Exploring Medium-Sized LLMs for Knowledge Base Construction

Tomás Cerveira da Cruz Pinto
University of Coimbra, CISUC/LASI, DEI
Coimbra, Portugal
tomaspinto@student.dei.uc.pt

Chris-Bennet Fleger
Hasso Plattner Institute
University of Potsdam, Germany
chris-bennet.fleger@uni-potsdam.de

Hugo Gonçalo Oliveira
University of Coimbra, CISUC/LASI, DEI
Coimbra, Portugal
hroliv@dei.uc.pt

Abstract

Knowledge base construction (KBC) is one of the great challenges in Natural Language Processing (NLP) and of fundamental importance to the growth of the Semantic Web. Large Language Models (LLMs) may be useful for extracting structured knowledge, including subject-predicate-object triples. We tackle the LM-KBC 2023 Challenge by leveraging LLMs for KBC, utilizing its dataset and benchmarking our results against challenge participants. Prompt engineering and ensemble strategies are tested for object prediction with pre-trained LLMs in the 0.5-2B parameter range, which is between the limits of tracks 1 and 2 of the challenge. Selected models are assessed in zero-shot and few-shot learning approaches when predicting the objects of 21 relations. Results demonstrate that instruction-tuned LLMs outperform generative baselines by up to four times, with relation-adapted prompts playing a crucial role in performance. The ensemble approach further enhances triple extraction, with a relation-based selection strategy achieving the highest F1 score. These findings highlight the potential of medium-sized LLMs and prompt engineering methods for efficient KBC.

1 Introduction

The creation of high-quality, machine-readable Knowledge Bases (KBs) is critical to advancements in Natural Language Processing (NLP) and Semantic Web Technologies (Weikum et al., 2021). These technologies enable us to structure information in formats that enhance accessibility and interoperability for both humans and machines. Large Language Models (LLMs) have recently emerged as powerful tools for a range of tasks, including the automation of knowledge extraction, particularly generating subject-predicate-object triples, that are fundamental components of knowledge graphs (AlKhamissi et al., 2022; Petroni et al., 2019). By contributing to the construction of structured KBs, LLMs play a crucial role in enabling semantic reasoning, querying, and web-based applications.

In this work, we investigate the potential of LLMs to automate Knowledge Base Construction (KBC) by exploring the Knowledge Bases from Pre-trained Language Models (LM-KBC) 2023 Challenge (Kalo et al., 2023). Although the study does not involve direct participation in the challenge, it was selected due to its ongoing relevance and potential for further research.

The dataset provided by the 2023 challenge includes 21 well-balanced relations, covering different topics such as geography, entertainment, or chemistry. These relations involve various categories, ensuring a wide range of diverse entities across different domains. The task is to predict an object given a subject-relation pair. For example, the subject "Andorra" and the relation *StateBordersState* should yield the set of objects "Spain, France". These subject-relation pairs are given to the LLMs to predict the corresponding set of objects.

Authors that participated in the LM-KBC 2023 Challenge had to pick one of two different tracks. Track 1 required the participants to use models with less than 1 billion parameters, whereas track 2 was open to models of any size, resulting in a choice of very powerful models, such as GPT-4 and LLaMA 2 with 70B parameters (Achiam et al., 2023; Touvron et al., 2023). As a result, there was a gap in the exploration of models with sizes close to the threshold (i.e., medium-sized), which this study aims to address, offering a valuable balance between computational efficiency and predictive power.

The 2023 challenge was chosen over the 2024 edition (Kalo et al., 2024) for several reasons. When we started this work, the proceedings of the 2024 edition had not been published yet, limiting the ability to draw insights from both the task and approaches. Furthermore, the dataset in the 2024 version comprises only 5 relations, which restricts the scope of experimentation.

Our main research goal is to explore how LLMs can enhance the automation of KBC, particularly by accurately predicting objects, ultimately forming subject-predicate-object triples. To achieve this, we define a set of subgoals.

We evaluate the performance of instruction-tuned LLMs in the prediction task, including Llama3.2-1B (Dubey et al., 2024), Gemma2-2B (Team et al., 2024), and Qwen2.5 (Bai et al., 2023), with 0.5B and 1.5B parameters. Additionally, we extended our analysis beyond instruction-tuned models to include DeepSeek-R1-Distill-Qwen-1.5B, as the DeepSeek family is revolutionizing the AI industry (Guo et al., 2025). This model falls within the same size range as our selected LLMs, making it a relevant addition for exploring its potential in the task. By analyzing these models, we aim to understand how medium-sized LLMs perform relative to both the smaller models from track 1 and the much larger models from track 2 used in the LM-KBC 2023 Challenge, offering insights into the trade-offs between model size and accuracy. Choosing not to rely on larger models offers advantages such as requiring fewer computational resources, enabling faster inference times, and potentially being run locally without the need for extensive infrastructure.

To refine knowledge extraction, we explore the impact of different prompt engineering strategies. Specifically, we investigate the impact of zero-shot and few-shot learning paradigms, designing tailored prompts for each relation type to optimize prediction accuracy. By structuring our prompts to align with the nature of each relation, we aim to improve object prediction while minimizing the need for computationally expensive fine-tuning.

