Automated Generation of Competency Questions Using Large Language Models and Knowledge Graphs*

Dario Di Nuzzo[1], Edlira Vakaj[1], Hadeel Saadany[1], Eglantina Grishti[2,*,†] and Nandana Mihindukulasooriya[3,†]

[1]*College of Computing, Birmigham City University*
[2]*University of Tirana, Tirana, Albania*
[3]*IBM Research USA, USA*

Abstract
This research presents a novel approach to automated competency question generation by integrating Large Language Models (LLMs) with Knowledge Graphs (KGs), particularly within the context of sustainability assessment standards like BREEAM. The study develops a comprehensive methodology combining natural language processing and knowledge representation to address the challenges of manual question generation in competency-based assessments. The methodology begins with text extraction from BREEAM standards, followed by preprocessing, transformation into graph documents, and the construction of a structured KG. Advanced LLMs, including GPT-4o and Mistral, are employed to generate competency questions based on entity-specific and community-focused retrieval methods. The system is rigorously evaluated using quantitative metrics such as cosine similarity scores and qualitative assessments using the "LLM-as-a-Judge" method. Results demonstrate that GPT-4 and Mistral models generate highly relevant, clear, and complex questions, highlighting the potential for scalable, domain-specific competency assessments. This research opens avenues for improving AI-driven educational technologies and personalised learning through automated, adaptive assessment tools.

Keywords
Competency Question, Large Language Models, Natural Language Processing, Artificial Intelligence

1. Introduction

The integration of LLMs with KGs marks a significant breakthrough in automating the generation of CQs, particularly for competency-based assessments. Existing approached of CQs design often face limitations in scalability, relevance, and efficiency, especially as knowledge evolve rapidly. This paper introduces an innovative methodology that leverages cutting-edge techniques in text extraction, data preprocessing, and KG construction to produce high-quality, contextually relevant CQs. Our approach is tested using the Building Research Establishment Environmental Assessment Method (BREEAM)[1] documents, standards, a leading sustainability assessment method for master planning projects, infrastructure, and buildings. By automating the generation of CQs, we aim to streamline the assessment process, ensuring it remains aligned with the latest sustainability standards and enhances the learning experience through personalisation and relevance. This methodology employs advanced natural language processing (NLP) techniques to extract and transform text from BREEAM Standards Technical Manuals into structured graph documents. The structured data is then processed to build a robust KG, which serves as the foundation for generating natural language summaries and CQs. By integrating the strengths of LLMs in language understanding and generation with the explicit and structured representation of KGs, our approach offers a scalable and efficient solution that transcends

NLP4KGC: 3rd International Workshop on Natural Language Processing for Knowledge Graph Creation in conjunction with SEMANTICS Conference,September 17-19, 2024, Amsterdam
*

*Corresponding author.
†These authors contributed equally.
Dario.Dinuzzo@mail.bcu.ac.uk (D. D. Nuzzo); edlira.vakaj@mail.bcu.ac.uk (. E. Vakaj); hadeel.saadany@mail.bcu.ac.uk (H. Saadany)
0000-0002-0877-7063 (D. D. Nuzzo); 0000-0002-0877-7063 (. E. Vakaj); 0000-0002-0877-7063 (H. Saadany)

[1]https://breeam.com/about

the capabilities of traditional methods. Therefore, this paper presents a novel approach that not only improves the efficiency and effectiveness of competency-based assessments but also ensures they remain up-to-date and relevant to evolving standards. Through the automation of CQ generation, we provide a foundation for more personalised and effective learning experiences, setting a new standard for competency assessments in various domains.

2. Related Works

2.1. Knowledge Graphs and Large Language Models

Knowledge graphs are structured representations of information, where entities are nodes and relationships between them are edges. They encode real-world knowledge in a machine-interpretable format, facilitating various AI applications, from search engines to recommendation systems [1]. A key advantage of KGs is their ability to explicitly represent and reason about relationships, making them powerful tools for organising and retrieving complex information. LLMs, such as OpenAI's GPT series and Google's BERT, have demonstrated remarkable capabilities in understanding and generating human-like text [2]. These models are trained on vast corpora of text data, enabling them to capture a wide range of linguistic patterns and knowledge implicitly [3]. However, LLMs often struggle with knowledge requiring precise and explicit representations, where KGs excel. The synergy between KGs and LLMs lies in their complementary strengths: while LLMs provide powerful language understanding and generation capabilities, KGs offer structured and explicit knowledge representation [4, 2, 3].

The integration of KGs and LLMs can be approached from two primary directions: using LLMs to enhance KGs and using KGs to augment LLMs. One approach involves leveraging LLMs to automatically extract and populate KGs with new information, thus enhancing the KG's coverage and accuracy. Techniques such as entity extraction, relation extraction, and link prediction are employed to extract structured knowledge from unstructured text [5, 6, 7]. Studies have shown that LLMs can significantly improve the quality and breadth of KGs, making them more robust and comprehensive. Conversely, KGs can be used to improve the performance and reliability of LLMs. For instance, incorporating KG-based information into a fine-tuned LLM can provide additional context and background knowledge, leading to better understanding and generation of contextually rich and accurate responses [8, 9]. These hybrid approaches ensure that language models are not solely reliant on the statistical patterns in the text but are also informed by structured, explicit knowledge.

However, the effectiveness of these strategies is contingent on the development of robust methodologies for embedding and integrating these diverse forms of knowledge [10, 11]. Current techniques are still evolving, and their efficacy varies significantly across different applications and domains [12]. Moreover, the computational demands of training and fine-tuning LLMs with KG-enriched datasets are substantial. The resources required for such processes are often prohibitive, limiting the accessibility and scalability of these advanced models [13]. This raises critical questions about the practicality and sustainability of the widespread implementation of such integrated systems.

2.2. Ontology Engineering and Competency Questions

Ontology engineering systematically develops formal representations of knowledge within a specific domain, defining entities, attributes, and relationships to support the organisation and interpretation of complex information [14]. This framework enables machines to process and understand data more effectively. Ontologies include classes for general concepts, instances to illustrate these classes, properties for describing attributes, and relationships to depict connections between entities. Competency questions are crucial in ontology engineering, assessing the coverage and effectiveness of an ontology [15]. They ensure the ontology accurately represents necessary knowledge and can support relevant queries. [16] describe an iterative process that begins with identifying key concepts and entities within a domain, followed by formulating questions that the ontology should be able to answer. This process refines the ontology based on its ability to address these questions.

[16] also introduce a semi-automatic framework for constructing KGs using LLMs. This pipeline involves formulating CQs, developing an ontology, constructing KGs, and evaluating the resultant KG with minimal human involvement. Although this approach reduces human effort, it underscores challenges in achieving full automation due to inaccuracies and the necessity for expert oversight.[17] explore the automatic generation of competency questions using LLMs, comparing different models under various settings. Their findings highlight variability in model performance, raising concerns about the consistency and reliability of CQ generation. While LLMs can enhance ontology development and validation, dependence on model-specific performance requires further refinement for standardised application across diverse ontologies.

[18] propose OntoChat, a framework that leverages conversational agents and LLMs to facilitate ontology engineering tasks. The chatbot supports requirement elicitation, CQ extraction, and analysis. However, reliance on conversational agents might lead to inconsistencies, particularly in specialised domains. While OntoChat can streamline interactions, its effectiveness is limited by the current capabilities of conversational [19] present DeepOnto, a Python package integrating deep learning techniques with ontology engineering tasks. The package utilises pre-trained LLMs for tasks like ontology alignment and completion. Although it shows significant potential in automating complex knowledge representation tasks, challenges such as model interpretability, potential biases in training data, and computational resource requirements need critical attention.

Integrating deep learning models into ontology engineering requires addressing these challenges to fully realise their potential [16]. While integrating LLMs and KGs in automating competency question generation shows significant promise, it raises critical concerns about reliability, consistency, and the need for human oversight [17]. The methodologies and tools discussed provide a solid foundation for advancing automated CQ generation, but their practical implementation requires careful consideration of the challenges and limitations inherent in current AI technologies. Ensuring the quality, relevance, and accuracy of generated CQs remains complex, necessitating ongoing refinement and validation to fully leverage the potential of LLMs and KGs in ontology engineering.

2.3. Retrieval Based KGs with Agents

The integration of retrieval-based systems and augmentation techniques in KG construction and maintenance has gained significant attention due to their potential to enhance coverage, accuracy, and relevance [20, 21]. These methods leverage automated agents to streamline processes, with retrieval-based approaches focusing on extracting relevant information from extensive data sources and augmentation techniques enriching existing KGs with new insights [22]. For instance, the Graph Retrieval-Augmented Generation (GraphRAG) approach exemplifies how retrieval-augmented generation combined with graph-based indexing systems can enhance the efficiency and comprehensiveness of summaries generated from large text corpora [23]. Once information is retrieved, augmentation techniques integrate this new data into the KG, involving processes such as entity resolution and relationship extraction. Additionally, Zhong et al. [24] demonstrated the importance of synthetic data generation for overcoming data scarcity and improving the integration of natural language with structured query languages through their SyntheT2C framework.

The application of retrieval and augmentation techniques is typically facilitated by intelligent agents that autonomously perform tasks such as querying data sources, extracting relevant information, and integrating it into KGs [22, 25, 24]. These agents leverage advanced algorithms and AI models to enhance both efficiency and scalability. Advanced AI methodologies such as Chain-of-Thought (CoT) reasoning, Reflexion, and ReAct have been integrated into these frameworks to further enhance capabilities [26, 27, 28]. CoT reasoning allows agents to decompose complex queries into manageable steps, improving systematic information retrieval and integration [27]. Reflexion techniques enable continuous evaluation and refinement of the agent's performance through self-reflection and feedback loops [26], while ReAct frameworks combine reasoning with action, allowing dynamic adjustment based on real-time analysis of actions' impacts [28].

