LLM based automatic relation between cost domain descriptions and IFC objects.

Chiara Gatto, chiara.gatto@polimi.it
Department of Architecture, Environment and construction Engineering, Polytechnic of Milan, Milan, Italy

Jacopo Cassandro, jacopo.cassandro@polimi.it
Department of Architecture, Environment and construction Engineering, Polytechnic of Milan, Milan, Italy

Claudio Mirarchi, claudio.mirarchi@polimi.it
Department of Architecture, Environment and construction Engineering, Polytechnic of Milan, Milan, Italy

Alberto Pavan, alberto.pavan@polimi.it
Department of Architecture, Environment and construction Engineering, Polytechnic of Milan, Milan, Italy

Abstract

Effective cost estimation for tendering plays a critical role in the building construction process, enabling efficient investment management and ensuring successful execution of the construction phase. The current practice involves the classification of building items, extracting all the quantities of the latter, collecting pricing information from construction price list documents and manually relate these data to the building items.

The objective of this paper is to support cost estimation activity by developing a tool that automates the process of assigning a cost domain description to IFC-based BIM building objects, in such a way as to minimize the human error when manually performing this activity and speed up the process. To handle the textual dataset, the authors propose a prompt-based framework, testing Mistral-7b language model to querying cost domain descriptions with data in IFC format. This approach is applied to two domains, each characterized by different semantics.

Keywords: prompt engineer, Natural Language Processing (NLP), Large Language Models (LLM), IFC, cost estimation, Structured Query Language (SQL).

1 Introduction

Accurate cost estimation is crucial for effective decision-making in construction project management, as highlighted by numerous studies.(Sepasgozar et al., 2022).

However, traditional construction cost estimation typically involves several manual processes and limited automation, which makes it time-consuming and prone to human errors (Akanbi & Zhang, 2021). This process can be summarized in some major stages consisting of: classifying all construction products that constitute a building project into assemblies or items, extracting the all quantities of the latters (quantity take off activity), collecting pricing information from construction specification, relating this data to assemblies and items and finally computing the project (Ma et al., 2013). Although BIM approaches are becoming more prevalent, the AEC industry still primarily exchanges information through paper-based documents. Written in natural language, these documents contain unstructured or semi-structured data, posing challenges for digital management. Manual data extracting can lead to discrepancies and inaccuracies in information, thereby posing financial risks, project delays, potential failure, and in the worst case scenario to judicial disputes (Jafari et al., 2021). These issues impact the effectiveness of the projects along with the credibility/reputation of the stakeholders. Moreover, the gap between traditional document-based (i.e., semi-structured and unstructured) and model-based information can lead to information loss and inconsistency (Opitz et al., 2014). Thus, effective data management turns out to be essential to the overarching project strategy.

To address the aforementioned issues, this research proposes a framework for automatizing the procedure of relating building elements extract from an IFC file with its respective cost item coming from price list textual documents. This is achieved through the application of Natural Language Processing (NLP) techniques, testing the effectiveness of large language models (LLM) models within this specific knowledge domain and task. The methodology focuses on using LLM as an instrument for automatizing the process of relate terms coming from the cost domain, through price list textual documents, with terms coming from geometric domain, coming from IFC files, with the aim of addressing the proper price list item to the building element. The proposed research activity follows the prior study focused on automating the process of structuring data from cost related textual documents (Gatto, Farina, et al., 2023), and a further study focused on verifying the consistency and the quality provided by price list tendering document (Gatto, Gholamzadehmir, et al., 2023). Specifically, this research shifts the focus on responding to the practitioner's need to facilitate the process of addressing the proper price item from textual sources to the building element extracted from IFC-based files.

This research is organized as follows: The initial section, "State of the Art," presents the research background. Following that, the "Research goal and methodology" and "Proposed framework" sections detail the study's approach and implementation. The subsequent segments, "Case study implementation demonstrate the practical application of the framework and its evaluation. Ultimately, the "Conclusion and future development" section encapsulates the key findings from the study.

2 State of the art

Cost estimation represents a fundamental stage in the effective decision-making process of a construction project. During this stage cost, time and other resources have to be predicted, not only for effectively plan the investments necessary for the building work realization, but also for successfully managing the construction phase (Choi et al., 2015). Building Information Modelling (BIM) tools assume a fundamental role in the cost estimation process, especially in the quantity take off step.

