2025 European Conference on Computing in Construction

CIB W78 Conference on IT in Construction

Porto, Portugal
July 14-17, 2025

PORTUGUESE CONSTRUCTION DATASET FOR AI: BOQ TEXT EXTRACTION AND SYNTHETIC DATA AUGMENTATION USING LLMS

Luís Jacques de Sousa[1,2], João Poças Martins[2], Luís Sanhudo[1], and João Miguel Silva[1]
[1]BUILT CoLAB - The Collaborative Laboratory for the Built Environment of the Future, Porto, Portugal
[2]CONSTRUCT-GEQUALTEC, Department of Civil and Georesources Engineering, Faculty of Engineering (FEUP), University of Porto, Porto, Portugal

Abstract

Manual classification of Bill of Quantities (BOQs) in construction procurement is labour-intensive and error-prone, limiting efficiency in bidding and contract management. Few structured open datasets for BOQ classification exist, restricting automation approaches. To address this, we present a labelled dataset of BOQ tasks from Portuguese public procurement contracts, structured for multilabel classification. Synthetic augmentation using GPT-4o Mini and cosine similarity-based batching mitigated class imbalance, expanding training data to 23,542 examples per fold (3 folds). This dataset provides a Portuguese construction corpus and enables Artificial Intelligence-driven BOQ task classification, fostering procurement automation and expanding automation routes in construction contract analysis.

Background

BOQs are essential in construction contracts, defining project scope, tasks, and costs. Despite their importance in tendering, BOQ processes remain manual, labour-intensive, and prone to errors, impacting bid quality and contract compliance (Luís Jacques de Sousa and Martins, 2023, IMPIC, 2019). AI-driven solutions can improve BOQ classification by matching project requirements with internal databases and enhancing cost estimation accuracy (Zheng et al., 2023, Du et al., 2024, Mahmud et al., 2025, Rabbani et al., 2025).

Artificial intelligence (AI) techniques such as machine learning (ML) and Natural Language Processing (NLP) have seen implementation in contract analysis, energy management, and structural optimization (Regona et al., 2022). However, their application remains limited due to fragmented documentation, lack of data availability and of awareness of AI capabilities (Sepasgozar and Davis, 2018, Saka et al., 2024, Ghimire et al., 2023). The unstructured nature of construction data requires extensive preprocessing before AI models can be effectively implemented (Saka et al., 2024, Jacques de Sousa et al., 2023b, Jacques de Sousa et al., 2023a).

Recent publications (Lee et al., 2023, Deza et al., 2022) illustrate the potential of structured BOQ datasets for streamlining classification and cost estimation. Building on this foundation, our work expands both the scope and accessibility of these data-driven methods by introducing a broader, fully open-source dataset and a novel synthetic data augmentation framework.

Large Language Models (LLMs) can help overcome data limitations in Construction (Gilardi et al., 2023). However, the use of AI-generated data raises challenges, as construction stakeholders may be reluctant to trust synthetic data-driven insights without transparent explainability features (Abioye et al., 2021). Additionally, training domain-specific LLMs requires substantial technical expertise and computational resources, posing significant barriers for smaller firms (Akinosho et al., 2020).

This research promotes AI adoption in Construction by developing a structured, labelled dataset for classification tasks and providing a Portuguese Construction Corpus that unlocks further NLP applications for contract analysis, procurement automation, and BOQ classification.

Methods

Data Retrieval and Classification Methodology

The methodology for data retrieval and classification, shown in Figure 1, began with the development of a web scraper to collect data from Portal Base, the Portuguese public procurement contract repository. Using Selenium and Chrome Driver, all contractual data related to public procurement projects in Portugal from 2015 to 2022 was scraped and downloaded. This dataset included both semi-structured statistical contract information—previously synthesized in another publication (Jacques de Sousa et al., 2023a), and unstructured file data of different file types.

After storing all the gathered data, the dataset processing presented in this paper focused on identifying and extracting Excel files (.xlsx and .csv) related to Bills of Quantities (BOQ) from 5,253 scraped contracts. The scope of this dataset is limited to structural construction tasks, a decision made to maintain a manageable classification dimensionality.

A rule-based algorithm was employed to detect BOQ files within the contracts, extract task descriptions, and store them in a structured repository. To facilitate

classification, a master BOQ template was developed in collaboration with Portuguese construction professionals. This standardized format structures BOQ task descriptions into a three-level hierarchy:

- Chapter Level – Represents broad construction categories (e.g., Reinforced Concrete Structures, Masonry Structures, Metal Structures).
- Subchapter Level – Defines more specific subcategories within the chapter (e.g., Slabs, Walls, Beams).
- Item Level – Details specific construction tasks (e.g., Solid Slabs, Solid Fungiform Slabs, Lightweight Fungiform Slabs).