Prompt engineering is crucial for the effectiveness of these processes, particularly in zero-shot and few-shot learning contexts. By designing prompts that guide intelligent agents' responses, the accuracy and relevance of retrieved and augmented data can be significantly improved. Zero-shot approaches enable agents to generate useful outputs without explicit examples in the training data, while few-shot approaches use a small number of examples to fine-tune responses [29, 30]. However, significant challenges must be addressed to fully realise these technologies' potential. One major challenge is the quality and reliability of the retrieved data. Not all sources are equally reliable, and integrating inaccurate information can degrade the KG's quality. [31] highlight LLMs struggle with using long contexts effectively, particularly when relevant information is positioned in the middle of the input context. This positional bias poses a significant hurdle for models processing extensive data inputs comprehensively. Additionally, the computational complexity of the retrieval and augmentation processes can be resource-intensive, requiring sophisticated optimisation techniques. Ensuring the scalability of these methods to handle continuously growing datasets remains another significant challenge [23, 20, 21, 24, 27, 31].

The combination of retrieval and augmentation techniques facilitated by intelligent agents therefore represents a powerful approach to advancing KG development. While challenges exist, the potential benefits are substantial, promising more comprehensive, accurate, and useful knowledge representations. The ongoing evolution of AI and related technologies, particularly with the integration of CoT reasoning, Reflexion, and ReAct frameworks, or new approaches, will be crucial in overcoming these challenges and unlocking the full potential of KGs in various applications.

3. Approach

3.1. Text Extraction and Preprocessing

The methodology employed in this paper involves a comprehensive and structured approach to extract, process, and transform text from various documents into a KG, followed by summarisation and competency question generation as shown in Figure 1. This multi-phase process ensures that the extracted data is accurately represented, contextually enriched, and ready for further analysis and application. The first step in the project methodology involves extracting text from PDF documents, specifically from BREEAM Standards Technical Manuals. This extraction was performed using the PyPDF2 library, which efficiently parses and retrieves text from each page of the PDF files, ensuring a comprehensive capture of the content. Once extracted, the text underwent a preprocessing stage where it was divided into manageable chunks of 500 words each. To maintain contextual continuity across these chunks, a 100-word overlap was employed. This chunking strategy is crucial for preserving the context and ensuring that subsequent processes, such as entity recognition and relationship extraction, have access to all necessary contextual information. This approach ensures that the text data is optimally prepared for transformation into structured graph documents in the next phase of our methodology.

3.2. Transformation to Graph Documents

After preprocessing, the chunked text was processed using LangChain's LLM Graph Transformer in conjunction with OpenAI's GPT-3.5-turbo. This step involved transforming the text chunks into structured graph documents. The LLM Graph Transformer and GPT-3.5-turbo applied advanced language understanding techniques to identify and classify entities and relationships within the text. As seen in Figure 1, these entities and relationships were then organised into a preliminary structure of nodes and edges, effectively representing the knowledge contained in the text. This transformation was essential for converting unstructured text into a structured format that can be further processed and analysed. The resulting graph documents form the foundation of our KG, enabling efficient querying and manipulation of the data in subsequent steps.

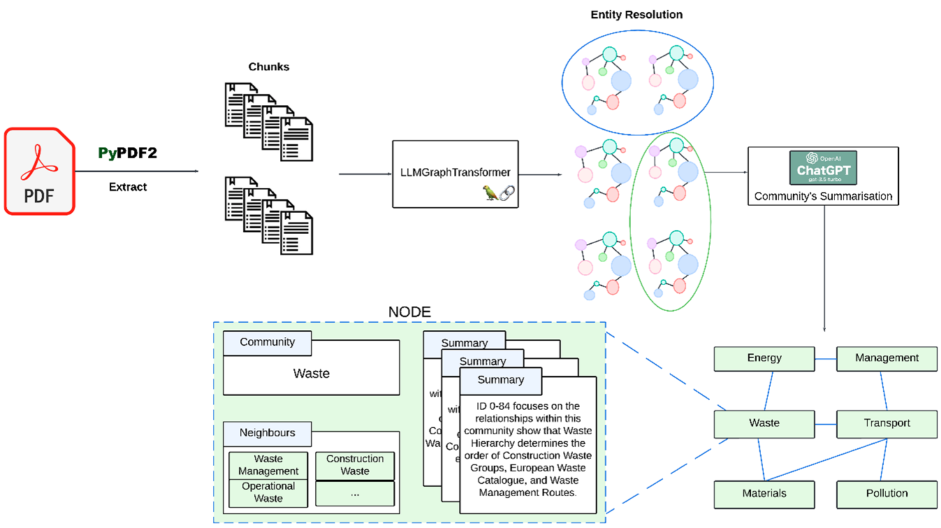

Figure 1: From PDF to KG

3.3. Building the Knowledge Graph

The initial graph documents served as the foundation for constructing the KG in Neo4j, a powerful graph database platform known for its efficient handling of complex data relationships [32]. As presented in Figure 1, the construction process began with entity resolution, which involved merging similar entities to ensure a clean and accurate representation of the data. This was achieved by utilising vector embeddings and similarity scoring techniques available in the Neo4j Graph Data Science (GDS) library, which allowed us to represent entities in a high-dimensional space and merge those with high similarity scores. Following entity resolution, we conducted community detection using the Leiden algorithm [33]. Community detection involves identifying clusters or groups of related entities within a network, helping to reveal the underlying structure and relationships among the entities.

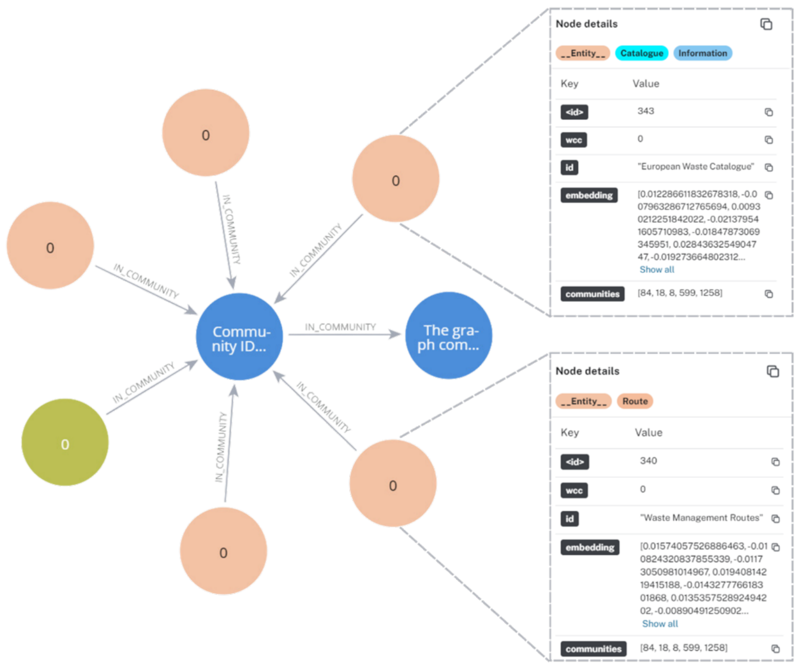

Figure 2: KG Data Representation

The Leiden algorithm is particularly effective for this purpose, ensuring accurate and meaningful grouping of related entities within large networks [23]. This algorithm facilitated the organisation of the KG into distinct communities, each representing a cohesive subset of related entities and relationships, as illustrated in Figure 2. The resulting KG was robust, easily searchable, and ready for the subsequent summarisation and competency question generation phases.

3.4. Natural Language Summarisation

The final step in our methodology involved generating natural language summaries for each community within the KG using OpenAI's GPT-3.5-turbo. Figure 3 demonstrates an example of a generated community. The GPT-3.5-turbo model was employed to create concise and coherent summaries for subsets of nodes and their relationships. These individual summaries were then combined to form a comprehensive summary for each community.

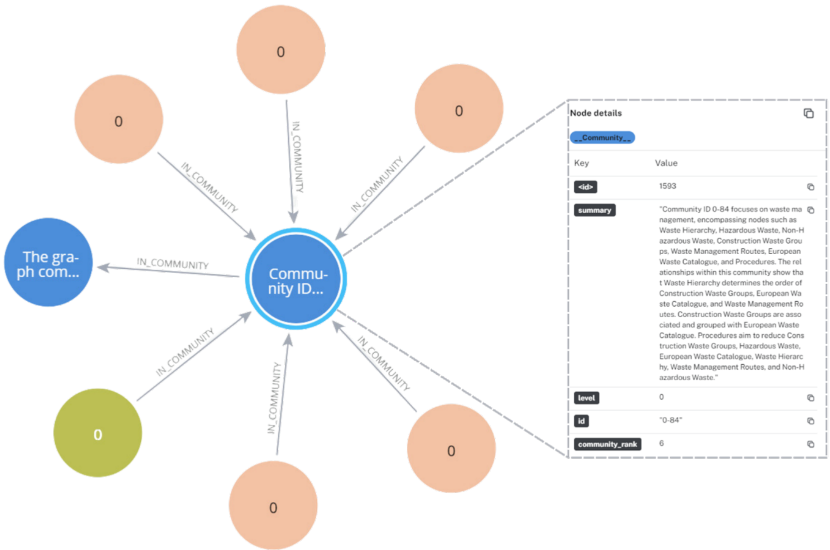

Figure 3: KG Summaries

The iterative summarisation process ensured that the generated summaries accurately reflected the content and context of the graph communities, providing a clear and comprehensive overview of the information encapsulated within each community. This step significantly enhanced the interpretability of the KG, making it easier to understand and utilise. Moreover, these natural language summaries were instrumental in facilitating the generation of competency questions that are well-aligned with the identified communities and their relevant contexts.