Within this chapter two research domain areas are examined with the aim of providing gaps and opportunities related to the BIM and the NLP applied to the AECO sector.

2.1 Natural language processing

Manual extraction of reporting requirements from extensive construction documents can lead to time and cost underestimations. In this direction, the application of NLP techniques has been increasingly adopted in the AEC sector to manage the information contained in documents (Jafari et al., 2021). NLP is mainly applied in four scenarios of information extraction, document organization, expert systems, and automated compliance checking (Wu et al., 2022).

Recently, NLP has been employed in the construction industry to aid cost estimation through improved document management (Tang et al., 2022). To automate the extraction of information from construction regulatory documents, a study was conducted using a semantic rule-based NLP approach for a text recognition algorithm based on semantic analysis. (Zhang & El-Gohary, 2016). A later study developed an automated framework that employs NLP and machine learning techniques to automatically identify and prioritize key contract terms, helping managers quickly and comprehensively understand contract agreements. (Hassan & Le, 2020). Furthermore, a model that automatically identifies the most relevant pairs of provisions from various specifications using semantic text similarity was developed (Moon et al., 2021). This assists practitioners by reducing the effort to complete tasks that involve written documents, enhancing the objectivity of outcomes, and minimizing human errors.

Several studies have explored the use of NLP techniques to convert natural language queries into structured queries in order to retrieve information from IFC based BIM models. To automate this process a study developed an approach based on parsing natural language queries according to the IFC ontology. The parser works following three steps: identifying the IFC concepts, then identifying the relationship among concepts and finally, identifying the value of the IFC concepts. Finally converting this data into a query for extracting information from IFC BIM models represented in RDF format (Yin et al., 2023).

In further study the aim of retrieving information from BIM models starting from a query in natural language have been performed by testing the effectiveness of GPT language model. The authors have developed a prompt generator for interpreting users natural language queries and a query manager setting for structuring the queries used for retrieving information from external databases (Zheng & Fischer, 2023).

2.2 Retrieving information from BIMs

Building Information Models capture a vast amount of multi-disciplinary data related to various stages of a construction project supporting information exchange based on unified standards. Consequently, those are also used for searching and retrieving information from models (Isikdag et al., 2007). Among them, the Industry Foundation Classes (IFC), an ISO standard for BIM, is the most widely used one. However, the high flexibility of IFC results in complexity and difficulty in understanding, leading to challenges in extracting information. Consequently, this hinders the application of BIM in many cases, causing some operators to still prefer manual operations.

The semantic search approach leverages the object-oriented structure of BIMs. For instance, a study have been carried out with the aim of enabling querying IFC-based BIMs with SPARQL (Beetz et al., 2009), in a further study the complexity of accessing models have been solved through the conversion of IFC models into OWL schema in order to simplify the process of retrieving information through SPARQL query and graph theory (Ruiz-Morales et al., 2022).

Additionally, a further study explored a more intuitive visual programming language for filtering IFC-based BIMS (Preidel et al., 2017).

2.3 Problem statement

The state of the art reveals a growing trend in applying artificial intelligence techniques combined with standardized data exchange languages in the construction industry. However, there is evidence of a lack of application of these techniques in the cost estimation sector.

This stage is critical, as it involves meticulous planning of material, time, and other resources necessary for the successful execution of the project. It demands a high level of expertise and attention from professionals, given the numerous manual tasks involved, such as consulting textual documents to extract valuable information for work quantification. Minimizing distractions and preventing semantic misinterpretation errors during this stage is essential.

3 Research goal and methodology

The methodology approach used during this study is explained in this section and synthesized as shows.

The current practice of building cost estimation involves the classification of building items, extracting all the quantities of the latter, collecting pricing information from construction priced list documents and manually relate these data to the building items. This work aims to optimize the final tasks mentioned above, specifically the extraction process of data from textual documents and the subsequent attribution of this information to building items extracted from IFC-based BIMs. The goal is to develop a framework which automatizes the retrieval and the coupling process of data contained in external textual documents with building elements in IFC-based BIM models.