In addition, we assess the effectiveness of ensemble methods in improving triple generation accuracy. We compare two ensemble strategies: relation-based model selection, which assigns the best-performing model for each relation, and majority voting, which selects the most frequently predicted object across models. By leveraging the complementary strengths of different LLMs, we aim to determine whether ensemble techniques provide a significant advantage over individual model predictions.

The main contributions of this work are summarized as follows:

- We investigate medium-sized models for KBC, offering a balance between computational efficiency and performance.
- We explore the integration of prompt engineering techniques, including relation-specific prompts and contextual enrichment, leveraging the strengths of instruction-tuned LLMs in enhancing task adaptability.
- We explore the synergistic potential of model ensembles combining the strengths of different models to improve overall performance.

This paper is structured as follows: In Section 2, we review findings in the field of KB construction using LLMs, focusing on the contributions from the 2023 LM-KBC Challenge. Section 3 outlines our methodology, including model selection, prompting techniques, and ensemble methods. In Section 4, we present our results, followed by a discussion in Section 5, where we discuss our findings, comparing them to prior work and highlighting key trends and limitations. Finally, Section 6 concludes the paper, summarizing the contributions and suggesting directions for future research. The code and experimental results of the study are available at https://github.com/TomasCCPinto/ldk25-medium-llms-kbc.

2 Related Work

In recent years, the construction of high-quality, machine-readable KBs has increasingly leveraged LLMs (Petroni et al., 2019), marking a paradigm shift from traditional dependence on structured data sources like Wikidata[1] (Vrandečić and Krötzsch, 2014) to models such as GPT-4 (Achiam et al., 2023), BERT (Devlin et al., 2019), and Llama 3 (Dubey et al., 2024). This shift has spurred significant progress in automating KBC, particularly in extracting structured subject-predicate-object triples directly from unstructured text. These efforts are exemplified by benchmarks like the LM-KBC Challenges, which have provided a compre-

[1] https://www.wikidata.org/

hensive framework for evaluating these capabilities.

2.1 LM-KBC Challenges and Their Evolution

The LM-KBC Challenges, introduced by Singhania et al. (2022), provide a framework for evaluating the ability of LLMs to generate accurate knowledge triples directly from their parameters by predicting the object(s) given a subject and a relation. For example, typical task instances might involve predicting "Nobel Prize in Physics" as the object given the subject "Albert Einstein" and the relation *PersonHasNoblePrize* or predicting "Spain" given the subject "Portugal" and the relation *CountryBordersCountry*. They emphasize extracting unique Wikidata entity identifiers, handling variable cardinalities, and resolving ambiguities, such as distinguishing between entities like "Paris, France" and "Paris, Texas."

The 2023 iteration refined this framework by dividing the task into two tracks based on model size, below and above 1 billion parameters, and incorporating complex relations (Kalo et al., 2023). Smaller models achieved respectable results through advanced prompt engineering and retrieval-based enrichment, while larger models consistently outperformed due to their capacity for richer contextual representations. Despite these advances, both tracks highlighted ongoing challenges, including difficulties with disambiguation, reliance on domain-specific training, and the necessity of extensive post-processing.

2.2 Commonalities and Innovations in Recent Approaches

The LM-KBC 2023 Challenge catalyzed a wide array of methodologies aimed at addressing the nuances of KBC.

Prompt Engineering and Context Enrichment were widely employed to align LLM outputs with the task objectives. High-performing approaches, such as LLMKE (Zhang et al., 2023), the winners of track 2, adopted multi-stage prompting strategies, including question-based prompts, triple completion, and context-enriched inputs incorporating entity information. Similarly, Li et al. (2023) utilized prompts enriched with Wikidata information related to the given relation. A strong emphasis was placed on crafting detailed task instructions, with some works incorporating task demonstrations (Biester et al., 2023), while others deliberately avoided demonstrations to test the limits of instruction-only setups (Ghosh, 2023).

Fine-tuning further boosted the performance of the models. For instance, the winners of the first track enriched their approach by fine-tuning BERT on the challenge's training set, in addition to pretraining it on a task-specific Wikipedia corpus (Yang et al., 2023). Additionally, Biswas et al. (2023) fine-tuned BERT's representations to align with a Wikipedia-derived entity embedding space, enabling the handling of multi-token entities and Wikidata ID linking.

Post-processing and Cleaning pipelines, such as entity validation and output reformatting, played a crucial role in improving object extraction, as LLM-generated responses often deviate from the expected output format. For example, the system by Li et al. (2023) implemented de-duplication and a Wikidata-based disambiguation process, improving precision and recall for challenging relations such as *PersonHasAutobiography*. Similarly, Ghosh (2023) employed manually designed cleaning steps, including linking extracted terms to Wikidata entities, disambiguating ambiguous objects, and applying relation-specific adjustments to ensure output conformity. While these techniques increase system complexity and require manual intervention, they proved highly effective.

2.3 Gaps in Existing Approaches

Despite impressive progress, existing approaches to KBC exhibit notable limitations. Most efforts have focused on either small models (under 1 billion parameters) or really large models (exceeding 70 billion parameters), leaving a gap in exploring models with intermediate parameter sizes. These models could offer a balance between computational efficiency and predictive performance, yet their potential remains under-investigated.