3.5. Retrieval Augmentation Generation (RAG) Phases

3.5.1. Retrieval Phase

The retrieval phase in our methodology involved two distinct approaches for extracting pertinent information from the KG to support the generation of competency questions: entity-focused retrieval and community-focused retrieval as presented in Figure 4. Initially, we targeted specific entities within the KG, identified based on their relevance to the predefined assessment criteria. Using the extract_top_topics_from_graph function, we queried the graph to obtain counts of entities and their corresponding mentions, providing an overview of the data's scope within the graph. For each targeted entity, the retrieve_relevant_chunks_from_graph function was employed to extract relevant text chunks.

This function utilised Cypher queries to identify documents linked to the entities and retrieved the pertinent content, ensuring that the extracted text was contextually rich and directly relevant to the entities in question. The use of the safe_neo4j_operation decorator ensured the robustness

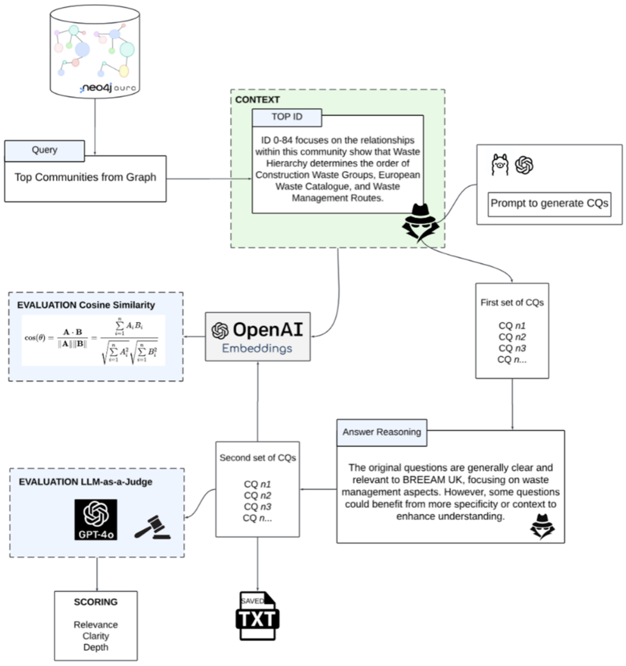

Figure 4: From KG to CQs

of the retrieval process by gracefully handling errors and maintaining continuity. In addition to entity-focused retrieval, we implemented a community-focused approach to capture broader contextual information. Using the retrieve_relevant_communities_from_graph function, we extracted summaries of top communities within the KG. Each community summary provided an aggregated view of the interconnected entities and their relationships, offering a comprehensive context for further processing. This approach allowed us to leverage the structural and relational information captured in the KG to generate more holistic and contextually relevant competency questions. Both retrieval approaches provided a robust foundation for the subsequent augmentation phase, where the extracted text would be further processed to enhance its relevance and informativeness. By combining entity-focused and community-focused retrieval methods, we ensured a rich and comprehensive data context for generating competency questions.

3.5.2. Augmentation and Generation Phases

The augmentation and generation phases of our methodology are crucial for refining the extracted data and ensuring the production of high-quality, relevant competency questions. These phases employ sophisticated natural language processing techniques to enhance the accuracy and contextual appropriateness of the generated questions. The augmentation and generation phases are integral to enhancing the relevance, informativeness, and quality of the competency questions. These phases involve several detailed steps that integrate advanced NLP techniques, including CoT reasoning and Reflexion techniques, to ensure the production of CQs.

The CoT technique involves guiding the agent through a sequence of intermediate reasoning steps that lead to the final generation of questions. This method is particularly effective for complex tasks as it helps the agent maintain logical consistency and coherence throughout the reasoning process [34]. By breaking down the task into smaller steps, CoT ensures that each step is thoroughly understood and processed, ultimately enhancing the quality and relevance of the generated CQs. After the initial generation of questions, the Reflexion technique prompts the agent to Self-Reflect reviewing and evaluating its own outputs as show in Appendices 7 to 9. This includes identifying any gaps or weaknesses in the questions, assessing their alignment with the context, and suggesting improvements. The Reflexion process is iterative, meaning that the agent continuously refines the questions through multiple cycles of evaluation and enhancement. This iterative refinement is crucial for ensuring that the final set of questions is comprehensive and of high quality.

Subsequently, we applied both zero-shot and few-shot learning approaches to the text chunks and community summaries retrieved from the KG. The zero-shot approach, as illustrated in 5, involves generating questions without any prior examples, relying solely on the persona's knowledge and the provided context [35]. In contrast, the few-shot approach, detailed in 6, involves providing the agent with a few examples of relevant questions before generating new ones [36]. This helps the agent to better understand the desired format and style, which would lead to a more accurate and contextually appropriate set of CQs. Both types of retrieval, entities and communities, were used in conjunction with these approaches. For entities, the agent focused on generating questions that are directly related to specific entities identified within the text. For communities, the questions were generated based on the broader context provided by community summaries, which represent groups of related entities and their interactions. By employing these techniques, we ensured that the generated questions were not only relevant to specific entities but also captured the broader thematic elements within the context.

4. Experiments

4.1. Experimental Settings

Following the extraction, transformation, and summarisation processes, the experimentation phase involved benchmarking both proprietary and open-source models to generate CQs. For proprietary models, we used GPT-4[2], GPT-4o[3], and GPT-4o-mini[4]. To run open-source models, we employed Ollama[5], a tool that enabled us to retrieve and run language models locally on the machine. The open-source models benchmarked included Llama 3.1 8B[6], Llama 3 8B[7], Gemma 2 9B[8], Mistral-Nemo 12B[9], Mistral 7B[10], Qwen 2 7B[11], and Phi 3 14B[12]. All experiments were conducted on a machine equipped with an NVIDIA GeForce RTX 4090 16GB GDDR6, providing the necessary computational power to handle large-scale model inferences. Each model was tasked with generating questions, which were then refined using the CoT and Reflexion techniques using zero and few-shots. This iterative refinement process ensured that the questions were contextually relevant, clear, and of appropriate complexity. It is important to note that only the ID 0-84 summary was retrieved for the question generation experiment. This summary serves as the ground truth for evaluating the effectiveness of the models in generating competency questions. Once the questions were refined to a satisfactory level, the next step was to save them using the save_questions function. This function documented the questions and saved them to a text file, ensuring that the final output was preserved for future use. The saved questions were formatted and organised, making them easily accessible and usable for assessments. This process ensured that the refined CQs generated through rigorous benchmarking and refinement were ready for practical application in competency-based assessments.

4.2. Evaluation Methods

In this chapter, we present a detailed examination of the methods used to evaluate our approach to entity and community retrieval, and question generation. Our evaluation framework integrates both

[2]https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4
[3]https://platform.openai.com/docs/models/gpt-4o
[4]https://platform.openai.com/docs/models/gpt-4o-mini
[5]https://ollama.com/
[6]https://ollama.com/library/llama3.1
[7]https://ollama.com/library/llama3
[8]https://ollama.com/library/gemma2
[9]https://ollama.com/library/mistral-nemo
[10]https://ollama.com/library/mixtral
[11]https://ollama.com/library/qwen2
[12]https://ollama.com/library/phi3

quantitative and qualitative analyses to ensure a thorough assessment of the model's performance. We first describe the quantitative measures, focusing on the relevance of generated questions using cosine similarity scores. Following this, we explore the qualitative aspects by employing the "LLM-as-a-judge" method to assess the relevance, clarity, and depth of the questions. This dual approach provides a comprehensive understanding of the effectiveness and quality of the questions generated by our model.

4.2.1. Quantitative Evaluation

The quantitative evaluation of our approach involved a comprehensive assessment of the effectiveness of entity and community retrieval methods, using both zero-shot and few-shot learning approaches. These methods were selected to evaluate how well the model could identify relevant information and generate CQs based on the provided context. To measure the relevance of the generated questions, we calculated cosine similarity scores [37]. This metric helps quantify the degree of alignment between the embeddings of the questions and the overall context. The process began by generating embeddings for both the questions and the context using the OpenAI embedding model text-embedding-ada-002 [38]. The cosine similarity scores were then computed to determine how closely each question matched the context. Three key metrics were derived from these scores: the minimum relevance score, which indicates the lowest degree of relevance among the generated questions; the maximum relevance score, which highlights the question with the highest relevance; and the average relevance score, which provides an overall measure of the relevance across all questions.

4.2.2. Qualitative Evaluation

In addition to cosine similarity scores, we employed the "LLM-as-a-judge" method to further evaluate the quality of the generated questions [39, 40, 41, 42]. This method involved using GPT-4o to rate each question on a scale from 1 to 5 across three criteria: relevance, clarity, and depth. Relevance measures how well the question pertains to the context and key concepts, ensuring that it is pertinent and meaningful. Clarity assesses how easily the question can be understood, which is crucial for ensuring that the questions are not ambiguous or confusing. Depth evaluates the level of insight and understanding required to answer the question, ensuring that the questions are challenging and promote a deep understanding of the subject matter. By combining cosine similarity scores with the "LLM-as-a-judge" method, we ensured a robust quantitative and qualitative evaluation of the generated questions. This comprehensive assessment allowed us to verify that the questions were not only contextually relevant but also clear and insightful, providing a solid foundation for CQs.

5. Results

5.1. Analysis of Top Performers

We present and analyse the results of the experiments focusing on the automated generation of CQs using various proprietary and open-source language models. The models were benchmarked across different retrieval and augmentation methodologies, assessing their performance based on relevance, clarity, and depth of the generated CQs. This analysis provides insight into the efficacy of our approach and identifies areas for improvement. Appendix 1 summarises the average relevance scores, along with minimum and maximum scores for each model and method combination. The relevance scores were computed using cosine similarity between the CQs and context embeddings, ensuring a robust quantitative evaluation. The experimental results in Appendix 2 highlights GPT-4o and Mistral 7B as the top performers in their respective categories.

GPT-4o demonstrated outstanding proficiency in entity-focused retrieval, achieving the highest average relevance scores of 0.8501 in zero-shot settings and 0.8435 in few-shot settings. This underscores GPT-4o's capability to generate precise and contextually relevant questions, especially when dealing with specific entities. Mistral 7B excelled in community-focused retrieval tasks, with impressive scores of 0.8759 in zero-shot and 0.8764 in few-shot settings. These results indicate Mistral 7B's strength in synthesising broader contexts to produce comprehensive and relevant questions, effectively handling complex interconnections within the data.

