rules for automated BIM rule checking development. Automation in Construction, 53:69–82.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B. et al. (2023). Large Language Models are not Fair Evaluators.

Xu, F., Song, Y., Iyyer, M., and Choi, E. (2023). A Critical Evaluation of Evaluations for Long-form Question Answering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.

Yin, M., Tang, L., Webster, C., Li, J., Li, H., Wu, Z. et al. (2023). Two-stage Text-to-BIMQL semantic parsing for building information model extraction using graph neural networks. Automation in Construction, 152:104902.

Zheng, J. and Fischer, M. (2023). Dynamic prompt-based virtual assistant framework for BIM information search. Automation in Construction, 155:105067.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y. et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.