Apart from that, few methods explore the synergistic potential of model ensembles. Most focus on optimizing individual models, leaving untapped opportunities for leveraging diverse model strengths.

Finally, few studies systematically compare the performance of the same model across different parameter sizes. By using Qwen2.5 with 0.5B and 1.5B parameters, we aim to address this gap, providing insights into how scaling parameters impact a model's ability to handle diverse relations.

These gaps motivate the need for methodologies that balance computational efficiency with robust performance across diverse settings.

Our approach aims to address these gaps by integrating instruction-tuned LLMs with prompt engineering and ensemble strategies. By leveraging relation-specific prompts and lightweight contextual enrichment, we optimize the adaptability of medium-sized models. Furthermore, our use of relation-based and majority-voting ensembles allows us to harness the complementary strengths of different models.

3 Methodology

In this section, we present a comprehensive overview of the methodology employed to address our study. This includes a detailed description of each phase of the work, from the selection and preparation of the LLMs to the extraction of knowledge.

3.1 Dataset

The dataset used was provided as part of the LM-KBC 2023 Challenge and follows the object prediction format described in Section 2.1, serving as the primary foundation for evaluation. This dataset was specifically designed to assess object prediction accuracy and contained 21 distinct relations, offering a diverse set of subjects and their associated ground-truth objects. For example, the dataset encompasses relations such as *CountryHasStates*, *PersonPlaysInstrument*, and *SeriesHasNumberOfEpisodes*, capturing a wide range of knowledge domains. Each relation includes a maximum of 100 unique subject entities across all data splits, with 17 of the relations achieving this maximum, while the remaining 4 relations feature approximately 60 subject entities each.

The object entities in the dataset cover a broad range of categories, including individuals (e.g., people), organizations, countries, counts, and in some cases, the placeholder "none" to signify the absence of a valid object.

A key feature of the dataset is its reliance on ground-truth identifiers from Wikidata, ensuring accurate disambiguation of object entities. These identifiers serve as precise references for evaluating model predictions, reducing the ambiguity inherent in natural language.

3.2 Model Selection

We selected four instruction-tuned LLMs: Llama3.2[2] (1B parameters), Gemma2[3] (2B), and Qwen2.5[4,5](0.5B and 1.5B). These models balanced advanced capabilities with computational feasibility, optimizing performance within the constraints of available hardware.

Initially, we planned to use the non-instruction-tuned versions of these models but found them limited in generating concise, accurate predictions or following prompt instructions, even with advanced techniques like zero-shot prompting and few-shot learning. Instruction tuning significantly enhances their ability to handle complex, task-specific queries, making them, in our view, well-suited for the tasks in the LM-KBC 2023 Challenge.

Despite the aforementioned concerns regarding performance inconsistencies, we further decided to consider DeepSeek-R1-Distill-Qwen-1.5B[6] in experimentation as well. This model is a distilled version based on a mathematical Qwen2.5 model, fine-tuned using outputs generated by DeepSeek-R1 and incorporating slight changes to the model configuration and tokenizer. Even though it is not an instruction-tuned focused model like the others, it falls within the model size range being explored and is part of a rapidly evolving model family that is gaining prominence in the AI landscape. This made it an interesting candidate to explore and compare for investigation purposes.

3.2.1 Setup adaptation

Our implementation is built upon the baseline setup provided by the LM-KBC 2023 Challenge organizers, which utilizes the Transformers library from Hugging Face[7]. While effective, this setup required several modifications to suit the autoregressive models and optimize performance.

We adjusted the generation process to ensure proper handling of the models and refined the post-processing pipeline for cleaner, more accurate results, addressing limitations in the baseline's ap-

[2] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

[3] https://huggingface.co/google/gemma-2-2b-it

[4] https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct

[5] https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct

[6] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

[7] https://huggingface.co/

proach. Additionally, we optimized GPU utilization, reducing runtime and improving efficiency for these large models.

The baseline's evaluation script was retained, as it demonstrated reliability in calculating metrics such as precision, recall, and F1 score.

The experiments were executed either on Google Colab using the freely available NVIDIA T4 GPU, or performed on a local machine that is equipped with an Apple M2 chip with integrated 10-core GPU and 16 GB RAM. This setup ensured sufficient computational resources for running inference efficiently.

3.2.2 Response Cleaning

Response cleaning was essential for the models to meet evaluation requirements focused on predicted objects. These models often included input fragments or extra text, requiring automatic removal to isolate object entities. For example, removing strings like "answer:" ensured cleaner outputs.

For specific relations, such as *PersonHasNumberOfChildren*, numerical responses were converted to strings to match ground-truth formats. For multi-object relations like *CountryHasStates*, strings were split into individual entities for accurate evaluation.

These steps ensured proper formatting and preserved the accuracy of extracted triples, making the model outputs suitable for evaluation.

3.3 Prompt Engineering

Prompt engineering was the central approach used to adapt the selected LLMs to the specific task of object prediction. Rather than fine-tuning the models, we focused on crafting and optimizing the prompts to guide the models in generating accurate and relevant responses. Our approach is similar to that of Ghosh (2023), who also emphasized prompt engineering to align LLM outputs with task-specific objectives, demonstrating its potential as a lightweight alternative to more resource-intensive strategies.