To achieve the goal, the effectiveness of large language models (LLM) was tested. In developing this article, research approach pioneered by (Zheng & Fischer, 2023) was used as a model. The developed procedure has been refined and tested on a practical case study. The methodology approach developed during this study is shown in Figure 1. Here are listed the main stages:

- preprocessing data related to building project coming from IFC-based BIMs and the data related to the cost domain coming from price lists documents stored in SQL databases;
- using prompt-engineering for establishing a semantic relation between the vocabulary coming from IFC-based BIMs and semi-structured textual documents stored in SQL database;
- extracting the specific terms to compile a SQL query;
- finally isolate only those price items compatible with the properties of the building element examined.

It's worth noting that during the refinement of the framework, the initial approach involved structuring the prompt by supplying it with building object IFC format data and a list of cost domain descriptions in natural language format, rather than organizing it within an SQL database. However, the large language model was unable to correctly assign the appropriate price item to each building element. Therefore, a more deterministic approach has been preferred by structuring the dataset in databases and using SQL queries for information retrieval, as shown in the methodology chart.

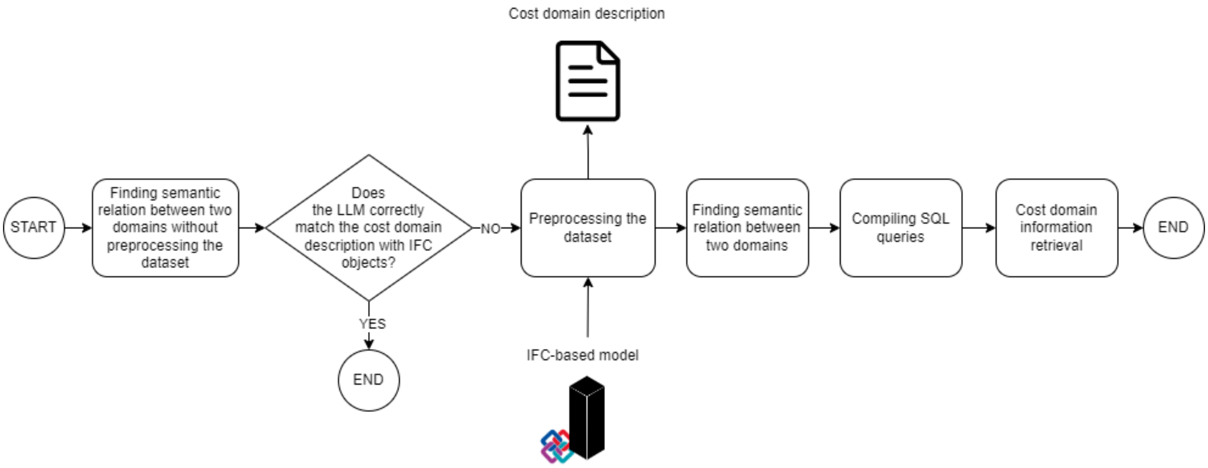

Figure 1 Methodology.

4 Proposed framework

In this section the process of the framework design and development is explained and synthesized, as Figure 2 shows. The object to be tent during this phase is to develop a procedure based on prompt engineering for improving the practitioners experience of retrieving textual information from textual documents when performing a building cost estimation.

The first step in the methodology is related to the preprocessing of the dataset, where both data related to the building project and the cost domain are structured. The aim is the one of isolate only the most relevant information from both dataset to allow their comparison.

Firstly, building elements coming from IFC-based BIMs needs to be isolated to identify IFC properties and attributes useful for characterizing the object from an economic perspective. Each object is characterized by a unique set of dimensional, geometric, performance, and physical attributes. Once these parameters have been identified, they can be isolated and extracted. The information selected at this stage is structured and collected within an SQL database, where the properties of an object are organized into columns, and each row corresponds to a unique building element. Secondly, cost items from textual documents are structured as well into an SQL database. At this stage the data are structured, however, direct comparison is not yet possible due to non-homogeneous semantics used in the two domains.