The comparison between zero-shot and few-shot learning approaches reveals a consistent trend: the few-shot method generally yielded slightly higher relevance scores across both proprietary and open-source models. This pattern underscores the importance of providing a few examples to guide the models, enhancing their ability to generate more precise and contextually appropriate questions. For instance, Mistral 7B's average relevance score improved from 0.8759 in zero-shot to 0.8764 in few-shot community-focused retrieval, while GPT-4o showed similar improvements in entity-focused tasks. These findings highlight the added value of example-based guidance in refining the models' performance and ensuring the generation of high-quality competency questions.

While the quantitative metrics provide a solid foundation for evaluating the models, qualitative analysis is equally important to assess the clarity and depth of the generated questions. Using the "LLM-as-a-Judge" method, GPT-4o rated each question on relevance, clarity, and depth. High scores in clarity ensure that the questions are easily understandable, avoiding ambiguity, while high depth scores reflect the complexity and insight required to answer the questions, ensuring they challenge the respondents' understanding effectively. As detailed in Appendices 3 and 4 for GPT-4o, the results showed a strong performance in all three criteria. Similarly, in Appendices 5 and 6, the Mistral 7B model demonstrated comparable high scores. This confirms that the questions generated by these models were not only relevant but also clear and thought-provoking, making them suitable for use in competency-based assessments.

6. Discussion

The integration of LLMs with KGs for the automated generation of competency questions has demonstrated considerable promise, as evidenced by the comprehensive experimental results. The proprietary models from the GPT series, particularly GPT-4o, showcased remarkable proficiency in generating highly relevant and contextually accurate questions. The zero-shot and few-shot approaches both produced strong results, with GPT-4o achieving an average relevance score of 0.8501 in zero-shot entity-focused retrieval. This performance highlights the model's ability to comprehend and process complex information without prior examples, suggesting a robust intrinsic understanding of the contextual relationships within the KG data.

In the community-focused retrieval tasks, GPT-4o maintained consistent performance, with average relevance scores around 0.8239 (zero-shot) and 0.8258 (few-shot). This consistency indicates that the model can handle broader, more generalised contexts effectively. The slight improvement seen with the few-shot approach underscores the added value of example-based guidance, which enhances the model's capacity to generate contextually enriched questions. This pattern was similarly observed in GPT-4 and GPT-4o-mini, albeit with marginally lower scores, affirming the overall efficacy of the proprietary models in both entity-specific and community-based contexts.

On the other hand, the open-source models presented a varied range of performances. Mistral 7B emerged as a standout, particularly in community-focused retrieval, where it achieved the highest relevance scores of 0.8759 in zero-shot and 0.8764 in few-shot settings. This indicates a strong ability to synthesise and contextualise information from broader summaries, making it highly effective in generating comprehensive and relevant competency questions. In entity-focused tasks, Mistral 7B also performed well, although not as dominantly as in the community-focused tasks, suggesting its particular strength lies in handling more complex, interconnected data.

The Llama series and other open-source models like Gemma 2 9B and Phi 3 14B showed reliable performance but generally lagged behind the proprietary models. For instance, Llama 3 8B's performance in community-focused few-shot retrieval (0.8461) was notable but still below the proprietary counterparts. This reflects the inherent differences in training data scale and model architecture between proprietary and open-source models, impacting their respective abilities to generate high-quality questions.

The few-shot learning approach consistently yielded slightly higher relevance scores compared to zero-shot across both proprietary and open-source models. This trend underscores the importance of providing a few examples to guide the models, enhancing their ability to generate more precise and contextually appropriate questions. This improvement is particularly crucial in educational and professional training contexts, where the relevance and clarity of competency questions significantly impact the effectiveness of assessments.

Qualitative evaluations further validated these quantitative findings. The "LLM-as-a-Judge" method revealed that top-performing models like GPT-4o and Mistral 7B not only excelled in relevance but also in clarity and depth. High clarity scores indicate that the questions generated were easily understandable and free of ambiguity, while high depth scores reflect the complexity and insight required to answer them, ensuring they challenge the respondents' understanding effectively.

Despite these promising results, several areas for improvement were identified. The selection of models for summarisation tasks, for example, could be enhanced by employing more advanced models like GPT-4 to improve the contextual quality of summaries used for question generation. Additionally, expanding the range of datasets for testing will provide a more comprehensive evaluation of the approach's robustness and applicability across different domains. Incorporating a detailed qualitative analysis of the generated questions will also help refine their clarity, relevance, and pedagogical value, ensuring that they meet the high standards required for effective competency-based assessments.

In summary, the integration of LLMs with KGs represents a significant advancement in automating the generation of competency questions, offering an efficient and scalable solution that maintains high relevance, clarity, and depth. The results from this study provide a strong foundation for further refinement and expansion, addressing the identified limitations to enhance the applicability and effectiveness of this innovative methodology in various educational and professional domains.

7. Conclusion

This paper presented a proof of concept for the automated generation of competency questions through the integration of LLMs with KGs. The methodology adopted involved a multi-stage process encompassing text extraction and preprocessing, transformation into graph documents, knowledge graph construction, and the subsequent generation of CQs. The approach was specifically designed to align with the Building Research Establishment Environmental Assessment Method standards, aiming to deliver a scalable and efficient solution for competency-based assessments.

7.1. Key Findings

The study effectively demonstrated the potential of combining LLMs and KGs to create a robust framework for CQ generation. The text extraction and preprocessing stage utilised the PyPDF2 library, efficiently parsing and segmenting text into 500-word chunks with overlaps to maintain context. This ensured that subsequent processing stages retained the necessary contextual information. The transformation phase employed LangChain's LLM Graph Transformer in conjunction with OpenAI's GPT-3.5-turbo, converting text chunks into structured graph documents. This step facilitated the identification and classification of entities and relationships within the text, forming the basis of the KG. The construction of the KG was achieved using Neo4j, a graph database platform adept at handling complex data relationships. Key processes included entity resolution and community detection, which were crucial for merging similar entities and organising the KG into meaningful clusters. Natural language summarisation was performed using GPT-3.5-turbo, generating concise summaries for each community within the KG. This step significantly enhanced the interpretability of the KG, making it easier to understand and utilise for CQ generation. The iterative application of advanced NLP techniques such as CoT reasoning and Reflexion techniques ensured the production of high-quality CQs, with both zero-shot and few-shot learning approaches enhancing the contextual relevance of the questions. Experimental validation involved benchmarking proprietary models (particularly GPT-4o) and open-source models (such as Mistral 7B). The performance of these models was assessed through quantitative evaluations using cosine similarity scores and the "LLM-as-a-Judge" method alongside assessments. The results demonstrated that these models could generate relevant, clear, and insightful questions, validating the effectiveness of the proposed methodology.

7.2. Contributions and Implications

The integration of LLMs with KGs for CQ generation represents a significant advancement in the field of NLP and AI [15, 14, 17]. This approach addresses several limitations of traditional assessment methods, offering a more dynamic, scalable, and contextually enriched solution. By automating the generation of competency questions, the process becomes more efficient and capable of adapting to evolving standards, such as BREEAM. This ensures that assessments remain current, relevant, and capable of providing a personalised learning experience. The findings of this study provide a robust foundation for further refinement and expansion of the methodology. The successful demonstration of integrating LLMs with KGs highlights the potential for broader applications in various educational and professional domains. This approach not only streamlines the assessment process but also enhances its effectiveness, offering a more comprehensive evaluation of competencies.

7.3. Future Work

Several areas of potential improvement and exploration have been identified for future work. Enhancing the models used for summarisation tasks, such as employing more advanced models like GPT-4, could improve the contextual quality of the summaries generated for question creation. Expanding the range of datasets on which the methodology is tested would help evaluate its robustness and applicability across different domains, ensuring that the approach is versatile and widely applicable. A more comprehensive qualitative evaluation could be undertaken, involving a panel of engineers with expertise in sustainable building practices and familiarity with BREEAM standards. This panel would assess the generated questions based on criteria such as relevance, clarity, complexity, coverage, and practicality. Detailed feedback from this evaluation could lead to iterative improvements, enhancing the quality and effectiveness of the competency questions. Addressing the computational demands associated with training and fine-tuning LLMs with KG-enriched datasets is another critical area for future work. Optimising resource usage and enhancing the scalability of the methodology would make it more accessible and practical for widespread implementation.

7.4. Summary

The integration of LLMs with KGs offers a promising and innovative approach to automating the generation of competency questions. This proof of concept marks a significant advancement in AI-driven educational technology, demonstrating the potential for creating more effective, scalable, and personalised competency-based assessments. The findings from this study suggest numerous possibilities for further advancements, promising to revolutionise the way competency assessments are conducted, ensuring they remain relevant, comprehensive, and tailored to individual learning needs. By continuing to refine and expand upon this methodology, there is a substantial opportunity to enhance educational and professional training outcomes across various domains.

8. Acknowledgement

This work is funded by the European Union's Horizon Europe research and innovation programme under grant agreement no 101056973 (ACCORD) and READ Diaspora Project. UK Participants in Horizon Europe Project [ACCORD] are supported by UKRI grant numbers [10040207] (Cardiff University), [10038999](Birmingham City University and [10049977] (Building Smart International).

References

[1] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, Artificial Intelligence Review 56 (2023) 13071–13102.

[2] D. Hagos, R. Battle, D. Rawat, Recent advances in generative ai and large language models: Current status, challenges, and perspectives, 2024.

[3] J. Z. Pan, S. Razniewski, J.-C. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliyanenko, W. Zhang, M. Lissandrini, R. Biswas, G. de Melo, A. Bonifati, E. Vakaj, M. Dragoni, D. Graux, Large language models and knowledge graphs: Opportunities and challenges, 2023. URL: https://arxiv.org/abs/2308.06374.