3.3.1 Relation Adapted Prompts

Similar to the work by Nayak and Timmapathini (2023), in zero-shot learning settings, we designed a distinct prompt for each relation, tailoring the instructions to align with the specific requirements of the relation. While the baseline setup provided a basic question template for each relation, our approach went further by appending instruction information to increase the likelihood of correctly formatted results. Figure 1 demonstrates how an example input is composed of these two parts for the relation *BandHasMember*.

Question Part: Who are the members of *{subject_entity}*?
Instruction Part: List only the members, separated by ", " with no extra text.

Example Input: Who are the members of The Beatles? List only the members, separated by ", " with no extra text.
Example Output: **John Lennon**, **Paul McCartney**, **George Harrison**, **Ringo Starr**

Figure 1: Example of a Relation-Specific Zero-Shot Prompt for Relation *BandHasMember*. The first box shows the template while the second box demonstrates the instantiation.

With our additional instruction information, we can handle special characteristics for each relation. For instance, some relations, such as *SeriesHasNumberOfEpisodes*, require numerical responses as objects, while others like *PersonHasSpouse* typically expect a single answer. Additionally, certain relations involve multiple possible answers (e.g., *CountryHasStates*), or may even allow for the possibility of no answer at all (e.g., *PersonCauseOfDeath* if the individual has not passed away). Table 5 in Appendix A.3 shows all of our zero-shot question prompts.

3.3.2 Few-shot Prompting

In addition to relation-specific zero-shot prompts, we designed few-shot question and triple prompts to further explore LLM performance. Few-shot prompts were composed of a task explanation, n randomly selected examples from the training set (formatted either as questions or triples), and the target task. Figure 2 shows an example of our few-shot prompting technique using the triple template for the relation *PersonPlaysInstrument*.

The examples provided for a given instance belong to the same relation as that instance. Moreover, three examples were always used, thus following a three-shot prompting approach. Our triple prompt template followed a structured format that explicitly included the subject entity and relation, followed by the expected object. The question prompt template used the questions presented in table 5 in Appendix A.3. This approach aimed to leverage

Task Explanation: “Please answer the question with your knowledge. Beforehand there are a few examples. The output format should be a list of possible answers prefaced by "Answer: ", also if there is no answer write Answer: ["]”

Demonstrations:
“Bo Burnham, PersonPlaysInstrument: piano”
“Marko Topchii, PersonPlaysInstrument: guitar”
“Leo Pellegrino, PersonPlaysInstrument: saxophone”

Question:
“Kevin Pabst, PersonPlaysInstrument: **trumpet**”

Figure 2: Few-Shot Prompting. The prompt consists of a task explanation, three demonstrations, and the target task formulated as a triple.

LLMs’ ability to generalize from provided examples while maintaining consistency across formats.

3.3.3 Subject Context

To address potential ambiguities in subject entities, we enhanced the zero-shot prompts by including contextual information about the subject entity. Specifically, we append the first paragraph from the Wikipedia[8] page associated with the entity’s identifier in the prompt. The information is introduced using a “Context:” prefix and placed before the question part of the prompt shown in the format previously illustrated, in Figure 1. This additional context aims to clarify which specific subject the model should consider when generating predictions.

For example, if the subject in question is Leonid Volkov, ambiguity arises as there are multiple notable individuals with that name: a politician, an ice hockey player, and a skydiver. By including the introductory paragraph from the Wikipedia page, the model can better differentiate between these individuals and generate more accurate object predictions.

3.4 Model Ensemble

A key aspect of our project was developing a model ensemble approach to combine the strengths of the selected LLMs. Each model demonstrated varying performance across different types of relations, making an ensemble strategy a promising way to enhance overall accuracy.

We combined three models, Gemma2, Qwen2.5-1.5B, and Llama3.2, selected based on their F1 scores on the training dataset, which followed the same format as the evaluation set. This allowed us to assess the models’ performance in a comparable setting and identify the top-performing models for inclusion in the ensemble. Tables 2 and 3 in Appendix A.1 present the detailed results on the training dataset following zero-shot settings.

We implemented two ensemble strategies: Relation-Based Ensemble and Majority Voting Ensemble. The first selected the best-performing model for each relation based on F1 scores from the training set. For instance, if Llama3.2 excelled at predicting *PersonHasProfession* but Gemma2 performed better on *PersonPlaysInstrument*, the outputs from the respective best-performing models were combined in the final results. This dynamic selection process allowed the ensemble to adapt to different relation types effectively.

In the Majority Voting Ensemble, the models’ outputs were compared, and the most frequently predicted object(s) were chosen as the final answer. If no majority agreement was reached, the fallback response came from the model with the highest F1 score for the specific relation on the training set, increasing the likelihood of selecting the correct output.

This ensemble approach leveraged the complementary strengths of the models, improving both precision and recall across diverse relation types.

4 Results

Our experiments demonstrate notable progress in using instruction-tuned LLMs for KBC, achieving results far exceeding those of generative baselines like GPT-3. Specifically, our best setting delivered up to four times better performance, highlighting the effectiveness of prompt engineering and contextual enhancements. Table 1 summarizes the F1 scores for all of our approaches, the baselines, and the best approaches in track 1 and 2 of the challenge.

