IFC-QA: A Benchmark for Evaluating LLM Reliability over Industrial Building Information Models

Anastasia Volkova [1 2] **Andrey Gromyko** [2 3] **Anton Gusarov** [2] **Evgeniy Maslov** [2] **Valentin Khrulkov** [2]
Andrey Kuznetsov [1 2] **Ivan Oseledets** [4]

[1]*Sberbank of Russia* [2]*FusionBrain Lab, AXXX* [3]*Stanford University* [4]*AXXX*. Correspondence to: Anastasia Volkova a.volkova@fusionbrainlab.com.

1. Abstract

Industry Foundation Classes (IFC) constitute one of the largest deployed domain ontologies in engineering, encoding building knowledge through over 800 entity types with typed attributes and reified relational predicates. Despite growing interest in applying large language models (LLMs) to enable natural language access to building information models, no benchmark systematically characterizes which aspects of IFC structure pose challenges for automated reasoning. We introduce IFC-QA, a benchmark of 99 natural language questions organized into five levels of reasoning complexity — from atomic fact retrieval to multi-hop relational inference and constraint checking — each paired with model-independent ground-truth extraction scripts that ensure reproducible evaluation across any conforming IFC file. We evaluate three answering architectures (neural, symbolic, and hybrid) over three structurally diverse building models, yielding 297 question–answer pairs. The hybrid symbolic–neural pipeline achieves the highest accuracy (69.0%), but performance degrades systematically with ontological depth. Error analysis reveals that 81% of failures originate in semantic concept grounding and relationship traversal rather than query syntax, identifying the natural-language-to-ontology interface as the primary bottleneck for AI-driven building information access.

2. Introduction

Industry Foundation Classes (IFC), standardized as ISO 16739 [1], define a domain ontology of over 800 entity types for building information modeling (BIM). IFC encodes architectural knowledge through typed entities, attributes, and reified relational predicates (e.g., `IfcRelContainedInSpatialStructure`), making it one of the largest deployed domain ontologies in engineering practice. Despite this formal richness, extracting information from IFC models remains technically demanding, typically requiring specialist software or programmatic querying.

2.1 Related work

Recent work applies large language models (LLMs) to enable natural language access to BIM data, including graph-based retrieval [2], ontology-guided query generation [3], code-generating agents [4, 5], and multi-agent workflows [6]. However, evaluations in these studies are typically tied to a single IFC model and lack systematic characterization of reasoning complexity. It remains unclear which aspects of IFC structure—instance lookup, aggregation, relational traversal, or constraint satisfaction—pose systematic challenges.

2.2 Figures and tables

We introduce **IFC-QA**, a benchmark of 99 natural language questions over IFC building models, organized into five levels of reasoning complexity and paired with model-independent ground-truth extraction procedures. Evaluation of three answering architectures—neural, symbolic, and hybrid—reveals that performance degrades systematically with ontological depth, and that errors concentrate in semantic concept grounding and multi-hop relationship traversal rather than in query formulation (Fig. 1).

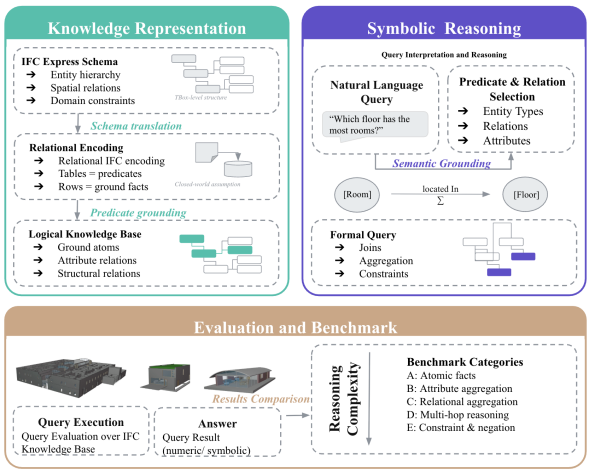

Fig. 1: Overview of the proposed IFC-QA framework. IFC EXPRESS schemas are translated into a relational knowledge base (tables as predicates, rows as facts). Natural language queries are semantically grounded into formal queries with joins, aggregation, and constraints, and evaluated over the IFC knowledge base using benchmark categories of increasing reasoning complexity.

Overview of the proposed IFC-QA framework. IFC EXPRESS schemas are translated into a relational knowledge base (tables as predicates, rows as facts). Natural language queries are semantically grounded into formal queries with joins, aggregation, and constraints, and evaluated over the IFC knowledge base using benchmark categories of increasing reasoning complexity.

3. The IFC-QA Benchmark

IFC-QA comprises 99 questions partitioned into five categories of increasing reasoning complexity: **A (Factual, 14 q.)**: single-predicate fact retrieval, e.g., "How many doors are in the building?"; **B (Attribute aggregation, 26 q.)**: aggregation over intrinsic properties, e.g., "Total floor area of spaces?"; **C (Relational aggregation, 42 q.)**: aggregation requiring explicit relation traversal, e.g., "Total area of curtain walls?"; **D (Multi-hop, 13 q.)**: composition of multiple relational predicates, e.g., "Rooms without windows?"; **E (Constraint, 4 q.)**: threshold or violation detection, e.g., "Spaces exceeding 50 m^2?"

A central design principle is *model-independent ground truth*: each question is paired with a deterministic Python script using `IfcOpenShell` that computes the correct answer from any conforming IFC file, without manual annotation. This decouples evaluation from particular building instances and ensures reproducibility. We evaluate on three structurally diverse open-source models provided by buildingSMART—Clinic Architectural and Duplex A (IFC2x3), and SampleHouse4 (IFC4)—yielding 297 question–answer pairs.

