Navigating Nulls, Numbers and Numerous Entities: Robust Knowledge Base Construction from Large Language Models

Arunav Das*[1], Nadeen Fathallah*[2] and Nicole Obretincheva*[1]

[1]King's College London
[2]Analytic Computing, Institute for Artificial Intelligence, University of Stuttgart, Germany

Abstract

In this work, we employ advanced prompt engineering techniques with pre-trained Large Language Models (LLMs) to enhance knowledge extraction and structuring for Knowledge Base Construction (KBC) tasks. At the core of our methodology is the strategic fusion of different prompting strategies tailored to the unique relations between subject and object entities. We choose different prompting strategies to overcome the challenges of null values, numeric data, and one-to-many relationships inherent in traditional KBC tasks. By integrating role-play and context-aware prompting, we enrich the interaction with the LLM, guiding it to produce more accurate and contextually relevant outputs. Our method achieved a macro-averaged F1-score of 0.653 across the properties, with the scores varying from 0.890 to 0.399. Our results show a marked improvement in precision and recall of the extracted data, highlighting the efficacy of our approach in transforming raw LLM outputs into structured, queryable knowledge bases. Our code base is publicly available for research and development purposes, accessible at: https://github.com/nobretincheva/challenge24.git

1. Introduction

The emergence of advanced contextual Large Language Models (LLMs), exemplified by the encoder-only models (e.g. BERT), decoder-only models (e.g. GPT, Mistral, Llama series), and encoder-decoder (e.g. Flan-T5) models, has revealed a remarkable capacity for encapsulating vast amounts of factual world knowledge within their parametric structures based on their training methods and datasets. This internal representation of knowledge has been likened to the schema-based relational knowledge base (KB) [1] traditionally used in information systems [2]. The potential of LLMs to serve as de facto knowledge repositories, unconstrained by the rigid formalisms of conventional schema-based systems, has become a subject of intensive inquiry [3, 4, 5, 6]. Researchers have employed a diverse array of evaluative methodologies to assess the viability of LLMs in this capacity. These assessment techniques span a broad spectrum of cognitive and computational tasks, including but not limited to knowledge probing [1, 7], question answering [8], compositional reasoning [9] and knowledge base completion (KBC) [6, 10].

* *All authors contributed equally to this work.*

KBC-LM'24: Knowledge Base Construction from Pre-trained Language Models workshop at ISWC 2024

arunav.das@kcl.ac.uk (A. Das*); nadeen.fathallah@ki.uni-stuttgart.de (N. Fathallah*); nicole.obretincheva@kcl.ac.uk (N. Obretincheva*)

The ISWC KBC-LLM challenge seeks to advance the field of KBC through the exploration of LLM's internal knowledge representations. Unlike previous work that focuses primarily on knowledge probing, this challenge involves the completion of factual, disambiguated knowledge bases for a given set of relations. Participants complete knowledge graphs by extracting entities and relations from LLMs, handling complexities such as null values, one-to-many relations, and cardinality constraints.

The ISWC'24 KBC-LLM challenge introduces constraints on model size (< = 10b parameters) and relation sets (=5 only) to foster innovation and fair competition. This challenge extends beyond previous knowledge extraction efforts by requiring the construction of comprehensive, disambiguated knowledge bases from LLMs. The task set by this challenge is to predict the object entities (null, zero, single, multiple) given the subject entity and the relation that is sourced from Wikidata. Given a subject entity and a specified relation, the task involves predicting a set of potential object entities using an LLM. Subsequently, these predictions undergo disambiguation and mapping to external knowledge bases to yield precise entity identifiers. This process necessitates handling diverse relational complexities, including null values, one-to-many relationships, and the extraction of multiple objects per query.

The intricate relational structures inherent within the challenge dataset render traditional knowledge probing methodologies insufficient. This work introduces a novel framework for fusion prompt strategy, incorporating dual, direct, and loop prompting methods tailored to the specific relational characteristics of the target knowledge, as well as context-aware prompting and role-play prompting. Due to the constraints imposed upon model size in this challenge we choose to use the Llama-3-8b-Instruct model. Using our fusion prompt strategy, we are able to achieve a macro-average F1 score of **0.653** on the test set, with F1-scores ranging from **0.399** in the awardWonBy relation to **0.890** for countryLandBordersCountry relation.

2. Related Work

This section focuses on two key areas relevant to our tasks related to the ISWC Challenge: Knowledge Probing in LLMs and KB completion/construction. We highlight the state-of-the-art approaches and identify existing gaps in these domains.