Based on these results, we make several observations. Regarding the performance of individual models:

- Gemma2 showed the best performance across most configurations, especially in the *0-shot + paragraph context* configuration, where it achieved an F1 score of 0.377.

- Llama3.2 exhibited significantly lower performance across configurations, with its highest

[8] https://www.wikipedia.org/

Model	Method	P	R	F1
BERT	Baseline	0.368	0.161	0.142
GPT-3	Baseline	0.126	0.060	0.061
VE-BERT	Winner of track 1 (Yang et al., 2023)	0.395	0.393	0.323
LLMKE	Winner of track 2 (Zhang et al., 2023)	0.715	0.726	0.701
Llama3.2 1B	0-shot	0.184	0.314	0.193
	0-shot + paragraph context	**0.258**	**0.401**	**0.271**
	3-shot question	0.185	0.237	0.153
	3-shot triple	0.295	0.329	0.268
Gemma2 2B	0-shot	0.279	0.336	0.259
	0-shot + paragraph context	**0.394**	**0.443**	**0.377**
	3-shot question	0.319	0.288	0.223
	3-shot triple	0.263	0.280	0.260
Qwen2.5 0.5B	0-shot	0.116	0.174	0.115
	0-shot + paragraph context	0.170	0.264	0.175
	3-shot question	0.119	0.208	0.106
	3-shot triple	**0.214**	**0.264**	**0.188**
Qwen2.5 1.5B	0-shot	0.187	0.257	0.188
	0-shot + paragraph context	**0.286**	**0.350**	**0.281**
	3-shot question	0.219	0.214	0.166
	3-shot triple	0.206	0.192	0.189
DeepSeek-R1 1.5B	0-shot	0.056	0.107	0.057
	0-shot + paragraph context	0.057	0.107	0.057
	3-shot question	0.100	0.170	0.068
	3-shot triple	**0.091**	**0.197**	**0.093**
Ensemble	0-shot + relation-based	0.348	0.412	0.334
	0-shot + majority voting	0.344	0.408	0.331
	0-shot + relation-based + paragraph context	**0.395**	**0.453**	**0.384**
	0-shot + majority voting + paragraph context	0.392	0.451	0.381

Table 1: Average Precision (P), Recall (R), and F1 Score (F1) for Each Model and Method.

F1 score being 0.271, achieved in the *0-shot + paragraph context* configuration, closely followed by the *3-shot triple* configuration.

- Qwen2.5 models displayed generally divergent performance, with the 1.5B model also achieving its best performance of F1 = 0.281 in the *0-shot + paragraph context* configuration. The 0.5B model consistently performed worse, with its highest F1 score of 0.188.
- DeepSeek-R1 showed the weakest performance, with an best F1 score of just 0.093, far below the other models and close to the GPT-3 baseline on most configurations, indicating significant limitations in this task.

Specifically on the ensemble methods:

- There were slight improvements over individual models, but the ensembles did not largely surpass the best individual model (Gemma2). The highest ensemble F1 score was 0.384, achieved by *0-shot + paragraph context + relation-based* prompting.
- The relation-based ensemble outperformed majority voting by less than 0.01 points.

Finally, on the performance of different types of prompts:

- *0-shot + paragraph context* consistently outperformed other configurations for most models, particularly for Gemma2, which exhibited the highest F1 scores of all individual models.
- *3-shot question* prompts were the least effective across models, exhibiting a notable decline in performance relative to other configurations.

Relation-specific performance varied widely, as shown in Appendix A.2 Table 4, which reports precision, recall, and F1 score for each relation under our best-performing configuration: *ensemble 0-shot + relation-based + paragraph context*. High-performing relations included *CountryBordersCountry* and *RiverBasinsCountry*. These relations likely benefit from their structured representations and prominence in KBs. Conversely, *PersonHasAutobiography*, *StateBordersState*, and others consistently exhibited lower F1 scores, reflecting challenges like data sparsity and ambiguity in text representations.

5 Discussion

The results of this study provide valuable insights into the performance of various language models and prompting strategies for knowledge-based tasks. This section aims to provide a detailed interpretation of the findings, highlighting connections to existing research and discussing potential areas for improvement.

5.1 Model Performance and Comparisons

The experimental results reveal significant differences in performance among the tested models. Our largest model, Gemma2 2B, consistently outperformed all other models, achieving its highest F1 score of 0.377. This performance highlights the model's ability to leverage structured input effectively, aligning with previous studies emphasizing the role of context in improving task performance for large models.

In contrast, Qwen2.5 0.5B performed poorly, with its best F1 score being only 0.188, highlighting that structured triple-based prompting was relatively more effective for this smaller model, compared to standard question-based prompts. Its underwhelming results suggest limitations in its capacity to process and utilize contextual information as effectively as larger models like Gemma2. These findings support observations in the literature that smaller models struggle with tasks requiring fine-grained reasoning and complex information extraction.

Despite the large size of the Qwen2.5 1.5B and Llama3.2, they achieved F1 scores of 0.281 and 0.271, respectively, failing to match Gemma2. This underscores that model size alone is not sufficient to guarantee high performance. Architectural differences, training data quality, and task-specific optimizations likely contributed to the performance gap.

