2025 European Conference on Computing in Construction

CIB W78 Conference on IT in Construction

Porto, Portugal
July 14-17, 2025

A GRAPH-BASED METHOD FOR GENERATING Q&A DATASETS FOR IFC INFORMATION RETRIEVAL

João M. Silva[1], Luis Jacques[1,2], Luis Sanhudo[1,2], João Poças Martins[2], and Rosaldo Rossetti[3]
[1]Built CoLAB, Porto, Portugal
[2]CONSTRUCT-GEQUALTEC, FEUP, University of Porto, Porto, Portugal
[3]LIACC, FEUP, University of Porto, Porto, Portugal

Introduction

Information retrieval (IR) in BIM is a critical research area for the AEC industry, particularly as projects grow in scale and complexity. The ability to respond to precise, context-specific queries on BIM supports decision-making and many downstream tasks for stakeholders (Jiang, 2024). Traditionally, retrieval has relied on using query languages within the backend of BIM applications (Mazairac, 2013), limiting access to predefined searches or users with deep technical understanding of the data schema. However, advances in Large Language Models (LLMs) are now enabling a shift toward natural language interfaces, making the process more accessible to end users (Zheng, 2023).

However, this shift exposes the need for a reliable method to evaluate the quality of such systems. The lack of a benchmark for LLM-based BIM IR has made it impossible to accurately compare different approaches. This is due to Q&A pairs being created manually (Mengtian, 2023a), making them a scarce resource; and due to the few publicly available datasets being incomplete – either missing ground truth answers or lacking the corresponding IFC models.

To address this, we introduce a new method that can automatically extract Q&A datasets from IFC, enabling the evaluation of IR performance across any input model. This flexibility is a key strength of our approach, allowing for crowdsourced dataset expansion: peers can use our algorithm on their files and contribute new Q&A. By covering different disciplines and typologies, we enhance semantic coverage and strengthen the benchmark's robustness.

Methodology

We conducted preliminary research by collecting Q&A pairs from public datasets (Wang, 2021; Mengtian, 2023b, 2023c, 2024) and classifying them by the scope of information needed to answer each query:

a) Property-based: focused on class properties or filtering elements by property values.
b) Class-based: involving comparative tasks such as ranking, counting, or averaging within element types.
c) Relation-based: involving relationships between objects, such as connectivity, inclusion, or shared attributes.

Table 1: Dataset sample for the three Q&A categories

Question	Answer	Type
How many IfcSlabType have a Function of Interior?	24	a)
What is the material of the plate with GlobalId "26T…"?	Glass	a)
How many IfcSlab have a Type Name of Generic 150mm?	3	b)
What are the IfcCurtainWall elements with a Base Offset value equal to 0.0?	['2D_O…', (…)]	b)
Is the IfcSlab with Global ID "3qq…" contained within the IfcBuildingStorey "Level 1"?	True	c)
What is the OverallWidth of the IfcWindow "K-LINE" that is contained within the wall with GlobalId "0O8…"?	2864.0	c)

We define question generation as identifying the required information context for a ground truth from which a relevant question can be formulated. The main challenge lies in extracting this context, as the question itself is created by providing a LLM (Llama 8B) with the context and ground truth (Namboori, 2024).

For property and class-based queries, direct navigation through the IFC schema is enough to extract property values from individual elements or all elements of a specific class. We developed two algorithms (using *IfcOpenShell*) that iterate through available classes and collect questions based on the type of property being randomly sampled: quantity and statistics related

questions for numerical values, and filtering or matching-related questions for categorical.

However, relation-based queries require a more flexible approach, as illustrated in Figure 1. The class-dependent nature of relationships in the IFC schema make it impractical to automatically trace a path between two classes without relying on a case-by-case logic. To address this, instead of explicitly searching for a path, we perform shortest walks through the schema and infer a question from the extracted subgraph. It is possible to specify target classes to link and require the path to include a specific class, effectively guiding the extracted subgraph to reflect a desired semantic relationship between the target classes.

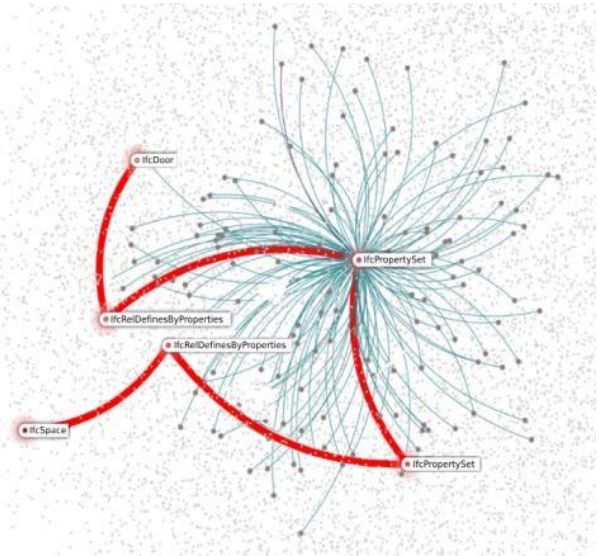

Figure 1: Shortest path for the question: "Is the IfcDoor with Global ID "04v…" contained within the IfcSpace "12"?".

Results

To assess data quality, we sample 2000 Q&A pairs created from a public IFC model (Wisén, 2024).

Figure 2 shows each query type forming a distinct, minimally overlapping cluster, indicating consistently different semantic content. Relation-based datapoints show the highest density, suggesting that using different mandatory classes along the path can improve semantic coverage.