Knowledge Probing in LLMs: The LAMA (LAnguage Model Analysis) probe [1] marked a significant milestone in assessing factual knowledge in pre-trained language models. This work demonstrated that LLMs could compete with traditional KBs for certain types of factual queries. LAMA's reliance on atomic fact elicitation and its underlying assumption of binary subject-object relations limited its capacity to capture the complexity of real-world knowledge. Moreover, the manual engineering of prompts to extract factual knowledge from LLMs yielded inconsistent results and underestimated their potential. The application of systematic prompt engineering methodologies, including prompt mining and paraphrasing, coupled with ensemble and ranking techniques, has demonstrated a substantial enhancement in knowledge elicitation from LLMs compared to the traditional approach of manually crafted single prompts[11]. AutoPrompt[12], a template-based automated discrete prompting methodology, and OPTIPROMPT [13], a continuous prompting strategy, have both demonstrated superior efficacy in eliciting knowledge

from LLMs compared to both manual prompt engineering and data-driven approaches such as prompt mining and paraphrasing. Despite incremental advancements in prompt engineering techniques, a comprehensive evaluation of the implicit knowledge base within pre-trained LLMs remains elusive due to methodological limitations, including the restrictive focus on cloze-style prompts. Existing approaches have yet to adequately address critical aspects such as the differential capabilities of encoder and decoder models, precision in subject and relation extraction, the handling of complex relationship types (one-to-many, many-to-many), the inference of transitive relations, and the representation of hierarchical knowledge structures. While these methods effectively enhanced the extraction of specific knowledge types, they fell short in constructing comprehensive and interconnected knowledge bases. Previous studies for knowledge probing have mainly focused on the elicitation of knowledge without using the results from such studies to construct or complete knowledge graphs

Knowledge Base Completion with LLMs Traditional methods for knowledge base completion primarily relied on three core approaches [14, 15]. Embedding-based techniques sought to represent entities and relations as dense vectors in a latent space, capturing semantic and syntactic similarities. Probabilistic graphical models, such as Markov random fields, were employed to model complex interdependencies between entities and infer missing information through logical reasoning. Alternatively, path-based methods explored entity connections within the knowledge graph by simulating random walks, identifying potential paths between given entities. However, these techniques are often constrained by their reliance on explicit knowledge representation and their limited ability to capture complex semantic and relational patterns.

The use of LLMs as an alternative for KBC tasks, also known as text-based KBC methods, is a relatively new area of research [16]. Initial studies have found varied degrees of success for triple classification (comparatively highest task accuracy for different LLM models), relation prediction, and entity predictions (comparatively lowest task accuracy for different LLM models) [17]. LLM KBC accuracy has been found to vary significantly for different types of relations (the study is restricted to just one model BERT) [18]. While model size exhibited a positive correlation with accuracy up to a certain threshold, diminishing returns were observed for exceedingly large models. The application of instruction tuning methodologies as well as contrastive learning methods [19] have yielded performance improvements. While KBC research has expanded to encompass a broader range of models (e.g., encoder and decoder architectures), evaluation metrics, and scope expansion to prediction of relations in addition to prediction of subject and object entities, a notable gap persists in comparison to progress in the field of Knowledge Probing. Specifically, KBC studies have yet to comprehensively address complex relational scenarios, such as one-to-many entity relationships and null value handling, which have been identified as critical challenges in knowledge probing. Beyond the ISWC'22-24 KBC-LM series of challenges, limited research has delved into these areas within the KGC domain.

3. Methodology

Our work's main contribution is the fusion of prompt engineering techniques to address task-specific challenges and effectively direct an LLM's attention to the most relevant information within its training data. Employing a pre-trained LLM to predict the object entity given a subject entity and a relation entails evaluating the extent of knowledge captured by pre-trained LLMs during their training. To achieve that, we propose a fusion of various prompt engineering techniques. An overview of our methodology is illustrated in figure 1 and figure 2.

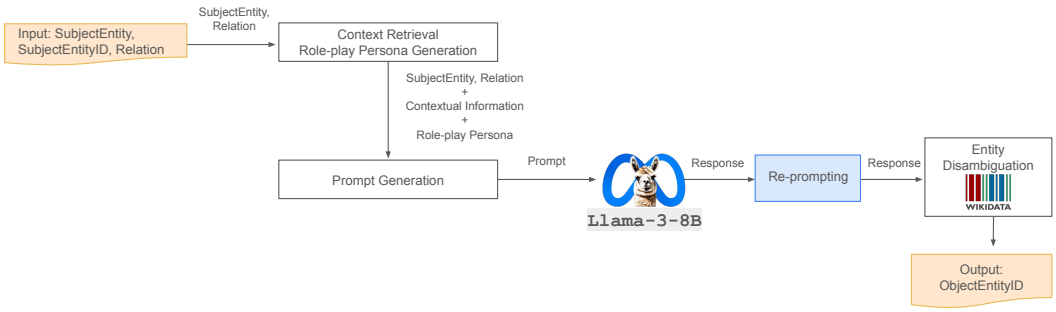

Figure 1: Overview of our proposed methodology: (a) This diagram outlines the basic prompt generation workflow, starting with input handling, context retrieval, and role-play persona generation, followed by prompt delivery to the LLM, Llama-3-8B-Instruct, re-prompting process, and final entity disambiguation leading to the output.