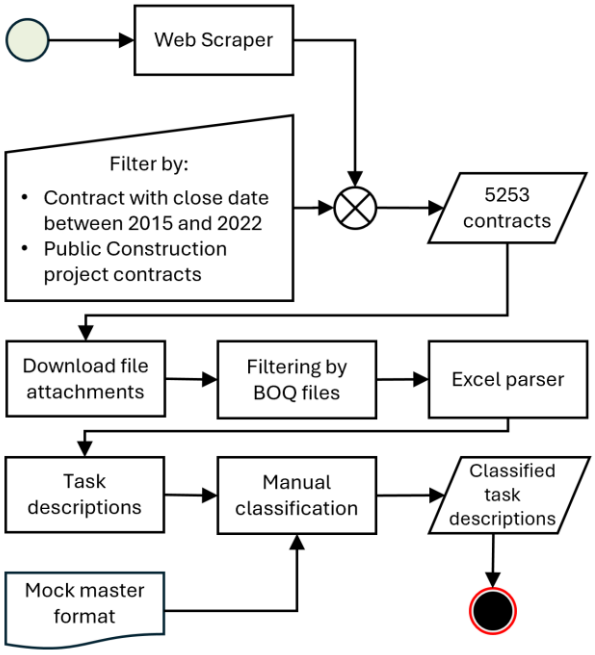


Figure 1: Data Retrieval and Classification Methodology

Each extracted task was manually classified following this hierarchical structure, resulting in 7,096 labelled task descriptions.

Synthetic Data Generation

The initial dataset of 7,096 task descriptions was deemed insufficient for AI applications due to data imbalance across the three hierarchical levels. Early tests revealed that certain subchapters and items were significantly underrepresented, motivating the need for synthetic data augmentation to expand the corpus while preserving its structure.

Given the scarcity of task-specific datasets and models, many researchers rely on LLMs to generate labelled data (Laurer, 2024). Annotating large-scale datasets is both time-consuming and costly, particularly in specialized domains (He et al., 2023). Recent studies show that LLMs now match or exceed human performance in data annotation, producing high-quality synthetic data comparable to expert-labelled datasets (Gilardi et al., 2023).

To address data limitations, the ChatGPT-4o mini API was used to expand the dataset. The ChatGPT-4o Mini API was chosen for its improved handling of non-English text, lower cost, and higher token limits compared to GPT-4 (Craig, 2024).

Inspired by the Synthetic Minority Over-sampling Technique (SMOTE), which enhances minority class representation by interpolating between samples (Preeti, 2024), our approach involved rewriting, extending, shortening, and modifying task descriptions to balance the dataset at the subchapter level. Unlike SMOTE, which is designed for numerical data, direct sentence interpolation is not feasible in NLP tasks. Instead, synthetic examples were generated using sentence-level transformations, mimicking SMOTE's balancing effect in a way suitable for text-based classification.

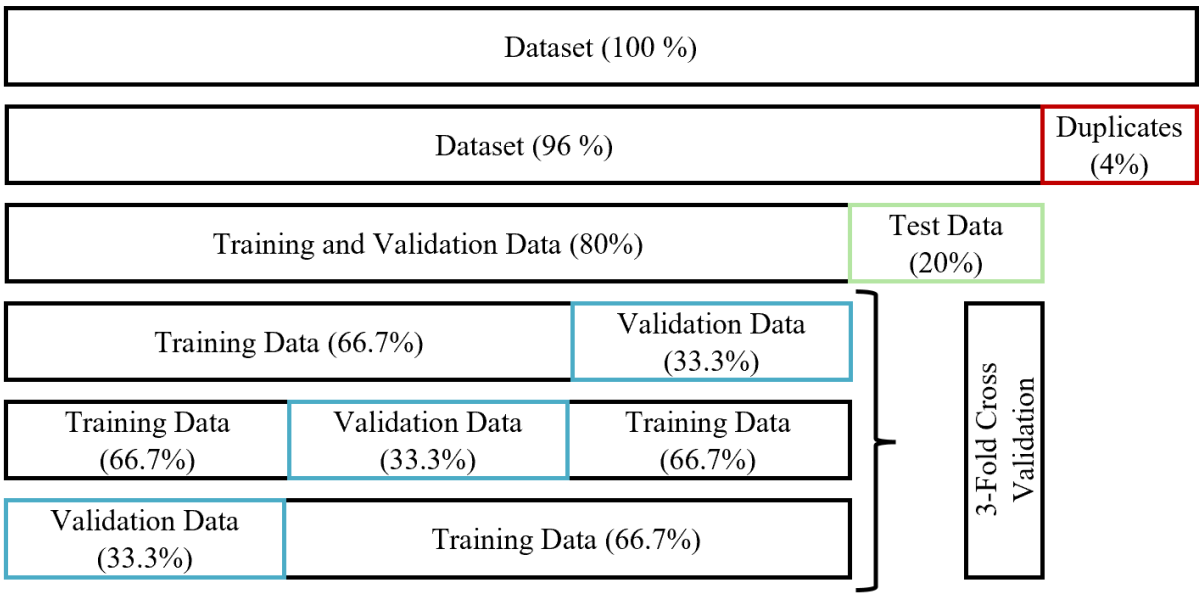


Figure 2: Dataset 3-fold split in preparation for LLM augmentation

As shown in Figure 2, the training and validation data were split using 3-fold cross-validation. Before splitting, the data was shuffled to ensure randomness, and a fixed random seed was applied to maintain reproducibility. The process iterated through the training and validation indices, creating distinct subsets for each fold.