In order to allow an automatic connection between two databases the subsequent stage of the proposed study is to create a semantic relationship between the data coming from BIMs domain and cost domain. To achieve this, the capabilities of large language models are tested by structuring a prompt whose goal is to relate database fields to project building elements and the database to price items. A few shots example prompt strategy is implemented leveraging the capabilities of LLM of generalize the task on specific domain (Dai et al., 2022). In order to provide the knowledge domain to the LLM the prompt is structured into five modules:

- role definition: information regarding the role the LLM is expected to carry out;
- dataset composition: details about the content and composition of the dataset;
- task instructions: instructions on the tasks to be performed. For this study four tasks have been defined:
    - coupling the objects from the building project database to the price item list database,
    - coupling the materials from the two databases,
    - understanding if the wall has a bearing function,
    - extracting the with parameter;
- examples: few examples demonstrating how the output is expected to be generated; the example serves also for providing the instruction of the output format in which the answer should be provided;
- input provision: the input for which we need the prompt to perform the previously mentioned tasks, which consists of providing it with the building object IFC-based data.

Once the LLM provides the required output, the answer data are extracted with regular expressions in order to subsequently compile a structured SQL query.

The final stage of the framework consists of using the SQL query for retrieving the only relevant price list items from the cost domain database. If more than one price list item is suitable for the building object, the final step is for the practitioner to choose the best option.

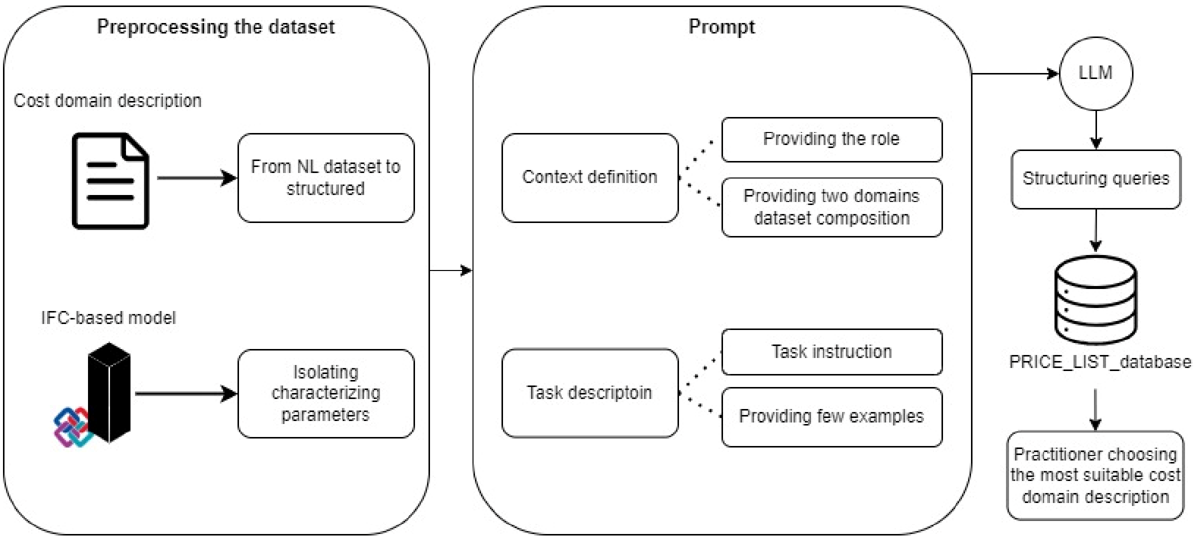

Figure 2 Framework workflow.

5 Case study implementation

In this section, the framework is tested on a case study focusing on a dataset composed of 6 walls modeled in Revit software and then exported in IFC format. The dataset preprocessing of the building elements and of the price list documents is shown, the definition of the prompt is explained and then the results are presented.

5.1 IFC-based dataset preprocessing

Focusing on the building elements extracted from BIM in IFC format, the dataset is composed of 6 typologies of walls, which varies one to the others in terms of width, material and bearing function. From the set of data set accompanying the IFC file, only those that characterize the object from an economic point of view are identified and filtered. Only the following properties "IfcBuildingElement," "IfcMaterial," and the attributes "LoadBearing," "Width" are then selected.

Table 2 present the content extracted from the IFC format for each case study. It is possible to see that the textual content is mixed in Italian language for the preset fields and in English formal language.