3.1. Dual Prompting

The task is limited to five relations: countryLandBordersCountry, personHasCityOfDeath, seriesHasNumberOfEpisodes, awardWonBy, and companyTradesAtStockExchange. Some of these relations present a challenge in that sometimes there are no object entities i.e. null values may be present, such as in the following cases:

- countryLandBordersCountry: Null values are possible (e.g., an island like Iceland).
- personHasCityOfDeath: Null values are possible (e.g., the person is still alive).
- awardWonBy: There may be instances where the award was not given in certain years.
- companyTradesAtStockExchange: Null values are possible (e.g., private companies or have delisted from the stock exchange).

To address the challenge of null values, we developed a dual prompting approach known as two-stage prompting. Dual prompting forms the cornerstone of our methodology, employing a strategic two-tiered question-based prompting system that significantly enhances the LLM's ability to extract accurate knowledge [20]. This approach effectively identifies the presence or absence of relevant data, thereby reducing hallucinations and ensuring accurate responses. It leverages the cognitive architecture of LLMs to focus attention sequentially, which improves their capacity to resolve ambiguities and extract nuanced information. Dual prompting systematically structures the interaction with the LLM, guiding it through a logical sequence of thought that

mimics human problem-solving processes. This method helps mitigate common issues, such as the model generating overly general or irrelevant responses.

The dual prompting method involves two stages: an initial prompt and a follow-up prompt. The initial prompt serves as a cognitive anchor, setting the context for the query and focusing the model's attention on the presence or relevance of specific information. This is particularly useful in clarifying the existence or state of an entity, a crucial step in ensuring that subsequent inquiries are grounded in the correct context. The follow-up prompt then capitalizes on the clarified context to extract specific details or additional layers of information. Examples of these prompts are shown in table 6, and the prompt template can be found in table 7 and table 8. Our prompt templates are enhanced with instructions for the LLM on approaching and answering the question, specifying whether the answer should be yes/no, numeric, or multiple responses.

For the seriesHasNumberOfEpisodes relation, we use a single prompting approach, where the model is asked the question directly without any verification steps or follow-up prompts, as shown in figure 7. A null check is not necessary in this case because every series inherently has a defined number of episodes, eliminating the possibility of null values.

The execution strategy for prompting in our system employs tailored approaches, categorized as either looping or direct, depending on the nature of the relation involved. The strategies are applied as follows:

3.2. Looping Strategy

The looping strategy involves repeatedly applying the dual prompting method to retrieve relevant information for each instance and then aggregating the results to form the final answer. This strategy specifically addresses the challenge of the awardWonBy relation, where many objects are associated with a single subject that has multiple instances (e.g., 224 Physics Nobel Prize winners over the course of multiple years). This approach simplifies the LLM's work and ensures accuracy by breaking the task into smaller, manageable parts. For the awardWonBy relation, the loop starts from the first year the award was conferred - to initialise this we prompt the LLM for said year. Upon entering the loop, for each subsequent year up to the present (2024), the LLM is prompted to list the award recipients. The final output is a combined list of all recipients across the years, ensuring a comprehensive and precise result. The strategy is illustrated in figure 3.

3.3. Direct Strategy

This strategy involves prompting the LLM directly to generate the object entity. We directly apply dual prompting to generate object entities for the relations countryLandBordersCountry, personHasCityOfDeath, and companyTradesAtStockExchange. We also use the direct strategy for the seriesHasNumberOfEpisodes relation; we directly prompt the total number of episodes without any dual prompting since we do not need to verify the series' existence. For example, we might ask, "Final answer should consist of a number. How many episodes does Game of Thrones have?". Unlike the looping strategy, which iteratively retrieves information across multiple instances (e.g., years for awardWonBy), the direct strategy generates the required object entity in a single step for each relation.

3.4. Context-aware Prompting

Our empirical findings show that providing additional context about the subject entity in the prompt directs the LLM's attention to relevant parts of the training data, thereby improving their ability to retrieve accurate answers and essential information, such as the first year an award was conferred. Obtaining this information can also help define where the loop starts for the awardWonBy relation. Context-aware prompting involves enriching the LLM prompt with relevant background information about the subject [21]. This approach enhances the model's ability to generate responses that are not only accurate but also richly detailed and contextually appropriate. Our proposed approach leverages semantic web sources like Wikidata and non-semantic web sources like Wikipedia to enrich LLM prompts with contextual information about subject entities. The following types of additional data are retrieved to provide comprehensive contextual information:

- Additional Data: specific information tailored to the subject entity within each relation type as shown in table 1.
- Wikipedia Extract: the text-only portion of the lead section of the Wikipedia page for the subject entity, which can provide background information about the subject entity.