This method enables the use of cross-validation for model evaluation while also serving a key purpose of this study, enhancing synthetic data generation. By providing varied training scenarios across folds, this approach improves the robustness and general applicability of the generated synthetic data.

As seen in Figure 3, to ensure validity, each fold in the 3-fold cross-validation originated its own set of synthetic data, as the training composition differed across folds. Only the training data from each fold was used as input for the LLM to generate synthetic examples.

Synthetic augmentation was applied exclusively to the training dataset, with new examples created to balance minority classes. This ensured that all synthetic data remained strictly confined to training, preventing any contamination of the test set.

As a result, three independent rounds of synthetic data generation were conducted, one for each fold. This approach maintains the integrity of training data variability while ensuring that the test data remains an unbiased benchmark for evaluating model performance.

Prompting Strategy

In the existing literature, there is no universally accepted methodology for generating synthetic data using LLMs across different domains. Nevertheless, this process typically begins by selecting authentic data examples that accurately represent specific classes, in this case, descriptions that correspond to subchapters (Vongthongsri, 2024). These examples serve as a foundational context from which the LLM can generate new, synthetic descriptions.

To facilitate this process, batches of context are created and fed into the LLM for synthetic data generation (Vongthongsri, 2024). In the batching strategy illustrated in Figure 4, task descriptions were grouped based on cosine similarity, ensuring that each batch contained semantically related tasks. Within these similarity-based batches, random selection was employed to choose specific chunks for processing.