[4] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: a roadmap, IEEE Transactions on Knowledge and Data Engineering (2024) 1–20. URL: https://arxiv.org/pdf/2306.08302.pdf?trk=article-ssr-frontend-pulse_x-social-details_comments-action_comment-text. doi:10.1109/tkde.2024.3352100.

[5] R. Yang, B. Yang, S. Ouyang, T. She, A. Feng, Y. Jiang, F. Lecue, J. Lu, I. Li, Graphusion: Leveraging large language models for scientific knowledge graph fusion and construction in nlp education, 2024. URL: https://arxiv.org/html/2407.10794v1.

[6] H. Wu, Y. Yuan, L. Mikaelyan, A. Meulemans, X. Liu, J. Hensman, B. Mitra, Structured entity extraction using large language models, 2024. URL: https://arxiv.org/abs/2402.04437. doi:10.48550/arXiv.2402.04437.

[7] H. Zhu, D. Luo, X. Tang, J. Xu, H. Liu, S. Wang, Self-explainable graph neural networks for link prediction, ArXiv (Cornell University) (2023). doi:10.48550/arxiv.2305.12578.

[8] J. Baek, A. F. Aji, J. Lehmann, S. J. Hwang, Direct fact retrieval from knowledge graphs without entity linking, 2023. URL: https://arxiv.org/abs/2305.12416. doi:10.48550/arXiv.2305.12416.

[9] Z. Xu, M. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, Z. Li, Retrieval-augmented generation with knowledge graphs for customer service question an-swering, ArXiv (2024). doi:10.1145/3626772.3661370.

[10] S. Schramm, C. Wehner, U. Schmid, Comprehensible artificial intelligence on knowledge graphs: a survey, Journal of Web Semantics (2024) 100806–100806. doi:10.1016/j.websem.2023.100806.

[11] C. Wehner, C. Iliopoulou, T. R. Besold, From latent to lucid: Transforming knowledge graph embeddings into interpretable structures, 2024. URL: https://arxiv.org/abs/2406.01759. doi:10.48550/arXiv.2406.01759.

[12] A. Sawczyn, J. Binkowski, P. Bielak, T. Kajdanowicz, Empowering small-scale knowledge graphs: a strategy of leveraging general-purpose knowledge graphs for enriched embeddings, 2024. URL: https://arxiv.org/abs/2405.10745. doi:10.48550/arXiv.2405.10745.

[13] M. Trajanoska, R. Stojanov, D. Trajanov, Enhancing knowledge graph construction using large language models, 2023. URL: https://arxiv.org/abs/2305.04676. arXiv:2305.04676.

[14] K. Q. Monfardini, J. S. Salamon, M. P. Barcellos, Use of competency questions in ontology engineering: a survey, International Conference on Conceptual Modeling Cham: Springer Nature Switzerland. (2023) 45–64. doi:10.1007/978-3-031-47262-6_3.

[15] R. Alharbi, V. Tamma, F. Grasso, T. Payne, An experiment in retrofitting competency questions for existing ontologies, in: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, SAC '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1650–1658. URL: https://doi.org/10.1145/3605098.3636053. doi:10.1145/3605098.3636053.

[16] V. K. Kommineni, B. König-Ries, S. Samuel, From human experts to machines: an llm supported approach to ontology and knowledge graph construction, 2024. URL: https://arxiv.org/abs/2403.08345. doi:10.48550/arXiv.2403.08345.

[17] Y. Rebboud, L. Tailhardat, P. Lisena, R. Troncy, Can LLMs Generate Competency Questions?, in: ESWC 2024, Extended Semantic Web Conference, Hersonissos, Greece, 2024. URL: https://hal.science/hal-04564055.

[18] B. Zhang, V. A. Carriero, K. Schreiberhuber, S. Tsaneva, L. S. González, J. Kim, J. de Berardinis, Ontochat: a framework for conversational ontology engineering using language models, 2024.

URL: https://arxiv.org/abs/2403.05921.
[19] Y. He, J. Chen, H. Dong, I. Horrocks, C. Allocca, T. Kim, B. Sapkota, Deeponto: a python package for ontology engineering with deep learning, 2024. URL: https://arxiv.org/abs/2307.03067. doi:10.48550/arXiv.2307.03067.
[20] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, B. Cui, Retrieval-augmented generation for ai-generated content: a survey, 2024. URL: https://arxiv.org/abs/2402.19473. doi:10.48550/arXiv.2402.19473.
[21] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: a survey, 2024. URL: https://arxiv.org/pdf/2312.10997.
[22] S. Li, Y. He, H. Guo, X. Bu, G. Bai, J. Liu, J. Liu, X. Qu, Y. Li, W. Ouyang, W. Su, B. Zheng, Graphreader: Building graph-based agent to enhance long-context abilities of large language models, 2024. URL: https://arxiv.org/abs/2406.14550. doi:10.48550/arXiv.2406.14550.
[23] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, J. Larson, From local to global: a graph rag approach to query-focused summarization, 2024. URL: https://arxiv.org/abs/2404.16130. doi:10.48550/arXiv.2404.16130.
[24] Z. Zhong, L. Zhong, Z. Sun, Q. Jin, Z. Qin, X. Zhang, Synthet2c: Generating synthetic data for fine-tuning large language models on the text2cypher task, 2024. URL: https://arxiv.org/abs/2406.10710#:~:text=15%20Jun%202024%5D-. doi:10.48550/arXiv.2406.10710.
[25] S. Zerhoudi, M. Granitzer, Personarag: Enhancing retrieval-augmented generation systems with user-centric agents, arXiv.org (2024). doi:10.48550/arXiv.2407.09394.
[26] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: Language agents with verbal reinforcement learning, 2023. URL: https://arxiv.org/abs/2303.11366. arXiv:2303.11366.
[27] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.
[28] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, 2023. URL: https://arxiv.org/abs/2210.03629. doi:10.48550/arXiv.2210.03629.
[29] X. Amatriain, Prompt design and engineering: Introduction and advanced methods, ArXiv (2024). doi:arXiv:2401.14423.
[30] D. Cheng, S. Huang, J. Bi, Y. Zhan, J. Liu, Y. Wang, H. Sun, F. Wei, D. Deng, Q. Zhang, Uprise: Universal prompt retrieval for improving zero-shot evaluation, 2023. URL: https://arxiv.org/pdf/2303.08518.
[31] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, ArXiv (2023). doi:10.48550/arxiv.2307.03172.
[32] J. Minder, L. Brandenberger, L. Salamanca, F. Schweitzer, Data2neo - a tool for complex neo4j data integration, ArXiv (2024). doi:https://arxiv.org/abs/2406.04995.
[33] S. Sahu, A starting point for dynamic community detection with leiden algorithm, 2024. URL: https://arxiv.org/abs/2405.11658. arXiv:2405.11658.
[34] R. Vacareanu, A. Pratik, E. Spiliopoulou, Z. Qi, G. Paolini, N. A. John, J. Ma, Y. Benajiba, M. Ballesteros, General purpose verification for chain of thought prompting, 2024. URL: https://arxiv.org/abs/2405.00204. doi:10.48550/arXiv.2405.00204.
[35] T. Hu, N. Collier, Quantifying the persona effect in llm simulations, 2024. URL: https://arxiv.org/abs/2402.10811. doi:10.48550/arXiv.2402.10811.
[36] B. Leite, H. L. Cardoso, On few-shot prompting for controllable question-answer generation in narrative comprehension, arXiv (2024). doi:10.48550/arXiv.2404.02800.
[37] G. Gorgun, O. Bulut, Exploring quality criteria and evaluation methods in automated question generation: a comprehensive survey, Education and Information Technologies (2024). doi:10.1007/s10639-024-12771-3.
[38] X. Li, A. Henriksson, M. Duneld, J. Nouri, Y. Wu, Evaluating embeddings from pre-trained language models and knowledge graphs for educational content recommendation, Future Internet 16 (2023)

12–12. doi:10.3390/fi16010012.
[39] A. Bavaresco, R. Bernardi, L. Bertolazzi, D. Elliott, R. Fernández, A. Gatt, E. Ghaleb, M. Giulianelli, M. Hanna, A. Koller, A. F. T. Martins, P. Mondorf, V. Neplenbroek, S. Pezzelle, B. Plank, D. Schlangen, A. Suglia, A. K. Surikuchi, E. Takmaz, A. Testoni, Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks, 2024. URL: https://arxiv.org/abs/2406.18403.
[40] G. H. Chen, S. Chen, Z. Liu, F. Jiang, B. Wang, Humans or llms as the judge? a study on judgement biases, 2024. URL: https://arxiv.org/abs/2402.10669.
[41] Z. Kenton, N. Y. Siegel, J. Kramár, J. Brown-Cohen, S. Albanie, J. Bulian, R. Agarwal, D. Lindner, Y. Tang, N. D. Goodman, R. Shah, On scalable oversight with weak llms judging strong llms, 2024. URL: https://arxiv.org/abs/2407.04622.
[42] L. Shi, W. Ma, S. Vosoughi, Judging the judges: a systematic investigation of position bias in pairwise comparative assessments by llms, 2024. URL: https://arxiv.org/abs/2406.07791.