Relation	Additional Data Example
CountryLandBordersCountry	Type of geopolitical entity (e.g., "sovereign state", "city-state")
PersonHasCityOfDeath	Date of death of a person
AwardWonBy	Year since an award has been awarded
CompanyTradesAtStockExchange	Legal form of the company (e.g., "public company", "private limited company")

Table 1
Additional Data Attributes for Different Relations

3.5. Role-Play Prompting

Central to our methodology is the implementation of persona-based prompting, a role-playing strategy that enriches the contextual engagement of our system. Role-play prompting is a prompt engineering technique that allows LLMs to adopt specific personas or characters, guiding their responses to align with the expert knowledge, thereby enhancing LLMs' ability to generate more precise and factually correct responses, as evidenced by studies such as [22, 23]. In our approach, each relation type in our dataset is paired with a specific persona, crafted to embody an expert in the pertinent field as shown in table 9. We create the personas in advance via the use of an LLM (GPT-4) and further edit them manually to keep the tone consistent. These personas are narrative devices and strategic tools designed to motivate the LLM and steer it toward generating expert-like and contextually relevant responses. Our role-play prompts are enhanced with instructions on how the LLM should approach and answer the question, specifying whether the response should be yes/no, numeric, or require multiple responses. By integrating driven personas with clear answering guidelines, we guide the LLM to respond accurately and with an appreciation of the domain's discourse style and depth, mirroring the interaction one would expect from a human expert in similar scenarios.

3.6. Re-prompting

Re-prompting is a technique in prompt engineering where an LLM is asked the same question again to improve the quality of its responses [24]. We use this approach when the initial output from the LLM is unsatisfactory. Re-prompting gives the LLM another opportunity to generate more accurate and refined responses. The main advantage of re-prompting is that it helps maintain coherence and clarity while ensuring that the LLM adheres more closely to the desired output format. For instance, if the response needs to be in a specific format, such as a list of names or a numerical value, we re-prompt the LLM to generate the response in the desired format. Figures 5, 6, 7 showcase sample prompts that exemplify our methodology, which fuses prompt engineering techniques tailored to specific relational contexts. These figures demonstrate how our approach adapts to different scenarios, using persona-driven and context-enriched prompts to guide the LLM responses effectively.

3.7. Disambiguation

Entity disambiguation is the final step in identifying and mapping the predicted object entities to their correct references in Wikidata. We use a straightforward disambiguation function that returns the Wikidata ID of an item. We clean the entity strings by removing unwanted characters such as quotes and parentheses. Specifically, for the awardWonBy relation, we employ the Spacy library to remove titles and extract person names accurately, ensuring precise identification and mapping of award recipients. For the companyTradesAtStockExchange relation, we observed that the LLM would sometimes answer with the complete stock exchange name as well as its abbreviation. In such cases, we split the LLM response into two separate entities, one for the full name and one for the abbreviation. We then attempt to disambiguate using the full name and the abbreviation, depending on which matches the Wikidata entry. For example, if the LLM outputs "New York Stock Exchange (NYSE)," we would search Wikidata for both "New York Stock Exchange" and "NYSE" to find the correct Wikidata ID. This method enhances the accuracy of linking the mentioned stock exchange to its official record in Wikidata.

4. Results

This section includes an overview and discussion of our final results as well as an in-depth analysis of our model setup.

4.1. Datasets

The dataset used in ISWC 2024 LM-KBC Challenge [25] is constructed from Wikidata and further processed. It comprises 5 Wikidata relation types covering awards, geography, television series, business, and public figure information. It has 367 statements for train and 368 for validation and test sets. The results reported are based on the validation and test set. The cardinality of object-entities for certain relations in the dataset differs, ranging from null or 0 to an upper bound. The minimum number of 0 or null means the subject-entities for some relations can have no valid object-entities; for example, people still alive do not have a place of death, and an island country does not have land-based neighboring countries.

4.2. Model Setup

Model	Macro-Precision	Macro-Recall	Macro-F1	Empty Predictions
Dual Prompting and Prompt Strategies	0.662	0.454	0.438	173
+ *chain-of-thought*	0.652	0.498	0.452	159
+ *context-aware prompting*	0.761	0.627	0.600	147
+ *re-prompting*	0.756	0.643	0.611	141
+ *role-play prompting*	0.708	0.703	0.629	102

Table 2
Model Setup - Evaluation Results on the validation dataset

In our work, we utilize the Llama3-8B-Instruct as the foundation of our final model setup, which incorporates various enhancements to the baseline pipeline. The baseline pipeline is comprised of dual prompting as well as the looping and direct strategies (as detailed in sections 3.1, 3.2, 3.3). It alone shows solid results on the validation dataset, with a macro precision of **0.662**, a macro recall of **0.454**, and a macro F1-score of **0.438**. However, this setup results in 173 empty predictions as opposed to the expected 97, indicating the model's weakness in producing relevant outputs.