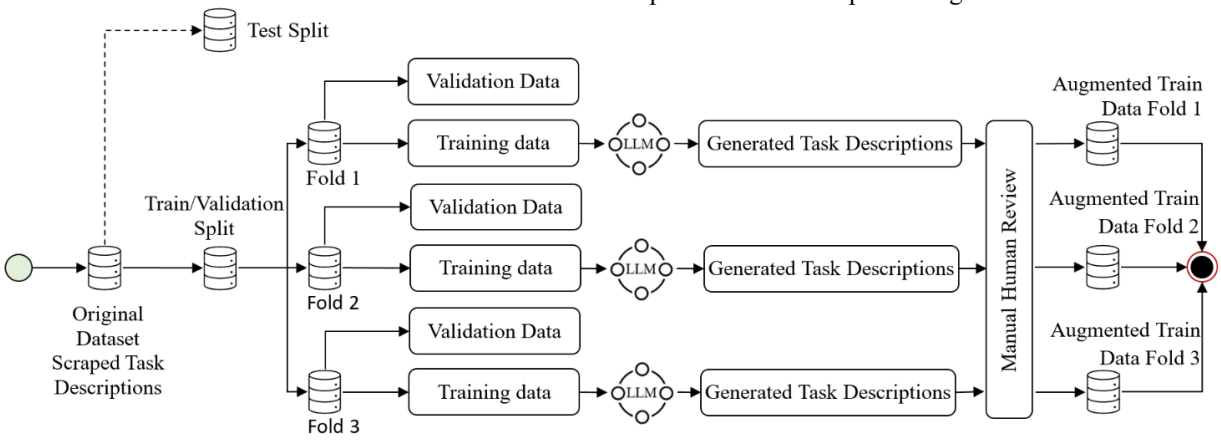


Figure 3: Three-fold LLM augmentation

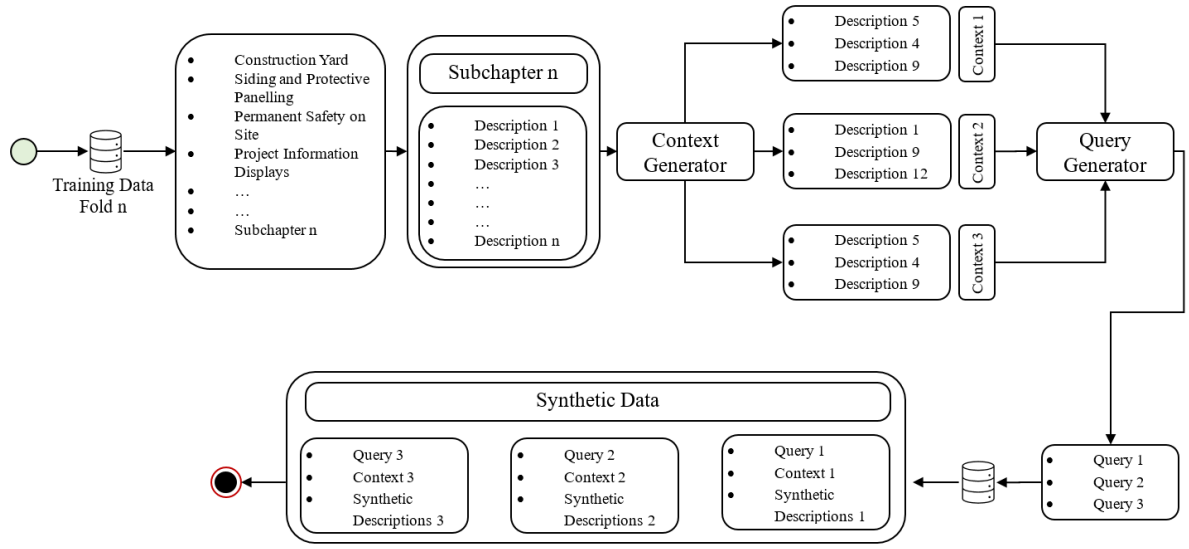
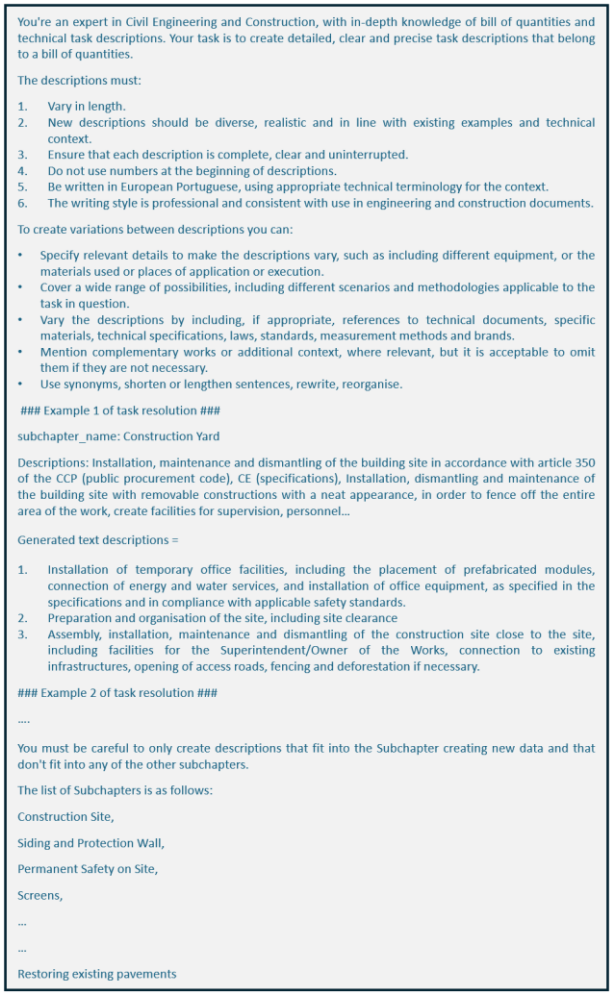


Figure 4: Synthetic Data Generator Model Architecture. Adapted from (Vongthongsri, 2024)

This method mitigates biases by promoting diversity in queries, preventing the model from generating repetitive or overly similar outputs. By randomizing context inputs within semantically similar groups, the approach enhances the variety of synthetic data, thereby improving the robustness and utility of the generated dataset. The synthetic data, when combined with its original context, forms the complete augmented dataset.

The version at the time of writing, of the ChatGPT-4o mini API operates by utilizing pairs of roles and content to determine the responses from the LLM. These pairs are typically assigned to establish the role of the system (the model's role) and to define its behaviour through specified content.

To define the model's behaviour, the system content (translated from Portuguese) in Figure 5 was specified after analysing the quality of the outputs during an iterative process.



You're an expert in Civil Engineering and Construction, with in-depth knowledge of bill of quantities and technical task descriptions. Your task is to create detailed, clear and precise task descriptions that belong to a bill of quantities.

The descriptions must:

1. Vary in length.
2. New descriptions should be diverse, realistic and in line with existing examples and technical context.
3. Ensure that each description is complete, clear and uninterrupted.
4. Do not use numbers at the beginning of descriptions.
5. Be written in European Portuguese, using appropriate technical terminology for the context.
6. The writing style is professional and consistent with use in engineering and construction documents.

To create variations between descriptions you can:

- Specify relevant details to make the descriptions vary, such as including different equipment, or the materials used or places of application or execution.
- Cover a wide range of possibilities, including different scenarios and methodologies applicable to the task in question.
- Vary the descriptions by including, if appropriate, references to technical documents, specific materials, technical specifications, laws, standards, measurement methods and brands.
- Mention complementary works or additional context, where relevant, but it is acceptable to omit them if they are not necessary.
- Use synonyms, shorten or lengthen sentences, rewrite, reorganise.

Example 1 of task resolution

subchapter_name: Construction Yard

Descriptions: Installation, maintenance and dismantling of the building site in accordance with article 350 of the CCP (public procurement code), CE (specifications), Installation, dismantling and maintenance of the building site with removable constructions with a neat appearance, in order to fence off the entire area of the work, create facilities for supervision, personnel...

Generated text descriptions =

1. Installation of temporary office facilities, including the placement of prefabricated modules, connection of energy and water services, and installation of office equipment, as specified in the specifications and in compliance with applicable safety standards.
2. Preparation and organisation of the site, including site clearance
3. Assembly, installation, maintenance and dismantling of the construction site close to the site, including facilities for the Superintendent/Owner of the Works, connection to existing infrastructures, opening of access roads, fencing and deforestation if necessary.

Example 2 of task resolution

....

You must be careful to only create descriptions that fit into the Subchapter creating new data and that don't fit into any of the other subchapters.