5.2 Cost domain descriptions preprocessing

Concerning the cost domain descriptions, the public tendering price list document of the Lombardy region was considered. The Public Italian Contracts Code imposes on each Italian region to annually provide a price list that contracting authorities must use for setting the project cost base for tenders. Therefore, each region provides practitioners with a price list containing work items and their respective cost (Sdino & Rosasco, 2021). The tool mainly stores data associated with construction activities, including their unit prices. This resource assists practitioners in generating estimated metric calculations. Additionally, to ensure more transparency in the composition of the price of construction works, the price list provides a catalog of elemental resources involved in the latter.

Therefore, there must be a full correspondence between works and elemental resources, otherwise inconsistency arises. Considering the need for annual updates, the price list is subjected to periodical revisions, consisting in the unit prices update, the addition of new work and elementary resource items or removal of outdated entries.

Table 1. Building elements under exam, masonry walls.

n° case study	Case study 1	Case study 2	Case study 3
IfcBuildingElement	IfcWall	IfcWall	IfcWall
LoadBearing	True	False	True
Width	240	120	300
IfcMaterial	Blocchi in laterizio	Blocchi in laterizio	Blocchi in laterizio

Table 2 Building elements under exam, concrete walls.

n° case study	Building element 4	Building element 5	Building element 6
IfcBuildingElement	IfcWall	IfcWall	IfcWall
LoadBearing	True	False	True
Width	240	120	300
IfcMaterial	Blocchi in calcestruzzo	Blocchi in calcestruzzo	Blocchi in calcestruzzo

Table 3 Cost domain description from price list document.

Database fields	Price item description
ID	5
extended_description	Muratura in blocchi di laterizio tipo "svizzero" portante, con malta cementizia o bastarda, compreso l'onere per la formazione di spalle, voltini, spigoli, lesene, piani di lavoro interni: con blocchi 10 x 25 x 13 cm, spessore 10 cm
object	muratura
material	laterizio
function	portante
parameter_width	spessore
walue_width	120
unitary_price	47,00
unit_of measurement	mc

Information is conveyed by the tool in verbal form: sentences composed by words and syntax delivering knowledge. Since each item is written in natural language and because the document doesn't follow a standard in providing information, a lack of homogeneity has been recorded between each item phrase structure and information typology transmitted.

Each price item has been stored in a database where the relevant information is structured and organized into columns, as presented in Table 3. Comparing the latter with Table 1 and highlights that the semantics used in the cost domain differ from those in the IFC format, making direct correlation challenging.

5.3 Structuring prompt

The prompt setting relied on the typology of LLM used. For this study, the Mistral-7b model was implemented in Google Colab environment. Mistral-7b is a 7-billion-parameter language model (Jiang et al., 2023), resulting in higher computational efficiency, lower cost, and characterized by greater accessibility. This model requires prompts to be formatted as a conversation between a user and an assistant, with the user initiating the dialogue. The structured prompt is showed in Figure 3.

Thus, the first entry to the prompt simulates the user's input, which in this case includes the sequence of information shown in Tables 1 and 2. The second entry consists of providing to the prompt:

- the role definition, highlighted in the red box,
- information regarding the dataset composition, highlighted in red box, where the semantic is provided by inserting the occurrences of the two databases in order to make the model acquire the morphology related to both of the IFC and cost domain,
- task instructions, highlighted in purple,
- the examples, highlighted in green box, demonstrate the format in which the output should be obtained; specifically, the output is presented within square brackets so that it can later be extracted using regular expressions;
- and finally, the input provision, highlighted in blue box.

5.4 Cost domain description retrieval

The final stage is based on structuring an SQL query by leveraging the output generated by LLM. Those are therefore extracted through regular expressions and subsequently used for compiling an SQL query. The structured query is then used for retrieving the price item descriptions from the PRICE_LIST_database.

A complete example that clarifies all the stages of the framework is shown in Figure 4.

6 Conclusion and future developments

The research described in this article contributes to the goal of supporting practitioners' activity in the cost estimation field. While LLMs show promise as a valuable tool, further testing is necessary to fully evaluate their effectiveness in this domain.

The primary goal was to develop a framework that automates the assignment of price items to IFC-based objects in BIMs. Initially, it was supposed to be possible without structuring the textual description dataset from the cost domain. However, in order to achieve effective results, the dataset still requires preprocessing and structuring textual information.