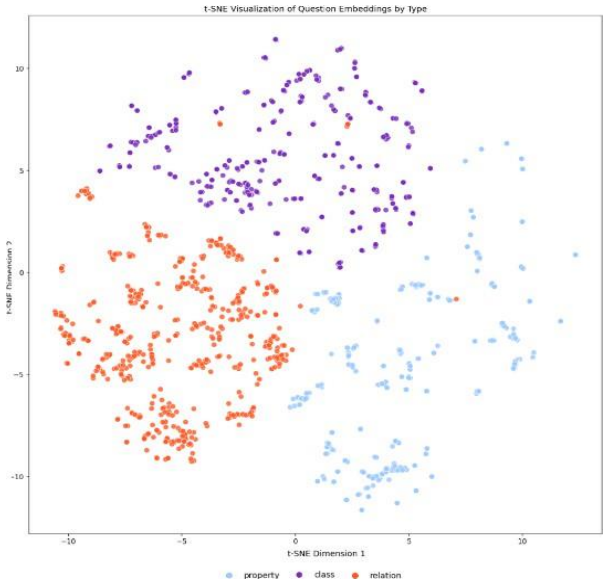

Figure 2: t-SNE visualization of question embeddings using the 'all-MiniLM-L6-v2' Sentence-BERT model.

Figure 3 shows an uneven distribution of Element classes, which can be explained by their original proportions in the model. The high frequency of links to *IfcRelContainedInSpatialStructure* and *IfcBuildingStorey* suggests a bias toward level-related class relations. We also found that generated questions often explore relationships unrelated to one or both target classes (paths average 3.98 classes in length), enhancing semantic diversity beyond what a structured approach can achieve.

Conclusions

Our approach enables scalable, automatic generation of high-quality Q&A datasets from IFC, supporting robust evaluation of LLM-based BIM IR systems. By capturing queries across property, class, and relational dimensions, and allowing contributions from practitioners using varied models, we aim to lay the foundation for a comprehensive, community-driven benchmark across many BIM use cases. While focus is on generating questions from known answers to evaluate how well LLM pipelines can retrieve information, the method is not intended to simulate how end users pose questions – an important but separate challenge.

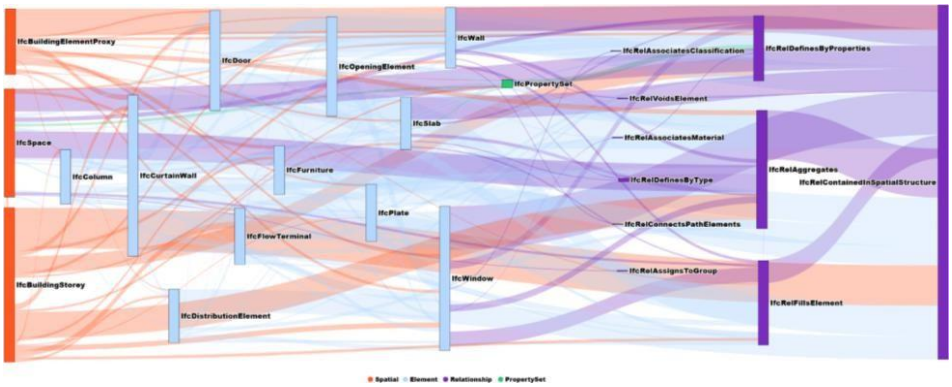

Figure 3: Co-occurrence of classes within shortest paths. Hight reflects frequency of appearance in the dataset, while links indicate classes that appear together in the same path.

Future work will focus on improving prompt engineering to increase the diversity of question phrasing and scope, further strengthening the benchmark's realism and coverage.

Acknowledgments

This work is co-funded by the PRR - Recovery and Resilience Plan and the European Union - www.recuperarportugal.gov.pt | PRR - Investment RE-C05-i02: Interface Mission – CoLAB.

References

Wisén, André. (2024). Samples IFC - Models [Data set]. Kaggle. doi.org/10.34740/KAGGLE/DSV/7951555

Jiang, S., Zhang, J., Shi, J., & Wu, Y. (2024). Adaptive information retrieval for enhanced building safety management leveraging BIM. Engineering, Construction and Architectural Management.

Mazairac, Wiet & Beetz, Jakob. (2013). BIMQL – An open query language for building information models. Advanced Engineering Informatics. 27. 444–456. 10.1016/j.aei.2013.06.001.

Zheng, J., & Fischer, M. (2023). BIM-GPT: A prompt-based virtual assistant framework for BIM information retrieval. arXiv preprint arXiv:2304.09333.

Mengtian Yin, Llewellyn Tang, Chris Webster, Shen Xu, Xiongyi Li, Huaquan Ying. (2023a). An ontology-aided, natural language-based approach for multi-constraint BIM model querying. Journal of Building Engineering, Volume 76, doi.org/10.1016/j.jobe.2023.107066

Namboori, A., Mangale, S., Rosenbaum, A., & Soltan, S. (2024). GeMQuAD: Generating multilingual question answering datasets from large language models using few shot learning. arXiv preprint arXiv:2404.09163.

Wang, Ning & Issa, Raja & Anumba, Chimay. (2021). A Framework for Intelligent Building Information Spoken Dialogue System (iBISDS).

Yin, Mengtian & Li, Haotian & Wu, Zhuoqian & Tang, Llewellyn. (2024). Information Requirement Analysis for Establishing BIM-Oriented Natural Language Interfaces.10.1007/978-981-99-7965-3_46.

Mengtian Yin, Llewellyn Tang, Chris Webster, Jinyang Li, Haotian Li, Zhuoquan Wu, Reynold C.K. Cheng. (2023b). Two-stage Text-to-BIMQL semantic parsing for building information model extraction using graph neural networks. Automation in Construction, Volume 152, https://doi.org/10.1016/j.autcon.2023.104902.

Mengtian Yin. BIM-NLQI. Github. (2023c). https://github.com/MengtianYin/BIM-NLQI/