The list of Subchapters is as follows:

Construction Site,

Siding and Protection Wall,

Permanent Safety on Site,

Screens,

...

...

Restoring existing pavements

Figure 5: System Role (translated from Portuguese)

After defining the content for the model's role, the content for the user's role is specified. This content varies depending on the subchapter being addressed, the task descriptions chunked according to cosine similarity, and the number of tasks to be created based on the deficit from the minority subchapter to the majority subchapter class.

The prompt provided to the generator incorporates several crucial elements:

- subchapter_name: which identifies the specific subchapter.
- descriptions: which are chunks of text grouped by cosine similarity to form a comprehensive context.
- deficit: which quantifies the deficit in the number of examples needed between the minority and majority subchapter classes.

The following prompt, defining the user's role, was provided to the LLM generator, as seen in Figure 6.

"Here are the task descriptions for the subchapter '{subchapter_name}': \n\n{descriptions}\n\n"

"Based on these examples, create {deficit} new task descriptions for this subchapter in European Portuguese."

Figure 6: User's Role (translated from Portuguese)

Finally, each generated task description was manually reviewed to ensure alignment with the master BOQ format classification and to confirm its validity within the corresponding construction theme. An example of this process can be found in the " Example of LLM-Generated Task Descriptions" section of this paper.

Results

Data Description

The dataset presented in this paper was collected from the Portuguese public procurement contract repository and includes contractual BOQ files from 2015 to 2022. The data is structured in tabular form, consisting of construction activity/task descriptions followed by a multilabel, one-hot encoded classification.

The dataset is pre-split into three folds, each containing separate training and testing data. The main directory includes three subfolders, each corresponding to one of the folds. Within each fold's folder, there are two files: one containing the training data and the other containing the testing data.

Each text description in the dataset is classified into three hierarchical levels: Chapter, Subchapter, and Item. The Chapter level comprises 12 classes, the Subchapter level includes 58 classes, and the Item level contains 72 classes. For instance, the hierarchical classification for a construction task can be represented as:

"Reinforced Concrete Structures → Slabs → Solid Slabs", where:

- Chapter: Reinforced Concrete Structures
- Subchapter: Slabs
- Item: Solid Slabs

Each chapter typically links to multiple subchapters, and each subchapter can connect to various items.

However, some subchapters do not descend to the item level, creating gaps in the hierarchical structure required by multi-output classification algorithms. To address this and still respect our master format, we introduced a “throwaway item” class (see Figure 7). This ensures that every hierarchical output (chapter, subchapter, and item) can be labelled consistently for both training and evaluation, preserving the integrity of the classification process.

Likewise, one of the chapters lacked any child classes, necessitating the creation of a “throwaway subchapter” class linked to the throwaway item class. These throwaway classes do not affect the generation of synthetic data, as they merely serve to fill structural gaps