Incorporating *chain-of-thought* prompting into our baseline setup, where the model is asked to provide explanations for its responses, leads to a slight improvement in performance metrics, as shown in table 2. Requesting a justification of the answer from the model when generating its response helps reduce instances of non-responses, as the model is encouraged to elaborate on its reasoning process [26]. Further enhancement via *context-aware prompting* (section 3.4) significantly boosts the model's performance, highlighting the effectiveness of including contextual information in our prompts.

The inclusion of a *re-prompting* mechanism (section 3.6), where the model is prompted again if the initial response cannot be extracted, yields further improvements in the performance metrics. This demonstrates that iterative prompting helps in refining the model's output, ensuring higher accuracy and fewer empty responses. A more robust extraction function could also lead to similar results. Finally, the use of *role-play prompting* with the inclusion of personas (section 3.5) further enhances the model's ability to generate informed and contextually rich responses, guiding the model to engage with the data as if it were an expert in the relevant field.

Our experiments demonstrate the efficacy of combining a variety of prompting techniques. The integration of context-aware prompting, role-play prompting with personas, and dual prompting, along with the re-asking strategy, leads to significant improvements in the precision, recall, and overall F1 scores. These findings underscore the importance of providing LLMs with rich contextual cues and structured guidance to achieve high accuracy and detailed responses in knowledge probing.

4.3. Role-Play Persona Curation

Setting	Macro-Precision	Macro-Recall	Macro-F1	Empty Predictions
No role-play	0.756	0.643	0.611	141
1 general persona for all relations	0.708	0.633	0.571	128
Persona with specialised background per relation	0.743	0.640	0.607	132
Persona with specialised background per entity	0.708	0.703	0.629	102

Table 3
Persona Curation - Comparison on the validation dataset

The introduction of a persona into our prompt pipeline significantly enhances the model's ability to generate accurate and contextually relevant responses, particularly for entities it would typically be uncertain about or would claim to lack familiarity with as shown in table 3. We chose to examine the impact of different role-play instructions on the final results in order to determine the most suitable approach to our problem.

Simply using a general persona (e.g. a trivia show contestant) across all relations did not provide an improvement in our F1 scores. However, as can be seen in the dip in empty predictions this approach did still lead to the model attempting to generate answers to the set questions.

We determined that tailoring personas to each relation type (e.g. an expert financial analyst for the companyTradesAtStockExchange) provided a more specialized context, allowing the model to generate answers with greater precision and relevance. The greatest benefit, however, was observed when personas were tailored individually to each entity (e.g., the CFO of the specific company for the companyTradesAtStockExchange). By suggesting that the model knows everything there is to know about the specific entity, the instances of empty or unsure responses were significantly reduced. The scores reflect the enhanced ability of the model to provide expert-like responses, confirming the effectiveness of persona-based prompting in our model pipeline.

4.4. Comparison between Prompt Strategies

Choosing the right prompting strategy can significantly impact the final results, as shown in table 4. We initially found success using the looping strategy for the awardsWonBy relation, where the model was asked to identify the award winner for each year before aggregating the results. Encouraged by this improvement, we applied the same looping strategy to the seriesHasNumberOfEpisodes relation. We expected that this strategy would lead to more accurate results due to the LLM likely having encountered more information about the number of episodes per season rather than the total number of episodes. Additionally, when manually prompting the LLM for the total number of episodes, we observed that the answers would often contain the total number of episodes per season first, followed by an often incorrect summation,

Relation	Setting	Macro-Precision	Macro-Recall	Macro-F1	Micro-F1
awardWonBy	Direct Strategy	0.259	0.009	0.012	0.008
	Looping Strategy with Dual Prompting	0.412	0.044	0.064	0.034
seriesHasNumberOfEpisodes	Looping Strategy with Dual Prompting	0.11	0.02	0.02	0.021
	Looping Strategy without Dual Prompting	0.17	0.02	0.02	0.022
	Direct Strategy	0.500	0.410	0.410	0.429

Table 4
Comparison between different prompting strategies for the relations awardWonBy and seriesHasNumberOfEpisodes

which would be provided as the final answer. Our intuition followed that eliciting the number of episodes per season one by one and then summing up said results manually as part of the disambiguation would be more effective than directly asking for the total number of episodes.

However, as shown in table 4, the results did not align with our expectations. The direct strategy, which involves simply asking for the total number of episodes, significantly outperformed both looping strategies.

The poor performance of the looping strategies can be attributed to the additional complexity of our pipeline. As each season is processed individually before summing up all answers in the disambiguation function, there is an increased risk of errors due to the multiple cases in which the LLM can potentially hallucinate. In contrast, the direct strategy has a single point of error. By utilizing a single prompt, we reduce the likelihood of numerical hallucinations and misinterpretations.