DeepSeek-R1 performed notably worse than the other models, achieving an F1 score of only 0.093. This poor performance was expected, given that it is not instruction-tuned, making it significantly less capable of following structured prompts and generating predictions in the required format. The model struggled to adhere to our task instructions, often producing incoherent or incorrectly formatted outputs. Its behavior supports our initial idea of not using the base versions of the other models tested, opting for instruct versions. Given that DeepSeek is part of a rapidly evolving model family, larger-scale or future instruction-tuned versions are likely to yield more competitive results.

5.2 Effectiveness of Ensemble Approaches

The most successful ensemble configuration, *0-shot + paragraph context + relation-based* prompting, achieved an F1 score of 0.384. While this achieved the highest total F1 score, it resulted in only a modest performance increase of 0.007 points compared to the individual performance of Gemma2. The similarity in results between the two ensemble methods indicates that both strategies were effective in leveraging model diversity. However, when model predictions diverge significantly, majority voting often defaults to the fallback strategy, selecting the best-performing model per relation, thereby approximating the behavior of the relation-based ensemble.

An important observation is that the effectiveness of an ensemble depends significantly on the relative performance of its constituent models. When one model, such as Gemma2, substantially outperforms the others, the ensemble tends to rely predominantly on that model's outputs across all relations. As a result, the ensemble offers limited improvements, as it essentially mirrors the strongest individual model.

Conversely, when models have more comparable performances (as observed in ensembles without paragraph context), the ensemble is better able to leverage the strengths of each model, with a performance increase of approximately 0.075 points of the best individual model. In such cases, the ensemble captures complementary knowledge and yields a more significant performance boost from the individuals by integrating the "good predictions" from all models.

This finding aligns with prior research suggesting that ensemble methods, while generally robust, require careful calibration to achieve significant performance gains (Biester et al., 2023). The modest improvements seen here highlight the need for further exploration into ensemble techniques, such as dynamic weighting or neural blending, to better harness the complementary strengths of individual models.

5.3 Insights from Prompting Strategies

The comparative analysis of prompting strategies revealed unexpected yet insightful patterns. Specifically, the *3-shot question* prompts configuration exhibited the weakest performance across most of the models. For instance, Llama3.2 recorded an F1 score of only 0.153 in this configuration, a significant performance drop compared to the *0-shot + paragraph context* or the *3-shot triple* prompts.

At first glance, this result seems counterintuitive, as one might expect the inclusion of examples in the prompts would enhance the model's performance by demonstrating the task more concretely. However, the discrepancy is attributable to the design of the prompts. The zero-shot prompts were carefully crafted with task instructions tailored specifically to each individual relation, ensuring the model was provided with precise, context-relevant guidance.

In contrast, the three-shot prompts relied heavily on the demonstrations to fulfill the task. Since answers for instances of the same relation can slightly vary, as for example, in terms of the number of answers or even the absence of an answer, performance may be affected without additional instructions. While the *3-shot triple* configuration could also be affected by similar variations, it was able to provide better results possibly because the triple format inherently offered a clearer and more straightforward way to present the relationship between entities. This structure likely minimized ambiguity, allowing the model to better understand the task and produce more accurate responses. This reinforces the importance of prompt structure in reducing confusion and enhancing model performance, especially in few-shot settings.

For zero-shot, we are aware that including the paragraph from Wikipedia may occasionally provide hints toward the correct answer in some instances. However, we do not see it as a threat to the experimentation goals, as the disambiguation benefits can be significant. Furthermore, the results suggest that the contextual grounding provided by paragraph-enhanced prompts significantly mitigated the need for examples, yielding the best results. This reinforces findings in the literature, where carefully designed zero-shot instructions have been shown to outperform few-shot approaches, particularly when the latter lacks alignment with the task's domain (Kojima et al., 2022).

5.4 Comparison with Participants in LM-KBC 2023

We compare our results with the performance of the participants in both tracks 1 (small model) and 2 (no limit) of the LM-KBC 2023 Challenge. This decision stems from the fact that the models we selected, although formally eligible for track 2, are still near the 1B parameter threshold.

We note that we could outperform the best result of track 1, showcasing the effectiveness of our methodology and the benefits of using slightly larger models. Another source of improvement may stem from the use of more recent models that were not available in 2023. Given the rapid progress in language model development, advances in pretraining and other techniques could also contribute to better model performance.

When comparing our results to those in track 2, the superior performance of larger models like GPT-3.5 Turbo and GPT-4 is unsurprising, given their substantial parameter count advantage. Also, some track 2 participants (Zhang et al., 2023; Nayak and Timmapathini, 2023) boosted performance by injecting vast Wikipedia knowledge directly into prompts. While effective, this raises concerns about whether the approaches are truly assessing the models' ability to extract knowledge on their own. Infoboxes and Wikidata triples, as used by the winners, already contain structured answers to many subject-relation pairs. However, since LLMs are already pre-trained on similar data, these concerns might be somewhat alleviated.

Our results demonstrate a strong balance between efficiency and effectiveness, achieving competitive performance. This reinforces the idea that strategic adaptations and well-tuned approaches can deliver meaningful outcomes even with limited computational resources.