To automate the task of manually assigning a textual description to a building element a prompt-based approach has been tested by providing the LLM with his role, the composition of the preprocessed dataset and few examples of the expected output. The framework has been tested on a limited sample, composed of six different walls obtaining successful results for all the tests. The model effectively identified semantic matches between domains characterized by different lexicons and the presence of terms in both English and Italian language, enabling the correct composition of SQL query for assigning the appropriate price item to the building objects.

The prompt used in this study is complex and composed as a database query rather than a more natural language. However, it is important to emphasize that the LLM module presented in this work is conceived to be used in back-end process, where the prompt is not manually crafted for each query but is dynamically generated changing the input part basing on the specific characteristics of the IFC objects being analyzed.

Future work could explore the potential for integrating more natural language prompts to enhance user interaction, especially in scenarios where human input is more directly involved. Additionally, the case study analyzed will be extended to test the effectiveness of the LLM on a more significant sample encompassing a broader range of semantic terms and a wider variety of building objects. This will also enable the production of evaluation metrics, not included in this article due to the limited sample studied. Furthermore, different models beyond Mistral-7b will be tested and compared, to obtain more robust results.

```
{"role": "user","content": "['entity': 'ifcWall', 'LoadBearing': 'True','ifcMaterial': 'blocchi in laterizio', 'width': '250']"},

{"role" :"assistant", "content": "You are a virtual operator that according to building elements coming from input classify the entity according to the objects occurrency of the PRICE_LIST_database.

Dataset information: IFC_database is a database containing data coming from an IFC format file, each building element occupy a raw and it is characterised by the following attributes ["entity", "LoadBearing", "IfcMaterial", "Width"].

PRICE_LIST_database is a database containing a list of price items coming from textual document, each price item occupy a raw and it is characterised by the following attributes ["extended_description" , "object" ,"material" , "function" , "udm" ,"unitary_price" , "parameter_width" , "udm_width", "value_width"]. The objects occurrency contained into "object" are ["muratura", "solaio", "scale", "intonaco", "finestra", "tetto"]. The materials occurrency contained into "material" are ["laterizio", "cemento", "legno", "acciao", "roccia","plastica", "vetro"].

Task instruction: your first task consists of matching the "entity" occurency from IFC_database to the "object" coming from PRICE_LIST_database. Your second task consists of matching the "material" form IFC-database to the material coming from PRICE_LIST_database. Your third task consists of matching the "loadbearing" from IFC-database to the function coming from PRICE_LIST_database. Your fourth task consists of isolate the width value from IFC-database.


Task 1 answer: ["IfcWall", "muratura"]. Task 2 answer: ["blocchi in calcestruzzo" ,"calcestruzzo"]. Task 3 anwser: ["LoadBearing" :"true" correspond to "function" : "portante"]. Task 4 answer: ["Width : 250" correspond to "spessore = 250"]"},


{"role": "user", "content": "['entity': 'ifcSlab', 'LoadBearing': 'False', 'ifcMaterial': 'legno in abete', 'width': '200']"},
{"role":"assistant", "content": "Task 1 answer:["IfcSlab", "solaio"]. Task 2 answer: ["legno di abete" , "legno"]. Task 3 anwser: ["LoadBearing" :"true" corresponds to "function" :"non-portante"]. Task 4 answer: ["Width : 200" correspond to "spessore = 200"]"},
{"role": "user", "content": "["entity" : "IfcSlab", "LoadBearing": "False", "IfcMaterial":"legno", "width": "200"]"},{"role": "assistant", "content": "Task 1 answer: ["IfcSlab", "solaio"]. Task 2 answer: ["legno di abete" , "legno"]. Task 3 anwser: ["LoadBearing" : "true" corresponds to "function":"non-portante"]. Task 4 answer: ["Width : 200" correspond to "spessore = 200"]"},


{"role": "user", "content": "['entity': 'ifcWall', 'LoadBearing': 'TRUE', 'ifcMaterial': 'blocchi in calcestruzzo', 'width': '120']"}
```

Figure 3 Prompt composition: role in yellow box, dataset composition in red box, task instruction in purple box, examples in green box, input provision in blue box.