4. Baseline Evaluation

Each IFC model is materialized into a relational SQLite database, mapping entity types to tables and instantiated entities to rows. Three pipelines translate questions to SQL: **S1 (LLM Agent)** performs schema discovery, SQL generation, and iterative repair (up to k=5 attempts) using Qwen3-235B-A22B. **S2 (Symbolic)** selects deterministic SQL templates via keyword-based intent extraction, requiring no language model. **S3 (Hybrid)** applies symbolic resolution first; on failure, falls back to the LLM agent.

Category	S1 (LLM)	S2 (Symb.)	S3 (Hybrid)
A: Factual	52.4%	83.3%	**88.1%**
B: Attr. aggregation	23.1%	59.0%	**82.1%**
C: Rel. aggregation	14.3%	48.4%	**65.1%**
D: Multi-hop	28.2%	23.1%	**30.8%**
E: Constraint	16.7%	75.0%	**83.3%**
Overall	23.9%	53.9%	**69.0%**

Table 1: Accuracy (%) by reasoning category, aggregated across three IFC models. The hybrid pipeline (S3) achieves the highest accuracy; performance degrades with ontological complexity.

Table 1 shows accuracy by reasoning category. The hybrid pipeline (S3) achieves the highest overall accuracy at 69.0%, demonstrating the complementary strengths of symbolic precision and LLM flexibility. Performance degrades systematically from Category A (88.1%) to D (30.8%), mirroring increasing demands on the representation layer. The pure LLM agent (S1) occasionally outperforms symbolic templates on Category D (28.2% vs. 23.1%), as it can discover traversal paths outside template coverage—partly because IFC's extensible `IfcPropertySet` mechanism introduces author-specific attributes that templates cannot anticipate.

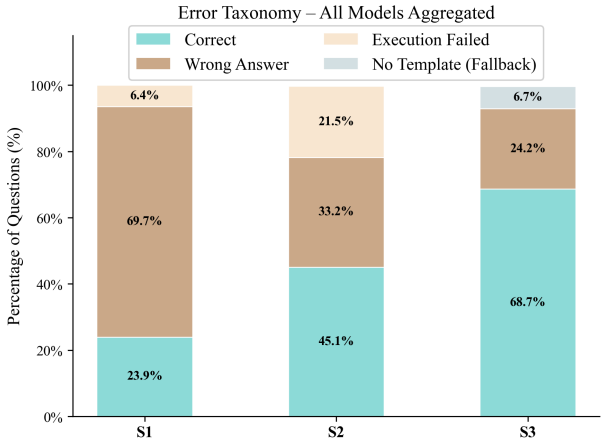

Fig. 2: Distribution of error types across the three answering pipelines. Semantic grounding errors (incorrect mapping of natural language terms to IFC predicates) and relationship traversal errors (failure to navigate multi-hop relational paths) dominate, indicating that the primary reasoning bottleneck lies at the natural-language-to-ontology interface.

Error analysis on Figure 2 reveals that 81% of failures stem from two sources: *semantic grounding errors* (incorrect mapping of natural language terms to IFC predicates) and *relationship traversal errors* (failure to navigate multi-hop paths in the entity graph). Query syntax errors are comparatively rare, indicating that the primary bottleneck is the natural-language-to-ontology interface. The symbolic-first strategy also yields substantial efficiency gains: average query time drops by four orders of magnitude (164.11 s for S1 vs. 0.01 s for the S3 symbolic path), while execution success rate improves from 93.6% to 99.0%.

5. Conclusion

IFC-QA provides a systematic, model-independent benchmark for evaluating natural language access to building information models across five levels of reasoning complexity. Our baseline evaluation demonstrates that reasoning performance correlates with ontological structure rather than query syntax, and that hybrid symbolic–neural architectures offer the best trade-off between accuracy, interpretability, and efficiency. The concentration of errors in semantic grounding and relationship traversal motivates richer knowledge representations—graph-based encodings, compositional rules, and more explicit logical structuring—to support natural language access to domain-specific engineering ontologies. IFC-QA and all evaluation code are publicly available.

References

[1] Mikael Laakso and Arto Kiviniemi. The IFC standard – a review of history, development, and standardization. *ITcon*, 17:134–161, 2012.

[2] Saeed Iranmanesh, Hesham Saadany, and Edlira Vakaj. LLM-assisted Graph-RAG information extraction from IFC data. In *Proceedings of the European Conference on Computing in Construction (EC3 2025)*, 2025.

[3] Meng Yin, Llewellyn Tang, Craig Webster, Shuo Xu, Xiao Li, and Hao Ying. An ontology-aided, natural language-based approach for multi-constraint BIM model querying. *Journal of Building Engineering*, 76:107066, 2023.

[4] Bin Liu and Hao Chen. BIMCoder: A comprehensive large language model fusion framework for natural language-based BIM information retrieval. *Applied Sciences*, 15(14):7647, 2025.

[5] Simon Hellin, Stavros Nousias, and Andre Borrmann. Natural language information retrieval from BIM models: An LLM-based agentic workflow approach. In *Proceedings of CIB W78 2025*, 2025.

[6] Dong Liu, Xiang Zhou, and Yiyu Li. Enhancing natural language retrieval of BIM data through integration of large language models with multi-agent systems. In *Proceedings of CAADRIA 2025*, 2025.