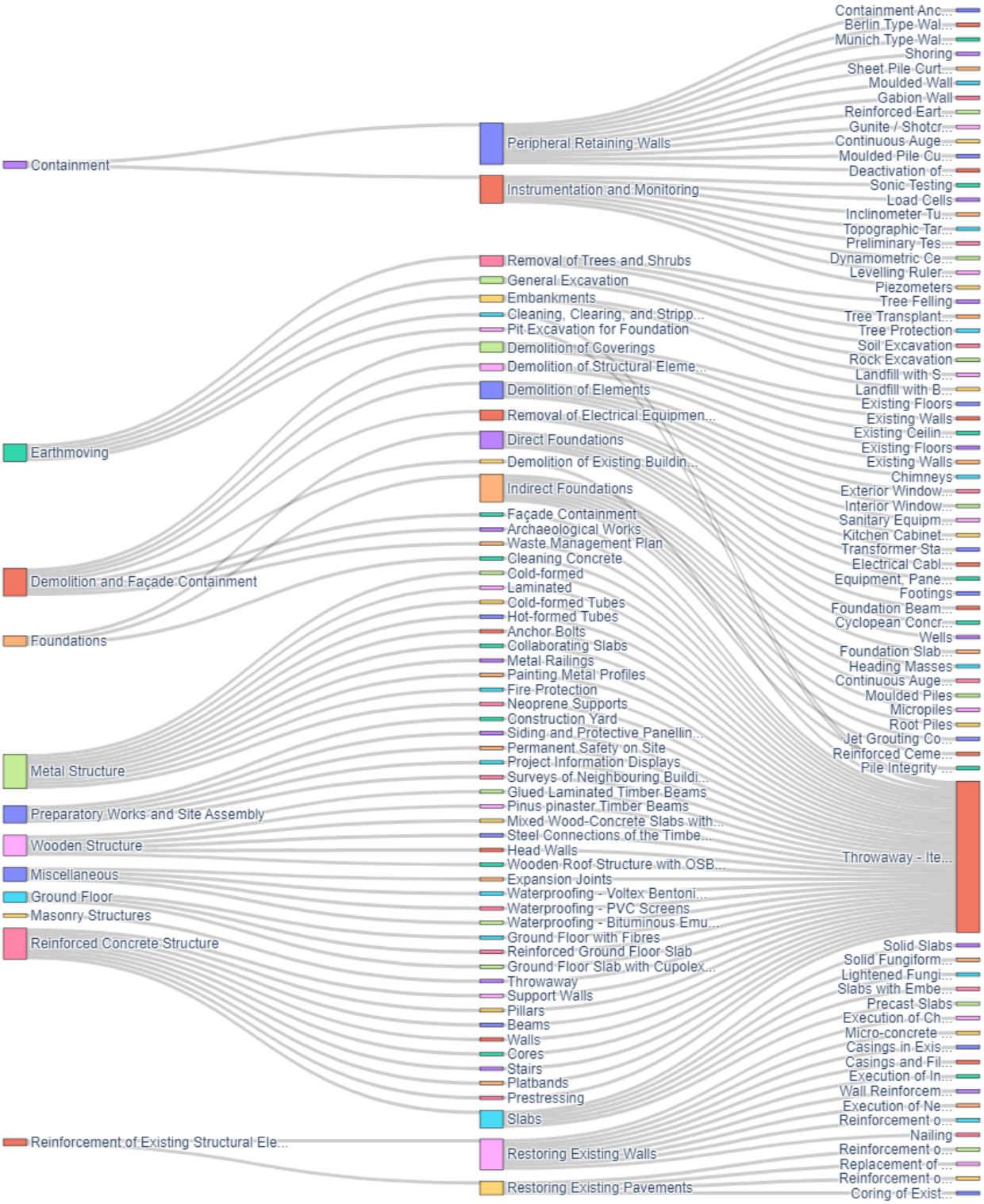


Figure 7: Relationship between hierarchical classes

without altering the semantic content derived from the original dataset. Consequently, while they uphold the classification framework's requirement for fully labelled hierarchies, they do not diminish the quality or realism of the remaining dataset.

Data Distribution

Regarding class distribution, as previously mentioned, synthetic data augmentation was applied to balance the dataset at the subchapter level. This means that the number of underrepresented subchapter classes was extended until the same number of examples in the most frequent subchapter class was reached. As a result, each subchapter class contains an equal number of instances (e.g., Fold 1: 427 examples per class).

Since the dataset follows a multilabel classification structure, it is inherently challenging to achieve perfect balance across all hierarchical levels. The interconnected nature of the hierarchy means that some labels appear more frequently than others due to the natural distribution of data. Figure 8 illustrates the distribution of chapter classifications, which remains relatively balanced.

At the item level, as seen in Figure 9, the dataset is significantly imbalanced. Many subchapters lack corresponding item classes, leading to a strong bias toward the "Throwaway Item" class.

Researchers utilizing this dataset for classification models should be mindful of this imbalance when designing and evaluating their models. The severe class imbalance at the item level, as shown in Figure 9, poses a significant challenge for classification models, as the dominance of the "Throwaway Item" class can lead to biased predictions and reduced generalization to less frequent item classes.

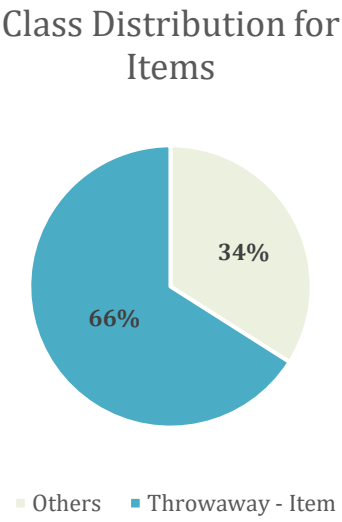


Figure 9: Item level classifications distribution

Example of LLM-Generated Task Descriptions

Our model architecture, combined with batching strategies and detailed prompt engineering, enabled the LLM to generate semantically similar task descriptions while preserving technical accuracy and adhering to construction-specific terminology. By structuring inputs around contextually relevant task groupings, our model maintained consistency with the original data while introducing meaningful variations aligned with the master BOQ classification.

For instance, consider the following real task description extracted from the dataset (it is important to note that, as the following descriptions were originally generated in Portuguese and translated into English, some linguistic