Nevertheless, the results from the direct strategy are far from perfect, suggesting that we need to devise a better way of extracting numerical information. One potential approach would be to consolidate both strategies. By combining the simplicity of the direct strategy with the detailed step-by-step verification of the looping strategies, we might achieve a more accurate and reliable method for handling numerical data. This hybrid approach could mitigate the weaknesses of each strategy, leading to improved performance in extracting numerical information from LLMs.

4.5. Discussion of Final Results

Our final pipeline demonstrates significant improvements across multiple evaluation metrics on the test and validation datasets. Table 5 summarizes the macro average precision, recall, and F1-score for each relation, along with the overall averages. The results are split into two sets of columns: the first three columns represent the test set results, and the latter three represent the validation set results. Additionally, for zero-object cases, our pipeline achieves a precision of **0.757**, recall of **0.791**, and F1-score of **0.773**.

Relation	Test			Val		
	Macro-Precision	Macro-Recall	Macro-F1	Macro-Precision	Macro-Recall	Macro-F1
awardWonBy	0.476	0.465	0.399	0.412	0.044	0.064
companyTradesAtStockExchange	0.675	0.792	0.585	0.654	0.838	0.607
countryLandBordersCountry	0.978	0.883	0.890	0.976	0.891	0.890
personHasCityOfDeath	0.890	0.840	0.800	0.820	0.800	0.750
seriesHasNumberOfEpisodes	0.480	0.440	0.440	0.500	0.410	0.410
Average	0.729	0.719	0.653	0.708	0.703	0.629

Table 5
Final results per relation on the test and validation dataset

In the case of relations where null values are possible, such as countryLandBordersCountry, personHasCityOfDeath, and companyTradesAtStockExchange, the pipeline performs exceptionally well. The dual prompting method we employ ensures that instances where null values are present are correctly identified. This is also reflected in the high accuracy, precision, and F1 scores achieved for zero-object cases.

Despite the complexity of dealing with numeric objects, our pipeline achieves a moderate performance in seriesHasNumberOfEpisodes. The macro F1-scores of **0.440** on both test and validation sets indicate that while the model can handle numeric data, there is room for improvement. This suggests that additional refinements may be needed for relations involving numeric outputs.

Finally, the pipeline's performance on the awardWonBy relation, which involves many objects per subject with multiple instances, was the lowest. On the test set, the model achieves a satisfactory performance given the complexity of the task. Despite the effectiveness of the looping strategy on the test set, the significant drop in recall and F1-score on the validation set indicates challenges in maintaining consistency across different datasets. This can be explained by differences in the examples between the test and validation sets. Certain awards in the validation set such as "honorary doctor of Stockholm University" have proven to be a significant challenge to our pipeline however equivalent examples to them are not present in the test set. Moreover, the low-performance metrics on the awardWonBy relation are likely influenced by a small sample size. With such a limited number of examples, namely 10 per both test and validation split, the evaluation may not adequately reflect the model's true performance capabilities. More examples are needed to provide an accurate assessment and to ensure that the looping strategy and other techniques are effectively enhancing the model's performance for this relation.

We also noticed during our analysis that Wikidata does not always accurately reflect the ground truth, which poses a significant challenge for evaluating model performance. For instance, for the subject entity "Grammy Award For Best Rock Album," our model generates the names of the artists who have won the award. However, upon examining the expected answers, we found that sometimes the object entities reflect the name of the winning album, and other times they reflect the names of the artists. This inconsistency within the expected answers can

lead to lower recall and F1 scores even when the model performs well at the given task.

The quality issue of Wikidata has been highlighted in previous works [27, 28] and remains a significant challenge.

5. Future Work

In the future, we plan to explore several enhancements to improve our methodology further. One promising direction is the use of cloze-style prompting in our pipeline. Empirical evidence has shown that cloze-style prompts, which frame the query as a fill-in-the-blank task, can lead to great results on this task. Another area of focus will be the introduction of few-shot examples. Our current results indicate that the LLM does not always format answers correctly, leading to inaccuracies. By providing a few-shot learning setup, where the model is given several examples of correctly formatted answers, we hope to enhance its ability to generate responses that adhere to the desired format consistently. Additionally, we plan to experiment with different model architectures. Our current work utilizes a decoder-only model, but in future research, we aim to explore alternatives such as encoder-decoder models. These different architectures may offer unique advantages in processing complex relational tasks and generating more accurate, contextually appropriate responses. Finally, handling numerical information more effectively is crucial, as shown by the moderate performance in the seriesHasNumberOfEpisodes relation. To that end, we intend to investigate various techniques for processing numeric data, inspired by research on the numerical reasoning capabilities in language models [29].