A. Appendices

A.1. LLMs Benchmark

Table 1
LLMs Benchmark

Model	Method	Avg. Relevance Score	Min. Score	Max. Score
GPT-4	Entities Zero-Shot	0.8354	0.7950	0.8721
GPT-4	Entities Few-Shots	0.8322	0.8062	0.8671
GPT-4	Communities Zero-Shot	0.8027	0.7465	0.8724
GPT-4	Communities Few-Shots	0.8266	0.7939	0.8615
GPT-4o	Entities Zero-Shot	**0.8501**	0.8256	0.8780
GPT-4o	Entities Few-Shots	**0.8435**	0.8068	0.8693
GPT-4o	Communities Zero-Shot	0.8239	0.7986	0.8365
GPT-4o	Communities Few-Shots	0.8258	0.7960	0.8724
GPT-4o-Mini	Entities Zero-Shot	0.8416	0.8186	0.8580
GPT-4o-Mini	Entities Few-Shots	0.8312	0.7944	0.8785
GPT-4o-Mini	Communities Zero-Shot	0.8365	0.8049	0.8622
GPT-4o-Mini	Communities Few-Shots	0.8370	0.8028	0.8675
Llama 3.1 8B	Entities Zero-Shot	0.8349	0.8118	0.8733
Llama 3.1 8B	Entities Few-Shots	0.8223	0.8013	0.8650
Llama 3.1 8B	Communities Zero-Shot	0.8148	0.7964	0.8808
Llama 3.1 8B	Communities Few-Shots	0.8099	0.7809	0.8312
Llama 3 8B	Entities Zero-Shot	0.8358	0.8035	0.8649
Llama 3 8B	Entities Few-Shots	0.8309	0.8182	0.8592
Llama 3 8B	Communities Zero-Shot	0.8135	0.7862	0.8637
Llama 3 8B	Communities Few-Shots	0.8461	0.8018	0.8785
Gemma 2 9B	Entities Zero-Shot	0.8211	0.7971	0.8656
Gemma 2 9B	Entities Few-Shots	0.8328	0.8053	0.8550
Gemma 2 9B	Communities Zero-Shot	0.8169	0.7835	0.8405
Gemma 2 9B	Communities Few-Shots	0.8626	0.8416	0.8869
Mistral-Nemo 12B	Entities Zero-Shot	0.8106	0.7271	0.8570
Mistral-Nemo 12B	Entities Few-Shots	0.8289	0.7271	0.8830
Mistral-Nemo 12B	Communities Zero-Shot	0.8172	0.7993	0.8433
Mistral-Nemo 12B	Communities Few-Shots	0.8237	0.7989	0.8582
Mistral 7B	Entities Zero-Shot	0.8435	0.8206	0.8692
Mistral 7B	Entities Few-Shots	0.8420	0.7963	0.8843
Mistral 7B	Communities Zero-Shot	**0.8759**	0.8514	0.9128
Mistral 7B	Communities Few-Shots	**0.8764**	0.8551	0.8955
Qwen 2 7B	Entities Zero-Shot	0.8323	0.8067	0.8676
Qwen 2 7B	Entities Few-Shots	0.8513	0.8183	0.8673
Qwen 2 7B	Communities Zero-Shot	0.8279	0.7912	0.8700
Qwen 2 7B	Communities Few-Shots	0.8028	0.7862	0.8228
Phi 3 14B	Entities Zero-Shot	0.8364	0.7856	0.8765
Phi 3 14B	Entities Few-Shots	0.8428	0.8206	0.8648
Phi 3 14B	Communities Zero-Shot	0.8363	0.8011	0.8778
Phi 3 14B	Communities Few-Shots	0.8448	0.8027	0.8954

A.2. Top Performers

Table 2
Top Performers

Model	Method	Avg. Relevance Score	Min. Score	Max. Score
GPT-4o	Entities Zero-Shot	0.8501	0.8256	0.8780
GPT-4o	Entities Few-Shots	0.8435	0.8068	0.8693
Mistral 7B	Communities Zero-Shot	0.8759	0.8514	0.9128
Mistral 7B	Communities Few-Shots	0.8764	0.8551	0.8955

A.3. Zero-Shot Prompt

```
prompt = f"""
    Based on the following context about Operational Waste, generate a
list of exactly 10 competency questions that would help to evaluate
understanding of the content. Each question should be numbered and on
a new line. Context: {context} Competency Questions:
"""
```

Figure 5: Zero-Shot Prompt

A.4. Few-Shot Prompt

```
few_shot_examples = """
   Context: Operational waste management focuses on minimizing
waste generation and ensuring proper disposal methods. It includes
strategies for waste reduction, segregation, recycling, and compliance
with regulatory standards
   Examples:
   1. What are the key strategies for reducing operational waste?
   2. How can waste segregation improve the efficiency of waste
management?
   3. What are the benefits of recycling operational waste?
   4. How does compliance with regulatory standards impact
operational waste management?
   5. What measures can be taken to minimize waste generation at the
source?

   Context: Effective operational waste management involves the use of
various facilities and technologies. It also addresses the environmental
and health impacts of waste, promoting sustainable practices and
resource conservation
   Examples:
   1. What facilities and technologies are commonly used in operational
waste management?
   2. How does operational waste management contribute to
environmental sustainability?
   3. What are the health impacts associated with improper waste
management?
   4. How can resource conservation be achieved through effective waste
management practices?
   5. What role does technology play in enhancing operational waste
management practices?
   """
```

Figure 6: Few-Shot Prompt

A.5. Output with Self-Reflection

Counting entities and mentions...
Total entities: 10
Total mentions: 188
Extracting top communities from the graph...

Top Community: ID 0-84, Summary: Community ID 0-84 focuses on waste
management, encompassing nodes such as Waste Hierarchy, Hazardous Waste,
Non-Hazardous Waste, Construction Waste Groups, Waste Management Routes,
European Waste Catalogue, and Procedures. The relationships within this
community show that Waste Hierarchy determines the order of Construction Waste
Groups, European Waste Catalogue, and Waste Management Routes. Construction

Waste Groups are associated and grouped with European Waste Catalogue.
Procedures aim to reduce Construction Waste Groups, Hazardous Waste, European
Waste Catalogue, Waste Hierarchy, Waste Management Routes, and Non-
Hazardous Waste.
Generating competency questions

>Entering new AgentExecutor chain...

Invoking: `generate_questions` with `{'input_data': {'context': 'Community ID 0-84 focuses on waste management, encompassing nodes such as Waste Hierarchy, Hazardous Waste, Non-Hazardous Waste, Construction Waste Groups, Waste Management Routes, European Waste Catalogue, and Procedures. The relationships within this community show that Waste Hierarchy determines the order of Construction Waste Groups, European Waste Catalogue, and Waste Management Routes. Construction Waste Groups are associated and grouped with European Waste Catalogue. Procedures aim to reduce Construction Waste Groups, Hazardous Waste, European Waste Catalogue, Waste Hierarchy, Waste Management Routes, and Non-Hazardous Waste.'}}`

First set of questions

['1. What are the primary components of the Waste Hierarchy within Community ID 0-84?', '2. How does the Waste Hierarchy influence the order of Construction Waste Groups, European Waste Catalogue, and Waste Management Routes?', '3. What are the key differences between Hazardous Waste and Non-Hazardous Waste as defined in this community?', '4. How are Construction Waste Groups categorized and associated with the European Waste Catalogue?', '5. What are the main Waste Management Routes identified in Community ID 0-84?', '6. How do Procedures aim to reduce the amount of Construction Waste Groups?', '7. In what ways do Procedures target the reduction of Hazardous Waste?', '8. How is the European Waste Catalogue utilized within the context of waste management in this community?', '9. What specific strategies are included in Procedures to minimize Non-Hazardous Waste?', '10. How do the relationships between Waste Hierarchy, Construction Waste Groups, and Waste Management Routes contribute to effective waste management?']

Invoking: `reflect_and_refine` with `{'input_data': ['1. What are the primary components of the Waste Hierarchy within Community ID 0-84?', '2. How does the Waste Hierarchy influence the order of Construction Waste Groups, European Waste Catalogue, and Waste Management Routes?', '3. What are the key differences between Hazardous Waste and Non-Hazardous Waste as defined in this community?', '4. How are Construction Waste Groups categorized and associated with the European Waste Catalogue?', '5. What are the main Waste Management Routes identified in Community ID 0-84?', '6. How do Procedures aim to reduce the amount of Construction Waste Groups?', '7. In what ways do Procedures target the reduction of Hazardous Waste?', '8. How is the European Waste Catalogue utilized within the context of waste management in this community?', '9. What specific strategies are included in Procedures to minimize Non-Hazardous Waste?', '10. How do the relationships between Waste Hierarchy, Construction Waste Groups, and Waste Management Routes contribute to effective waste management?']}`
responded: I have generated the initial set of competency questions based on the context provided. Now, I will proceed to reflect on and refine these questions. Let's refine them to ensure they are clear, relevant, and cover the key aspects of Operational Waste in Community ID 0-84.

{'reflection': '### Reflection on Original Questions\n\n#### Quality, Coverage, and Relevance\n1. **Quality**: The original questions are generally clear and specific, but some could benefit from more precise language and focus.\n2. **Coverage**: The questions cover a broad range of topics within operational waste, including the Waste Hierarchy, Construction Waste Groups, the European Waste Catalogue, and Waste Management Routes. However, they could delve deeper into specific operational practices, regulatory compliance, and technological innovations.\n3. **Relevance**:

Figure 7: Output with Self-Reflection

A.6. Output with Self-Reflection Continued

The questions are relevant to operational waste management but could be more comprehensive by including aspects like waste reduction technologies, economic impacts, and stakeholder roles.\n\n#### Gaps or Underrepresented Areas\n1. **Technological Innovations**: There is no mention of how technology can aid in waste management.\n2. **Economic Impact**: The financial aspects of waste management are not covered.\n3. **Regulatory Compliance**: The role of regulations and compliance in waste management is not addressed.\n4. **Stakeholder Roles**: The involvement of different stakeholders in waste management processes is missing.\n5. **Sustainability Metrics**: There is no focus on how waste management practices contribute to sustainability goals.\n\n#### Suggestions for Improvement\n1. **Incorporate Technological Aspects**: Include questions about the role of technology in waste management.\n2. **Economic Considerations**: Add questions that address the economic impact of waste management practices.\n3. **Regulatory Compliance**: Include questions about regulatory requirements and compliance.\n4. **Stakeholder Involvement**: Add questions that explore the roles of various stakeholders.\n5. **Sustainability Metrics**: Include questions that focus on sustainability and environmental impact.\n\n### Final Set of Refined Questions\n\n1. **What are the primary components of the Waste Hierarchy, and how do they apply to operational waste management?**\n2. **How does the Waste Hierarchy influence the categorization and management of Construction Waste Groups and Waste Management Routes?**\n3. **What are the key differences between Hazardous Waste and Non-Hazardous Waste, and how are these differences managed operationally?**\n4. **How are Construction Waste Groups categorized according to the European Waste Catalogue, and what are the implications for waste management practices?**\n5. **What are the main Waste Management Routes, and how do they contribute to effective waste management?**\n6. **What procedures and technologies are implemented to reduce the amount of Construction Waste Groups?**\n7. **What strategies and technologies are employed to minimize the generation and impact of Hazardous Waste?**\n8. **How is the European Waste Catalogue utilized to ensure compliance and efficiency in waste management practices?**\n9. **What specific strategies are included in procedures to minimize Non-Hazardous Waste, and how are these strategies evaluated for effectiveness?**\n10. **How do the relationships between the Waste Hierarchy, Construction Waste Groups, and Waste Management Routes contribute to sustainable waste management practices?**\n\n### Substantially Different Questions\n\n1. **What role do technological innovations play in enhancing waste management practices and reducing operational waste?**\n2. **How do economic factors influence the strategies and effectiveness of waste management practices?**\n3. **What are the regulatory requirements for waste management, and how do organizations ensure compliance?**\n4. **How do different stakeholders, including government, businesses, and communities, contribute to effective waste management?**\n5. **What metrics are used to evaluate the sustainability and environmental impact of waste management practices?**\n\n### Final Set of 10 Refined Questions\n\n1. **What are the primary components of the Waste Hierarchy, and how do they apply to operational waste management?**\n2. **How does the Waste Hierarchy influence the categorization and management of Construction Waste Groups and Waste Management Routes?**\n3. **What are the key differences between Hazardous Waste and Non-Hazardous Waste, and how are these differences managed operationally?**\n4. **How are Construction Waste Groups categorized according to the European Waste Catalogue, and what are the implications for waste management practices?**\n5. **What are the main Waste Management Routes, and how do they contribute to effective waste management?**\n6. **What role do technological innovations play in enhancing waste management practices and reducing operational waste?**\n7. **How do economic factors influence the strategies and effectiveness of waste management practices?**\n8. **What are the regulatory requirements for waste management, and how do organizations ensure compliance?**\n9. **How do different stakeholders, including government, businesses, and communities, contribute to effective waste management?**\n10. **What metrics are used to evaluate the sustainability and environmental impact of waste management practices?**\n\nThese refined questions provide a more comprehensive and detailed exploration of operational waste management, addressing previously identified gaps and ensuring a broader coverage of relevant topics.', 'refined_questions': []}

Invoking: `save_questions` with `{'input_data': {'questions': ['What are the primary components of the Waste Hierarchy, and how are they applied in operational waste management?', 'How does the Waste Hierarchy influence the categorization and prioritization of Construction Waste Groups and Waste Management Routes?', 'What are the key differences between Hazardous Waste and Non-Hazardous Waste, and how are these differences managed in operational practices?', 'How are Construction Waste Groups categorized according to the European Waste Catalogue, and what are the implications for waste management?', 'What are the main Waste Management Routes, and how are they selected and implemented in operational waste management?', 'What procedures are in place to reduce the amount of Construction Waste, and how effective are these procedures in practice?', 'What strategies are employed to minimize the generation of Hazardous Waste, and

Figure 8: Output with Self-Reflection Continued

A.7. Output with Self-Reflection Continued

what are the outcomes of these strategies?', 'How is the European Waste Catalogue utilized in the context of operational waste management, and what benefits does it provide?', 'What specific strategies are included in procedures to minimize Non-Hazardous Waste, and how are these strategies monitored and evaluated?', 'What are the practical challenges and solutions in implementing the Waste Hierarchy in operational waste management?']}}`

('Questions successfully saved to generated_questions.txt', ['What are the primary components of the Waste Hierarchy, and how are they applied in operational waste management?', 'How does the Waste Hierarchy influence the categorization and prioritization of Construction Waste Groups and Waste Management Routes?', 'What are the key differences between Hazardous Waste and Non-Hazardous Waste, and how are these differences managed in operational practices?', 'How are Construction Waste Groups categorized according to the European Waste Catalogue, and what are the implications for waste management?', 'What are the main Waste Management Routes, and how are they selected and implemented in operational waste management?', 'What procedures are in place to reduce the amount of Construction Waste, and how effective are these procedures in practice?', 'What strategies are employed to minimize the generation of Hazardous Waste, and what are the outcomes of these strategies?', 'How is the European Waste Catalogue utilized in the context of operational waste management, and what benefits does it provide?', 'What specific strategies are included in procedures to minimize Non-Hazardous Waste, and how are these strategies monitored and evaluated?', 'What are the practical challenges and solutions in implementing the Waste Hierarchy in operational waste management?'])

I have generated, reflected on, and refined a set of competency questions about Operational Waste in Community ID 0-84. Here are the final set of 10 refined questions:

1. What are the primary components of the Waste Hierarchy, and how are they applied in operational waste management?
2. How does the Waste Hierarchy influence the categorization and prioritization of Construction Waste Groups and Waste Management Routes?
3. What are the key differences between Hazardous Waste and Non-Hazardous Waste, and how are these differences managed in operational practices?
4. How are Construction Waste Groups categorized according to the European Waste Catalogue, and what are the implications for waste management?
5. What are the main Waste Management Routes, and how are they selected and implemented in operational waste management?
6. What procedures are in place to reduce the amount of Construction Waste, and how effective are these procedures in practice?
7. What strategies are employed to minimize the generation of Hazardous Waste, and what are the outcomes of these strategies?
8. How is the European Waste Catalogue utilized in the context of operational waste management, and what benefits does it provide?
9. What specific strategies are included in procedures to minimize Non-Hazardous Waste, and how are these strategies monitored and evaluated?
10. What are the practical challenges and solutions in implementing the Waste Hierarchy in operational waste management?

These questions cover a wide range of topics related to waste management practices in Community ID 0-84.

Figure 9: Output with Self-Reflection Continued

A.8. GPT-4o Entities Zero-Shots CQs

Table 3
GPT-4o Entities Zero-Shots CQs with LLM-as-a-Judge Feedback

N	Competency Questions	R	C	D	LLM-as-a-Judge Feedback
1	What are the specific regulatory requirements for labeling operational waste storage areas?	5	5	4	**Relevance**: Highly relevant as it addresses compliance with waste management regulations. **Clarity**: Very clear and straightforward question. **Depth**: Fairly deep, but could be more comprehensive by specifying which regulatory bodies or types of waste storage areas.
2	How can compliance with local and national waste management regulations be ensured?	5	5	5	**Relevance**: Extremely relevant to waste management practices. **Clarity**: Clear and easy to understand. **Depth**: Comprehensive as it covers both local and national regulations and implies multiple strategies for compliance.
3	What health and safety standards must be adhered to in the management of operational waste?	5	5	4	**Relevance**: Very relevant for ensuring safe waste management practices. **Clarity**: Clear and specific. **Depth**: Covers essential standards, although it could go deeper into specific standards for different types of waste.
4	How should operational waste storage areas be designed to minimize health risks to personnel?	5	5	5	**Relevance**: Highly relevant for operational safety. **Clarity**: Very clear and direct. **Depth**: Deep and specific, focusing on design aspects to minimize risks.
5	What technological solutions can be implemented to improve the efficiency of operational waste management?	5	5	4	**Relevance**: Highly relevant as technology plays a crucial role in waste management. **Clarity**: Clear and to the point. **Depth**: Fairly deep but could specify types of technology solutions (e.g., software, machinery).
6	How can data analytics be used to optimize waste collection and disposal processes?	5	5	5	**Relevance**: Very relevant, given the growing importance of data analytics. **Clarity**: Clear and specific. **Depth**: Comprehensive, focusing on the application of data analytics.
7	What training programs should be provided to staff involved in operational waste management?	5	5	4	**Relevance**: Highly relevant for effective waste management. **Clarity**: Clear and straightforward. **Depth**: Covers essential training but could delve into specific types of programs.
8	How can awareness and education about waste segregation and recycling be promoted among residents?	4	5	4	**Relevance**: Relevant, though slightly less so than operational-specific questions. **Clarity**: Very clear. **Depth**: Fairly deep but could elaborate on different strategies for promotion.
9	What measures can be taken to reduce the environmental impact of operational waste management practices?	5	5	5	**Relevance**: Extremely relevant to sustainability in waste management. **Clarity**: Very clear and direct. **Depth**: Comprehensive as it implies multiple measures for impact reduction.
10	What criteria should be used to determine the appropriate capacity for operational waste storage in different types of facilities?	5	5	5	**Relevance**: Highly relevant for effective waste management. **Clarity**: Clear and specific. **Depth**: Comprehensive as it addresses capacity determination across various facility types.