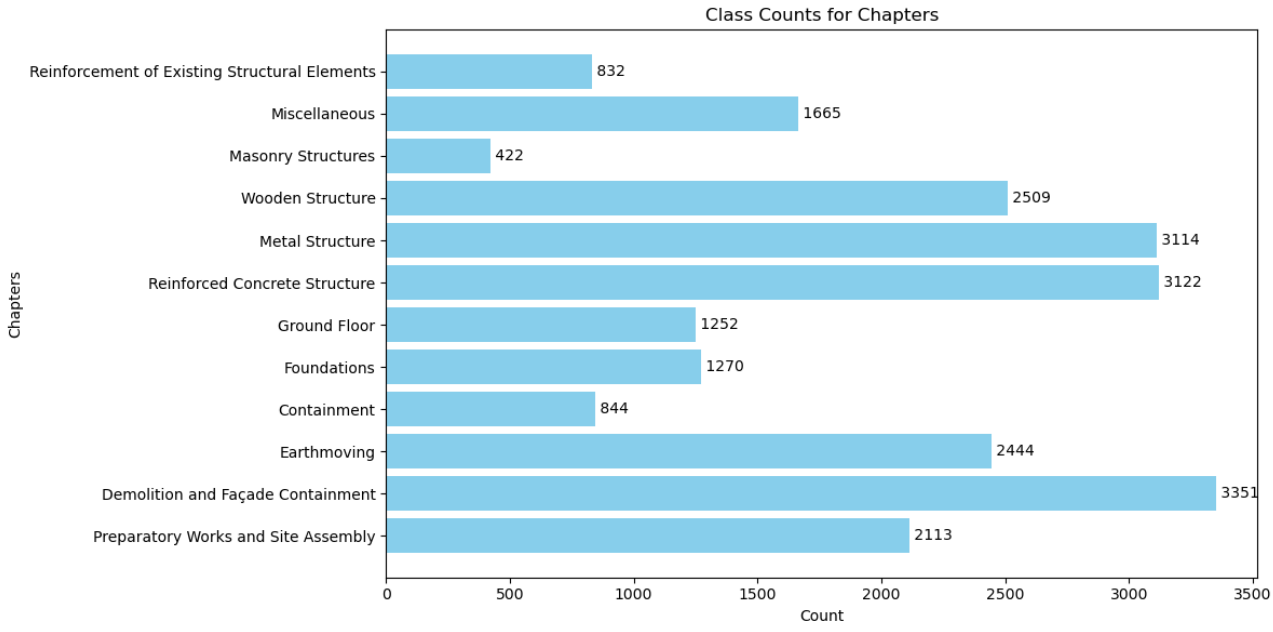


Figure 8: Chapter level classification distribution

nuances and variations may not be as impactful in this version as in Portuguese):

"*Dismantling of external metal staircases for replacement, including the safe disconnection of associated electrical or hydraulic connections, and the transportation of removed components for reprocessing or disposal in accordance with current regulations.*"

Based on this context, our model generated the following alternative task descriptions:

- *"Disassembly and removal of metal staircases in public access areas, with separation of materials for recycling and transportation of waste to an appropriate final destination, in accordance with construction industry regulations."*
- *"Removal and dismantling of external metal staircases, including material treatment for reuse in a new location according to project guidelines, as well as loading and transportation of metal waste to an authorized waste management operator."*

These examples illustrate how the synthetic task descriptions maintain alignment with the original while introducing subtle variations in phrasing and additional contextual details.

Notably, the generated descriptions rephrase tasks (e.g., "Dismantling" → "Disassembly and removal") while preserving their intended meaning. They also incorporate additional process details, such as material handling and disposal procedures. Furthermore, some descriptions introduce project-specific considerations, such as "reuse in a new location according to project guidelines", enhancing variability within the task descriptions of a given class.

Limitations

Although the authors are domain experts, the dataset may still be subject to bias due to the manual classification of tasks into chapters. Since classification was performed manually, subjective interpretations could have influenced the labelling process.

Additionally, given the large volume of text data, some degree of human error in data processing may have occurred. Despite efforts to ensure accuracy, inconsistencies or misclassifications might still be present. These factors should be considered when using the dataset for AI applications, as they negatively impact model performance and generalization.

Conclusion

The use of ML and NLP in the AEC sector remains in its early stages, largely due to difficulties in accessing relevant and reliable data. Although construction projects generate vast amounts of information, the unstructured nature of these records poses a substantial barrier to AI-driven applications. In Portugal, where procurement files are mandatorily submitted to online repositories, opportunities exist to harness these publicly accessible datasets for innovative AI research.

This study addresses the data-sourcing challenge by developing a web-scraping algorithm that successfully extracted 5,253 public procurement contracts, yielding a large, diverse dataset containing labelled BOQs tasks. Beyond simply providing a new dataset, we introduce a novel framework for synthetic text data augmentation in Construction that combines real BOQ data with LLMs. By leveraging LLMs to generate diverse synthetic examples, our approach effectively mitigates the issue of data scarcity in Construction research and lays the groundwork for more robust classification algorithms. This framework, which remains relatively unexplored in the sector, can be adapted for various applications with text data for Construction.

Our dataset is structured under a master format with three hierarchical levels (Chapters, Subchapters, and Items). It includes Portuguese-language BOQ tasks classified in a multi-labelled manner, thus reflecting the inherent ambiguity of real-world Construction BOQ translation. This openly available corpus is a valuable resource for domain-specific NLP applications, such as automated contract analysis and intelligent search systems. It facilitates further development of AI models trained on Portuguese-language corpora.

Future research can capitalise on this dataset for text-classification tasks specific to the Construction industry. Additionally, refinement of the classification structure through clustering algorithms could produce more granular or optimised task groupings. Another promising avenue involves integrating an AI agent to automatically verify the suitability of new synthetic examples, further improving dataset reliability. Finally, translating the dataset into other languages would broaden its applicability and enable cross-linguistic AI applications within the AEC sector.