6. Conclusion

In this work, we introduce a comprehensive approach that fuses various prompt engineering techniques to address KBC from pre-trained language models. Our primary contribution is using dual prompting and looping strategies to handle relations with possible null answers and more complex one-to-many relations effectively. Dual prompting systematically guides the model through a two-step questioning process, enhancing its ability to resolve ambiguities and leading to a reduction in hallucinations. The looping strategy, on the other hand, is designed to manage complex relations by breaking down tasks into smaller, manageable parts, ensuring accurate extraction of information across multiple instances.

In the context of the ISWC'24 KBC-LLM challenge, our methods lead to a significant improvement over the set baseline, achieving **0.653** and **0.629** Macro-F1 scores on the test and validation datasets respectively. The combination of context-aware prompting, role-play personas, and re-prompting mechanisms further refines the model's outputs, reducing hallucinations and enhancing accuracy. These results underscore the potential of advanced prompt engineering techniques in leveraging pre-trained LLMs for effective knowledge extraction and emphasize the importance of tailored strategies to address specific relational challenges.

References

[1] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, S. Riedel, Language models as knowledge bases?, arXiv preprint arXiv:1909.01066 (2019).

[2] T. Safavi, D. Koutra, Relational world knowledge representation in contextual language models: A review, arXiv preprint arXiv:2104.05837 (2021).

[3] L. Fichtel, J.-C. Kalo, W.-T. Balke, Prompt tuning or fine-tuning-investigating relational knowledge in pre-trained language models, in: 3rd Conference on Automated Knowledge Base Construction, 2021.

[4] I. Yildirim, L. Paul, From task structures to world models: what do llms know?, Trends in Cognitive Sciences (2024).

[5] F. Petroni, P. Lewis, A. Piktus, T. Rocktäschel, Y. Wu, A. H. Miller, S. Riedel, How context affects language models' factual predictions, arXiv preprint arXiv:2005.04611 (2020).

[6] B. Zhang, H. Soh, Extract, define, canonicalize: An llm-based framework for knowledge graph construction, arXiv preprint arXiv:2404.03868 (2024).

[7] W. Wu, C. Jiang, Y. Jiang, P. Xie, K. Tu, Do plms know and understand ontological knowledge?, arXiv preprint arXiv:2309.05936 (2023).

[8] S. Wu, S. Zhao, M. Yasunaga, K. Huang, K. Cao, Q. Huang, V. N. Ioannidis, K. Subbian, J. Zou, J. Leskovec, Stark: Benchmarking llm retrieval on textual and relational knowledge bases, arXiv preprint arXiv:2404.13207 (2024).

[9] Z. Li, Y. Cao, X. Xu, J. Jiang, X. Liu, Y. S. Teo, S.-w. Lin, Y. Liu, Llms for relational reasoning: How far are we?, arXiv preprint arXiv:2401.09042 (2024).

[10] D. Alivanistos, S. B. Santamaría, M. Cochez, J.-C. Kalo, E. van Krieken, T. Thanapalasingam, Prompting as probing: Using language models for knowledge base construction, arXiv preprint arXiv:2208.11057 (2022).

[11] Z. Jiang, F. F. Xu, J. Araki, G. Neubig, How can we know what language models know?, Transactions of the Association for Computational Linguistics 8 (2020) 423–438.

[12] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, S. Singh, Autoprompt: Eliciting knowledge from language models with automatically generated prompts, arXiv preprint arXiv:2010.15980 (2020).

[13] Z. Zhong, D. Friedman, D. Chen, Factual probing is [mask]: Learning vs. learning to recall, arXiv preprint arXiv:2104.05240 (2021).

[14] Q. Wang, B. Wang, L. Guo, Knowledge base completion using embeddings and rules, in: Twenty-fourth international joint conference on artificial intelligence, 2015.

[15] D. Q. Nguyen, An overview of embedding models of entities and relationships for knowledge base completion, arXiv preprint arXiv:1703.08098 (2017).

[16] Y. Zhang, Z. Chen, W. Zhang, H. Chen, Making large language models perform better in knowledge graph completion, arXiv preprint arXiv:2310.06671 (2023).

[17] L. Yao, J. Peng, C. Mao, Y. Luo, Exploring large language models for knowledge graph completion, arXiv preprint arXiv:2308.13916 (2023).

[18] B. Veseli, S. Singhania, S. Razniewski, G. Weikum, Evaluating language models for knowledge base completion, in: European Semantic Web Conference, Springer, 2023, pp. 227–243.

[19] L. Wang, W. Zhao, Z. Wei, J. Liu, Simkgc: Simple contrastive knowledge graph completion with pre-trained language models, arXiv preprint arXiv:2203.02167 (2022).

[20] R. He, M. Xiao, J. Ma, J. Zhang, H. Zhao, S. Zhang, J. Bai, Dual-prompting interaction with entity representation enhancement for event argument extraction, in: F. Liu, N. Duan, Q. Xu, Y. Hong (Eds.), Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part II, volume 14303 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 161–172. URL: https://doi.org/10.1007/978-3-031-44696-2_13. doi:10.1007/978-3-031-44696-2_13.