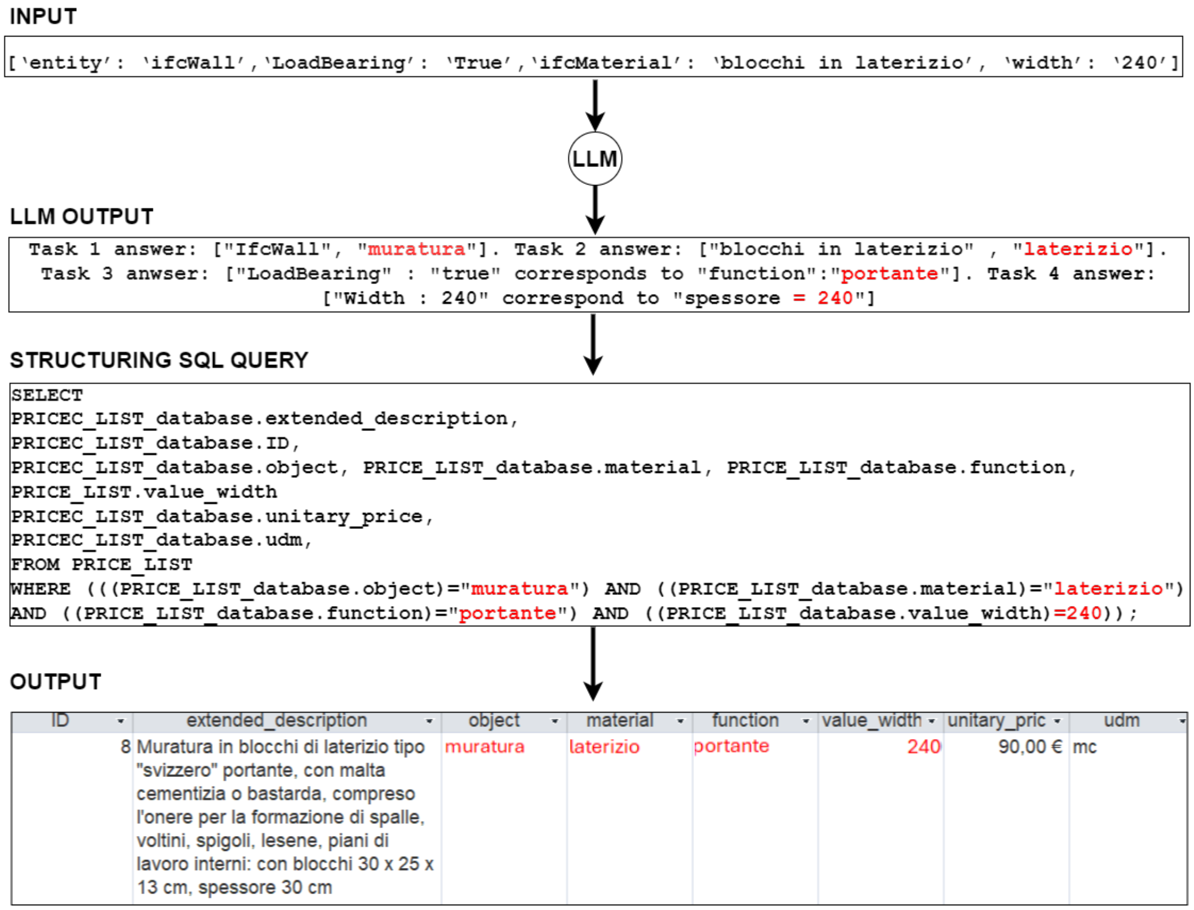

Figure 4 Complete example of the framework tested on a case study.

References

Akanbi, T., & Zhang, J. (2021). Design information extraction from construction specifications to support cost estimation. *Automation in Construction*, *131*, 103835.

Beetz, J., Van Leeuwen, J., & De Vries, B. (2009). IfcOWL: A case of transforming EXPRESS schemas into ontologies. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM*, *23*(1), 89–101. https://doi.org/10.1017/S0890060409000122

Choi, J., Kim, H., & Kim, I. (2015). Open BIM-based quantity take-off system for schematic estimation of building frame in early design stage. *Journal of Computational Design and Engineering*, *2*(1), 16–25.