All data and code for synthetic data generation are publicly accessible through the provided channels in this paper and the author's GitHub page: https://github.com/LuisJSousa/PORTUGUESE-CONSTRUCTION-DATASET-FOR-AI

Acknowledgements

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. This work is also financed by the Base Funding—UIDB/04708/2020 and Programmatic Funding—UIDP/04708/2020 (CON-STRUCT), funded by national funds through the FCT/MCTES (PIDDAC)

References

Abioye, S. O., Oyedele, L. O., Akanbi, L., Ajayi, A., Davila Delgado, J. M., Bilal, M., Akinade, O. O. & Ahmed, A. 2021. Artificial intelligence in the construction industry: A review of present status,

opportunities and future challenges. Journal of Building Engineering, 44, 103299.

Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O. & Ahmed, A. A. 2020. Deep learning in the construction industry: A review of present status and future innovations. Journal of Building Engineering, 32, 101827.

Craig, L. 2024. GPT-4o vs. GPT-4: How do they compare? [Online]. Available: https://www.techtarget.com/searchenterpriseai/feature/GPT-4o-vs-GPT-4-How-do-they-compare#:~:text=OpenAI's%20testing%20indicates%20that%20GPT,idioms%2C%20metaphors%20and%20cultural%20references.) [Accessed December 2024].

Deza, J. I., Ihshaish, H. & Mahdjoubi, L. 2022. A Machine Learning Approach to Classifying Construction Cost Documents into the International Construction Measurement Standard. ArXiv, abs/2211.07705.

Du, C., Esser, S., Nousias, S. & Borrmann, A. E. 2024. Text2BIM: Generating Building Models Using a Large Language Model-based Multi-Agent Framework. ArXiv, abs/2408.08054.

Ghimire, P., Kim, K. & Acharya, M. 2023. Generative AI in the Construction Industry: Opportunities & Challenges. ArXiv, abs/2310.04427.

Gilardi, F., Alizadeh, M. & Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences of the United States of America, 120.

He, X., Lin, Z.-W., Gong, Y., Jin, A., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N. & Chen, W. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. North American Chapter of the Association for Computational Linguistics, 2023.

Impic 2019. Anual Public Procurement in Portugal. IMPIC.

Jacques De Sousa, L., Poças Martins, J. & Sanhudo, L. 2023a. Portuguese public procurement data for construction (2015–2022). Data in Brief, 48, 109063.

Jacques De Sousa, L., Poças Martins, J., Santos Baptista, J. & Sanhudo, L. Towards the Development of a Budget Categorisation Machine Learning Tool: A Review. In: GOMES CORREIA, A., AZENHA, M., CRUZ, P. J. S., NOVAIS, P. & PEREIRA, P., eds. Trends on Construction in the Digital Era, 2023// 2023b Guimarães, Portugal. Springer International Publishing, 101-110.

Laurer, M. 2024. Synthetic data: save money, time and carbon with open source [Online]. Available: https://huggingface.co/blog/synthetic-data-save-costs#1-the-problem-there-is-no-data-for-your-use-case [Accessed December 2024].

Lee, G., Lee, G., Chi, S. & Oh, S. 2023. Automatic Classification of Construction Work Codes in Bill of Quantities of National Roadway Based on Text Analysis. Journal of Construction Engineering and Management, 149, 04022163.

Luís Jacques De Sousa, M. L. S., João Poças & Martins, L. S., Jorge Moreira Da Costa 2023. Statistical Descriptive Analysis of Portuguese Public Procurement Data from 2015 to 2022. CivilEng.

Mahmud, D., Hajmohamed, H., Almentheri, S., Alqaydi, S., Aldhaheri, L., Khalil, R. A. & Saeed, N. 2025. Integrating LLMs With ITS: Recent Advances, Potentials, Challenges, and Future Directions. IEEE Transactions on Intelligent Transportation Systems.

Preeti. 2024. Data Augmentation Techniques for Imbalanced Datasets [Online]. Medium. Available: https://medium.com/@preeti.rana.ai/data-augmentation-techniques-for-imbalanced-datasets-9214374ffe44 [Accessed December 2024].

Rabbani, S. B., Kowsar, I. & Samad, M. D. Transfer Learning of Tabular Data by Finetuning Large Language Models. 2025.

Regona, M., Yigitcanlar, T., Xia, B. & Li, R. Y. M. 2022. Opportunities and Adoption Challenges of AI in the Construction Industry: A PRISMA Review. Journal of Open Innovation: Technology, Market, and Complexity, 8, 45.

Saka, A., Taiwo, R., Saka, N., Salami, B. A., Ajayi, S., Akande, K. & Kazemi, H. 2024. GPT models in construction industry: Opportunities, limitations, and a use case validation. Developments in the Built Environment, 17, 100300.

Sepasgozar, S. M. E. & Davis, S. 2018. Construction Technology Adoption Cube: An Investigation on Process, Factors, Barriers, Drivers and Decision Makers Using NVivo and AHP Analysis. Buildings, 8, 74.

Vongthongsri, K. 2024. Using LLMs for Synthetic Data Generation: The Definitive Guide [Online]. Confident AI. Available: https://www.confident-ai.com/blog/the-definitive-guide-to-synthetic-data-generation-using-llms [Accessed December 2024].

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E. & Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. ArXiv, abs/2306.05685.