A.9. GPT-4o Entities Few-Shots CQs

Table 4
GPT-4o Entities Few-Shots CQs with LLM-as-a-Judge Feedback

N	Competency Questions	R	C	D	LLM-as-a-Judge Feedback
1	What are the essential characteristics and design requirements for spaces designated for operational waste storage and processing?	5	5	4	This question is highly relevant as it addresses fundamental aspects of waste management. It is clearly stated and focuses on specific criteria (characteristics and design requirements). However, it could be expanded to include more detailed subtopics or examples.
2	Which types of facilities and equipment are integral to effective operational waste management, and what are their specific functions?	5	5	4	This question is very relevant and clear, asking for specific information about facilities and equipment. It covers a broad range of potential answers, which is good, but it might benefit from further specificity to enhance depth.
3	How should spaces for operational waste be labeled to ensure compliance with regulatory standards and enhance operational efficiency?	4	5	3	The question is relevant and clear but somewhat narrow in focus. While labeling is important, the depth could be improved by including aspects like different types of labels, color-coding systems, or technology used in labeling.
4	Why is accessibility crucial for spaces designated for operational waste, and what best practices should be followed to ensure it?	5	5	4	This question is highly relevant and clear, addressing both the importance and the best practices for accessibility. The depth is good but could be enhanced by specifying different types of accessibility (e.g., physical, logistical).
5	What are the environmental impacts of different operational waste management practices, and how can they be mitigated?	5	5	5	This question is highly relevant, clear, and in-depth. It covers both the impacts and mitigation strategies, offering a comprehensive scope for discussion.
6	How do static waste compactors and composting vessels contribute to the efficiency and sustainability of operational waste management?	5	5	4	The question is relevant and clear, addressing specific technologies. It provides a good depth but could include more details about other types of equipment or comparative effectiveness.
7	What are the key regulatory guidelines and compliance requirements for managing operational waste, and how can organizations ensure adherence?	5	5	4	This question is very relevant and clear, dealing with critical regulatory issues. The depth is good, but it could be improved by specifying which regions or types of regulations are being referred to.
8	Why is the provision of home composting facilities significant in waste management, and what are the best practices for their implementation?	4	5	4	The question is relevant and clear, focusing on home composting. While it is slightly less relevant to operational waste management, it is still important. The depth is good and could be enhanced by discussing specific implementation strategies or challenges.
9	What considerations must be made for the design and maintenance of internal storage areas for operational waste to ensure safety and efficiency?	5	5	5	This question is highly relevant and clear, addressing both design and maintenance. It offers good depth by focusing on safety and efficiency.
10	How should waste collection vehicles be designed and operated to optimize access to internal storage areas and minimize disruption?	5	5	4	The question is highly relevant and clear, focusing on practical aspects of waste collection. The depth is good but could be enhanced by discussing specific design features or operational protocols.

A.10. Mistral Communities Zero-Shot CQs

Table 5
Mistral Communities Zero-Shot CQs with LLM-as-a-Judge Feedback

N	Competency Questions	R	C	D	LLM-as-a-Judge Feedback
1	How do waste minimization strategies affect project costs and labor productivity within a construction project?	5	4	4	**Relevance**: It is highly relevant as waste minimization is essential for reducing project costs, improving labor productivity, and promoting sustainability. **Clarity**: The question is clear but could be improved slightly by specifying whether the focus is on initial construction or the entire lifecycle of a project. **Depth**: A reasonable level of depth as it touches on cost savings and labor productivity, but further discussion on specific strategies and their impacts would add more depth.
2	What are the key factors to consider when implementing effective waste segregation practices on construction sites, and what challenges might arise?	5	4	4	**Relevance**: Waste segregation is critical for proper waste management and resource recovery. **Clarity**: The question is clear but could benefit from a more specific definition of effective waste segregation. **Depth**: A good level of depth, but further discussion on common challenges and potential solutions would add more depth.
3	In addition to the principles of the waste hierarchy, discuss other guiding frameworks for construction waste management that prioritize sustainability and resource efficiency.	5	4	3	**Relevance**: Waste segregation is critical for proper waste management and resource recovery. **Clarity**: The question is clear but could benefit from a more specific definition of effective waste segregation. **Depth**: A good level of depth, but further discussion on common challenges and potential solutions would add more depth.
4	How can a comprehensive waste management plan contribute to cost savings, improved labor productivity, and enhanced environmental performance in a construction project?	5	5	3	**Relevance**: Comprehensive waste management plans are essential for achieving sustainability goals in construction projects. **Clarity**: The question is clear and concise. **Depth**: A reasonable level of depth, but further discussion on specific strategies and their impacts would add more depth.
5	Can you elaborate on the role of lifecycle analysis in assessing the environmental impact of waste management strategies and resource use in construction projects?	5	4	4	**Relevance**: Lifecycle analysis is crucial for understanding the environmental impact of construction projects. **Clarity**: The question is clear, but it could be made more specific to focus on a particular aspect of lifecycle analysis or resource use. **Depth**: A reasonable level of depth, but further discussion on the key steps and considerations involved in lifecycle analysis would add more depth.
6	What are the best practices for hazardous waste treatment, disposal, and recycling in the context of construction and demolition activities?	5	4	4	**Relevance**: Hazardous waste management is an essential aspect of sustainable construction practices. **Clarity**: The question is clear but could benefit from a more specific definition of best practices. **Depth**: A reasonable level of depth, but further discussion on specific strategies and their effectiveness would add more depth.
7	How does the integration of circular economy principles into construction waste management plans lead to innovative and sustainable resource recovery solutions?	5	4	4	**Relevance**: The circular economy is a key concept for promoting sustainability in construction. **Clarity**: The question could be more specific about the focus of these innovative solutions. **Depth**: A reasonable level of depth, but further discussion on examples and case studies would add more depth.
8	How can waste data analysis help inform and improve waste minimization strategies and resource efficiency within a construction project lifecycle?	5	4	4	**Relevance**: Waste data analysis is essential for optimizing waste management practices in construction projects. **Clarity**: The question is clear but could benefit from a more specific definition of waste minimization strategies. **Depth**: A reasonable level of depth, but further discussion on common data analysis techniques and their impacts would add more depth.
9	What are some successful case studies that demonstrate effective partnerships between construction projects and licensed external waste contractors in promoting sustainable waste management practices?	5	5	4	**Relevance**: Case studies can provide valuable insights into best practices for collaboration. **Clarity**: The question is clear and concise. **Depth**: A reasonable level of depth, but further discussion on key takeaways from these case studies would add more depth.
10	In the context of BREEAM, how can waste management tools and guidelines be leveraged to optimize both resource efficiency and overall environmental performance across various project phases?	5	4	4	**Relevance**: BREEAM is a widely recognized tool for assessing sustainability in construction projects. **Clarity**: The question is clear but could benefit from a more specific focus on particular waste management tools and guidelines. **Depth**: A reasonable level of depth, but further discussion on the practical application of these tools and guidelines would add more depth.

A.11. Mistral Communities Few-Shot CQs

Table 6
Mistral Communities Few-Shot CQs with LLM-as-a-Judge Feedback

N	Competency Questions	R	C	D	LLM-as-a-Judge Feedback
1	What are the key components of an effective waste hierarchy implementation for efficient operational waste management in Community ID 0-84, and how do they contribute to overall efficiency?	5	5	4	**Relevance**: 5/5 **Clarity**: The question is clear, but it could be more specific about what constitutes a "key component" **Depth**: Answering this question requires knowledge of waste management practices and the waste hierarchy
2	In construction waste management, how are Construction Waste Groups defined and categorized, and why is this classification important for waste reduction strategies?	5	5	4	**Relevance**: 5/5 **Clarity**: The question is straightforward and relevant to waste management in construction **Depth**: Understanding the classification of Construction Waste Groups is crucial for waste minimization efforts, but more could be asked about specific categories or their significance
3	How does the European Waste Catalogue classify operational waste types and what implications does this have for their management in Community ID 0-84?	5	5	5	**Relevance**: The question is highly relevant to waste management practices within the specified communities **Clarity**: The question is clear and specific **Depth**: Answering this question requires knowledge of the European Waste Catalogue and its implications for waste management in the mentioned communities
4	What measures can be taken to reduce, reuse, recycle, recover, and dispose of Construction Waste Groups, Hazardous Waste, and Non-Hazardous Waste according to the procedures outlined in Community 0-84, while promoting a circular economy approach?	5	5	5	**Relevance**: 5/5 **Clarity**: The question is clear and relevant **Depth**: Answering this question requires knowledge of waste management procedures in Community 0-84 and the circular economy approach
5	In what ways does a Resource Management Plan facilitate waste minimization and circular economy principles within construction waste management practices, and how should it be developed for optimal impact?	5	5	4	**Relevance**: 5/5 **Clarity**: The question is clear but could be more specific about the impact of a Resource Management Plan **Depth**: Answering this question requires understanding of waste management practices and Resource Management Plans
6	How can the Circular Economy Concept be implemented across various waste streams, including construction waste and data center waste, in Community ID 0-84 to promote resource efficiency and circularity?	5	5	4	**Relevance**: The question is highly relevant to promoting a circular economy approach in the mentioned communities **Clarity**: The question is clear and specific **Depth**: Answering this question requires understanding of the Circular Economy Concept and its application across different waste streams
7	What are the key performance indicators for monitoring and optimizing accurate data records on waste arisings and waste management routes according to Community 0-84, and how do they promote efficient waste management and circular economy principles?	5	5	4	**Relevance**: The question is highly relevant to waste management practices within Community 0-84 **Clarity**: The question is clear and specific **Depth**: Answering this question requires understanding of key performance indicators for waste management in Community 0-84
8	In addition to reuse strategies, what other waste minimization techniques can be employed in both communities to reduce operational waste generation, promote a circular economy approach, and minimize the environmental impact of waste streams?	4	5	4	**Relevance**: The question is highly relevant to promoting a circular economy approach and minimizing environmental impact **Clarity**: The question is clear and specific **Depth**: Answering this question requires knowledge of waste minimization techniques beyond reuse strategies
9	How can data center managers optimize waste management practices for resource recovery and contribute to a circular economy approach within their facilities, while also reducing energy consumption and carbon footprint?	5	5	5	**Relevance**: While relevant, the question could be more specific about the connection between waste management and energy consumption in data centers **Clarity**: The question is clear and specific **Depth**: Answering this question requires knowledge of waste management practices in data centers and circular economy principles
10	What best practices can office managers follow when implementing waste minimization measures in recycling programs, as outlined in Community 0-84, while incentivizing employee participation, maximizing the effectiveness of these initiatives, and promoting a culture of sustainability within the workplace?	5	5	4	**Relevance**: The question is highly relevant to waste management practices in offices and promoting sustainability **Clarity**: The question is clear and specific **Depth**: Answering this question requires understanding of best practices for implementing waste minimization measures, employee incentives, and promoting a culture of sustainability in the workplace