6 Conclusion

This study explored the use of LLMs for KBC, focusing on their ability to predict object entities within the context of the LM-KBC 2023 Challenge. Through a systematic evaluation of multiple models, mostly instruction-tuned, and leveraging techniques such as prompt engineering and ensemble methods, we derived several key insights.

Our best F1 score, which stems from our ensemble configuration, is higher than that of the winner of track 1, proving the effectiveness of our approach.

Furthermore, we observe a strong correlation between parameter size and model performance, also within the Qwen2.5 model itself. We also see performance differences based on the prompting method used: Zero-shot prompting, tailored to each relation, achieved superior results compared to few-shot approaches. Triple prompts consistently outperformed our question prompts. In addition, contextual enhancements, particularly through paragraph-level information, proved critical in improving F1 scores across all models, demonstrating the value of incorporating external knowledge.

Ensemble techniques, when one model is clearly dominant, marginally improved performance. The limited gains suggest further refinement is needed to enhance effectiveness.

Our work contributes to the growing body of research on Natural Language Processing and Semantic Web Technologies, demonstrating the viability of medium-sized LLMs for efficient KBC. By achieving results proportionally competitive with those of larger models under resource constraints, we underscore the value of methodological innovation over raw computational scale. However, despite our methods showing promising results, they still fall short of the standards required for robust KBC. This underscores that LLMs, in their current state, are not yet capable of replacing structured KBs, but rather complement them.

To build on our work, future research could focus on incorporating contextual knowledge into 3-shot prompts and exploring their use within ensemble models. Additionally, investigating advanced ensemble techniques, such as dynamic weighting or neural blending, as well as leveraging larger, more diverse datasets, could significantly enhance LLM performance in KBC.

Acknowledgements

This work is financed through national funds by FCT - Fundação para a Ciência e a Tecnologia, I.P., in the framework of the Project UIDB/00326/2025 and UIDP/00326/2025; and by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Fabian Biester, Daniel Del Gaudio, and Mohamed Abdelaal. 2023. Enhancing knowledge base construction from pre-trained language models using prompt ensembles. In *KBC-LM/LM-KBC@ ISWC*.

Debanjali Biswas, Stephan Linzbach, Dimitar Dimitrov, Hajira Jabeen, and Stefan Dietze. 2023. Broadening BERT vocabulary for knowledge graph construction using Wikipedia2Vec. In *KBC-LM/LM-KBC@ ISWC*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shrestha Ghosh. 2023. Limits of zero-shot probing on object prediction. In *KBC-LM/LM-KBC@ ISWC*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jan-Christoph Kalo, Tuan-Phong Nguyen, Simon Razniewski, and Bohui Zhang. 2024. Preface: LM-KBC challenge 2024. In *2nd Workshop on Knowledge Base Construction from Pre-Trained Language Models*. CEUR. ws.

Jan-Christoph Kalo, Sneha Singhania, Simon Razniewski, Jeff Z Pan, et al. 2023. LM-KBC 2023: 2nd challenge on knowledge base construction from pre-trained language models. In *Joint proceedings of 1st workshop on Knowledge Base Construction from Pre-Trained Language Models (KBC-LM) and the 2nd challenge on Language Models for Knowledge Base Construction (LM-KBC)*, volume 3577.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Xue Li, Anthony James Hughes, Majlinda Llugiqi, Fina Polat, Paul Groth, Fajar J Ekaputra, et al. 2023. Knowledge-centric prompt composition for knowledge base construction from pre-trained language models. In *KBC-LM/LM-KBC@ ISWC*.

Anmol Nayak and Hari Prasad Timmapathini. 2023. LLM2KB: constructing knowledge bases using instruction tuned context aware large language models. *arXiv preprint arXiv:2308.13207*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Sneha Singhania, Tuan-Phong Nguyen, and Simon Razniewski. 2022. LM-KBC: Knowledge base construction from pre-trained language models. *the Semantic Web Challenge on Knowledge Base Construction from Pre-trained Language Models*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Gerhard Weikum, Xin Luna Dong, Simon Razniewski, Fabian Suchanek, et al. 2021. Machine knowledge: Creation and curation of comprehensive knowledge bases. *Foundations and Trends® in Databases*, 10(2-4):108–490.

Dong Yang, Xu Wang, and Remzi Celebi. 2023. Expanding the vocabulary of BERT for knowledge base construction. *arXiv preprint arXiv:2310.08291*.

Bohui Zhang, Ioannis Reklos, Nitisha Jain, Albert Meroño Peñuela, and Elena Simperl. 2023. Using large language models for knowledge engineering (LLMKE): A case study on wikidata. *arXiv preprint arXiv:2309.08491*.

A Appendix

A.1 Results on the Training Dataset

To leverage our ensemble method, we selected the top three models based on their performance on the training dataset. Table 2 presents results without contextual knowledge, while Table 3 includes prompts enriched by Wikipedia paragraphs. Llama3.2 1B, Gemma2 2B, and Qwen2.5 1B were chosen for our ensemble.