Dai, Z., Zhao, V. Y., Ma, J., Luan, Y., Ni, J., Lu, J., Bakalov, A., Guu, K., Hall, K. B., & Chang, M.-W. (2022). *Promptagator: Few-shot Dense Retrieval From 8 Examples*. https://arxiv.org/abs/2209.11755v1

Gatto, C., Farina, A., Mirarchi, C., & Pavan, A. (2023). Development of a framework for processing unstructured text dataset through NLP in cost estimation AEC sector. *Proceedings of the 2023 European Conference on Computing in Construction and the 40th International CIB W78 Conference*, *4*, 0–0. https://doi.org/10.35490/EC3.2023.232

Gatto, C., Gholamzadehmir, M., Zampogna, M., Mirarchi, C., & Pavan, A. (2023). *An Automated Framework for Ensuring Information Consistency in Price List Tendering Document*. 837–847. https://doi.org/10.36253/979-12-215-0289-3.83

Hassan, F. ul, & Le, T. (2020). Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, *12*(2), 04520009. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000379

Isikdag, U., Aouad, G., Underwood, J., & Wu, S. (2007). *Building information models: a review on storage and exchange mechanisms*.

Jafari, P., Al Hattab, M., Mohamed, E., & Abourizk, S. (2021). Automated Extraction and Time-Cost Prediction of Contractual Reporting Requirements in Construction Using Natural Language Processing and Simulation. *Applied Sciences 2021, Vol. 11, Page 6188*, *11*(13), 6188. https://doi.org/10.3390/APP11136188

Ma, Z., Wei, Z., & Zhang, X. (2013). Semi-automatic and specification-compliant cost estimation for tendering of building projects based on IFC data of design model. *Automation in Construction*, *30*, 126–135.

Moon, S., Lee, G., & Chi, S. (2021). Semantic text-pairing for relevant provision identification in construction specification reviews. *Automation in Construction*, *128*, 103780. https://doi.org/10.1016/J.AUTCON.2021.103780

Opitz, F., Windisch, R., & Scherer, R. J. (2014). Integration of Document- and Model-based Building Information for Project Management Support. *Procedia Engineering*, *85*, 403–411. https://doi.org/10.1016/J.PROENG.2014.10.566

Preidel, C., Daum, S., & Borrmann, A. (2017). Data retrieval from building information models based on visual programming. *Visualization in Engineering*, *5*(1). https://doi.org/10.1186/S40327-017-0055-0

Ruiz-Morales, C., Antonio Vera-Sánchez, J., González-López -, A., Xinglei, X., Zhiliang, M., Huanmin, H., & Peng, X. (2022). Automated extraction of attributes of IFC objects based on graph theory and SPARQL query. *IOP Conference Series: Materials Science and Engineering*, *1218*(1), 012024. https://doi.org/10.1088/1757-899X/1218/1/012024

Sdino, L., & Rosasco, P. (2021). The Regional Price Lists for Estimating the Costs of Construction. *GREEN ENERGY AND TECHNOLOGY*, 213–229. https://doi.org/10.1007/978-3-030-49579-4

Sepasgozar, S. M. E., Costin, A. M., Karimi, R., Shirowzhan, S., Abbasian, E., & Li, J. (2022). BIM and Digital Tools for State-of-the-Art Construction Cost Management. *Buildings 2022, Vol. 12, Page 396*, *12*(4), 396. https://doi.org/10.3390/BUILDINGS12040396

Tang, S., Liu, H., Almatared, M., Abudayyeh, O., Lei, Z., & Fong, A. (2022). Towards Automated Construction Quantity Take-Off: An Integrated Approach to Information Extraction from Work Descriptions. *Buildings 2022, Vol. 12, Page 354*, *12*(3), 354. https://doi.org/10.3390/BUILDINGS12030354

Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., & Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, *134*, 104059.

Yin, M., Tang, L., Webster, C., Xu, S., Li, X., & Ying, H. (2023). *An ontology-aided, natural language-based approach for multi-constraint BIM model querying*. https://arxiv.org/abs/2303.15116v1

Zhang, J., & El-Gohary, N. M. (2016). Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking. *Journal of Computing in Civil Engineering*, *30*(2). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000346

Zheng, J., & Fischer, M. (2023). Dynamic prompt-based virtual assistant framework for BIM information search. *Automation in Construction*, *155*, 105067. https://doi.org/10.1016/J.AUTCON.2023.105067