Model	P	R	F1
Llama3.2 1B	0.176	0.313	**0.188**
Gemma2 2B	0.299	0.371	**0.290**
Qwen2.5 0.5B	0.116	0.179	0.116
Qwen2.5 1.5B	0.185	0.247	**0.182**
DeepSeek-R1 1.5B	0.060	0.108	0.060

Table 2: Scores for the zero-shot question setting on the training dataset.

Model	P	R	F1
Llama3.2 1B	0.255	0.411	**0.270**
Gemma2 2B	0.399	0.445	**0.383**
Qwen2.5 0.5B	0.169	0.256	0.173
Qwen2.5 1.5B	0.298	0.358	**0.291**
DeepSeek-R1 1.5B	0.062	0.109	0.060

Table 3: Scores for the zero-shot question plus paragraph context setting on the training dataset.

A.2 Relation-specific performance

Table 4 presents the precision, recall, and F1 score for each relation using our best-performing method, the relation-based ensemble with *0-shot + paragraph context*. The results show significant variability in performance across different relations.

Relation	**P**	**R**	**F1**
BandHasMember	0.407	0.367	0.370
CityLocatedAtRiver	0.345	0.366	0.343
CompanyHasParentOrganisation	0.280	0.715	0.277
CompoundHasParts	0.402	0.416	0.404
CountryBordersCountry	0.727	0.786	0.739
CountryHasOfficialLanguage	0.615	0.704	0.615
CountryHasStates	0.303	0.149	0.185
FootballerPlaysPosition	0.280	0.648	0.358
PersonCauseOfDeath	0.680	0.680	0.680
PersonHasAutobiography	0.112	0.120	0.114
PersonHasEmployer	0.202	0.256	0.206
PersonHasNoblePrize	0.130	0.510	0.130
PersonHasNumberOfChildren	0.270	0.210	0.210
PersonHasPlaceOfDeath	0.495	0.495	0.495
PersonHasProfession	0.303	0.274	0.261
PersonHasSpouse	0.320	0.320	0.320
PersonPlaysInstrument	0.440	0.473	0.433
PersonSpeaksLanguage	0.602	0.768	0.646
RiverBasinsCountry	0.899	0.746	0.789
SeriesHasNumberOfEpisodes	0.305	0.310	0.307
StateBordersState	0.173	0.199	0.175
Average	**0.395**	**0.453**	**0.384**

Table 4: Precision (P), Recall (R), and F1 score per relation for the best result.

A.3 Prompt Templates

We crafted input prompts for zero-shot and few-shot prompting settings. Few-shot used either triples, as shown in Figure 2, or the question parts presented in Table 5. Zero-shot prompts used both the question and instruction parts.

Relation Name	Question Part	Instruction Part
Band Has Member	Who are the members of {subject_entity}?	List only the members, separated by ’, ’ with no extra text.
City Located At River	Which river is {subject_entity} located at?	List only the river(s), separated by ’, ’ with no extra text.
Company Has Parent Organisation	What is the parent organization of {subject_entity}?	Answer with the parent organization only or respond with ” if none, with no extra text.
Country Borders Country	Which countries border {subject_entity}?	List only the countrie(s), separated by ’, ’ with no extra text.
Country Has Official Language	What is the official language of {subject_entity}?	List only the language(s), separated by ’, ’ with no extra text.
Country Has States	Which states are part of {subject_entity}?	List only the states / provinces, separated by ’, ’ with no extra text.
Footballer Plays Position	What position does {subject_entity} play in football?	Provide the position(s), separated by ’, ’ with no extra text.
Person Cause Of Death	What caused the death of {subject_entity}?	Provide only the cause, or respond with ” if unknown, with no extra text.
Person Has Autobiography	What is the title of {subject_entity}’s autobiography?	Answer with the title, with no extra text.
Person Has Employer	Who is {subject_entity}’s employer?	List only the employer(s), separated by ’, ’ with no extra text.
Person Has NoblePrize	In which field did {subject_entity} receive the Nobel Prize?	Answer with the field only, or ” if none, with no extra text.
Person Has Number Of Children	How many children does {subject_entity} have?	Answer with the number only.
Person Has Place Of Death	Where did {subject_entity} die?	Provide only the place, or respond with ” if unknown, with no extra text.
Person Has Profession	What is {subject_entity}’s profession?	Answer with the profession(s), separated by ’, ’ with no extra text.
Person Has Spouse	Who is {subject_entity} married to?	List only the spouse name, with no extra text.
Person Plays Instrument	What instrument does {subject_entity} play?	List the instrument(s), separated by ’, ’ with no extra text.
Person Speaks Language	What languages does {subject_entity} speak?	List the language(s), separated by ’, ’ with no extra text.
River Basins Country	In which country can you find the {subject_entity} river basin?	Answer with the country name, or ” if none, with no extra text.
Series Has Number Of Episodes	How many episodes does the series {subject_entity} have?	Answer with the number only.
State Borders State	Which states border the state of {subject_entity}?	List only the state(s), separated by ’, ’ with no extra text.
Compound Has Parts	What are the components of {subject_entity}?	List the components, separated by ’, ’ with no extra text.

Table 5: Relation-specific Zero-Shot Question Prompts. For the question part, the question prompt template, as provided by the authors of the LM-KBC 2023 Challenge, is looked up for each relation individually and the instruction part is appended to increase the chance of correctly formatted results when querying the LLM.