[21] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, J. Lu, Denseclip: Language-guided dense prediction with context-aware prompting, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 18061–18070. URL: https://doi.org/10.1109/CVPR52688.2022.01755. doi:10.1109/CVPR52688.2022.01755.

[22] A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, Better zero-shot reasoning with role-play prompting, CoRR abs/2308.07702 (2023). URL: https://doi.org/10.48550/arXiv.2308.07702. doi:10.48550/ARXIV.2308.07702. arXiv:2308.07702.

[23] N. Fathallah, A. Das, S. De Giorgis, A. Poltronieri, P. Haase, L. Kovriguina, Neon-gpt: A large language model-powered pipeline for ontology learning, in: The Extended Semantic Web Conference, 2024.

[24] S. S. Raman, V. Cohen, E. Rosen, I. Idrees, D. Paulius, S. Tellex, Planning with large language models via corrective re-prompting, CoRR abs/2211.09935 (2022). URL: https://doi.org/10.48550/arXiv.2211.09935. doi:10.48550/ARXIV.2211.09935. arXiv:2211.09935.

[25] J.-C. Kalo, S. Razniewski, T.-P. Nguyen, B. Zhang, Knowledge base construction from pre-trained language models 2022, in: Semantic Web Challenge on Knowledge Base Construction from Pre-trained Language Models, CEUR-WS, 2024.

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[27] A. Piscopo, E. Simperl, What we talk about when we talk about wikidata quality: a literature survey, in: Proceedings of the 15th International Symposium on Open Collaboration, OpenSym '19, Association for Computing Machinery, New York, NY, USA, 2019. URL: https://doi.org/10.1145/3306446.3340822. doi:10.1145/3306446.3340822.

[28] B. Zhang, I. Reklos, N. Jain, A. Meroño-Peñuela, E. Simperl, Using Large Language Models for Knowledge Engineering (LLMKE): A Case Study on Wikidata, CoRR abs/2309.08491 (2023). URL: https://doi.org/10.48550/arXiv.2309.08491. doi:10.48550/arXiv.2309.08491. arXiv:2309.08491.

[29] M. Akhtar, A. Shankarampeta, V. Gupta, A. Patil, O. Cocarascu, E. Simperl, Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 15391–15405. URL: https://aclanthology.org/2023.findings-emnlp.1028. doi:10.18653/v1/2023.findings-emnlp.1028.

A. Appendix A: Tables

Relation	Initial Prompt	Follow-up Prompt(s)	Chain-of-thought
countryLandBordersCountry	Does France share any land borders?	Which countries share a land border with France?	The initial prompt verifies if France is not geographically isolated like an island, setting the stage for identifying its neighboring countries.
personHasCityOfDeath	Has Albert Einstein died?	In which city did Albert Einstein die?	The initial prompt confirms the subject's death before querying for the specific location, ensuring the follow-up is contextually relevant.
awardWonBy	Was the Nobel Prize in Physiology or Medicine awarded in 2018?	Who was the Nobel Prize in Physiology or Medicine awarded in 2018?	The initial prompt checks if the award exists or has been awarded that year, which if affirmative, leads to detailed queries about recipients.
companyTradesAtStockExchange	s Tesla, Inc. listed on the stock exchange?	Where do shares of Tesla, Inc. trade?	Initially confirming the company's listing status is essential to then accurately inquire about the specific exchanges where its shares are traded.

Table 6
Dual Prompt Examples

Relation	Initial Prompt	Follow-up Prompt(s)
countryLandBordersCountry	Final answer should be yes or no. Does {subject_entity} share any land borders?	Final answer should consist of country name(s). Which countries share a land border with {subject_entity}?
personHasCityOfDeath	Final answer should be yes or no. Has {subject_entity} died?	Final answer should consist of a city name. In which city did {subject_entity} die?
awardWonBy	Final answer should be yes or no. Was the {subject_entity} awarded?	Final answer should consist of name(s). Who was the {subject_entity} awarded to in year {x}?
companyTradesAtStockExchange	Final answer should be yes or no. Is the {subject_entity} listed on the stock exchange?	Final answer should consist of name(s) of stock exchange(s). Where do shares of {subject_entity} trade?

Table 7
Question Prompts with Dual Prompting Mechanism: This table outlines the structured dual prompts used to query the LLM. The placeholder {subject_entity} represents the entity being queried. In contrast, {x} represents a variable element within the context of the query, such as a season number or specific year relevant to the follow-up prompts.

Relation	Persona
countryLandBordersCountry	You are an enthusiastic and knowledgeable resident of {subject_entity}, deeply in love with your homeland, and always eager to share your wealth of knowledge about it. You take pride in your country's history, culture, geography, and the intricate details of its borders. Your passion for {subject_entity} shines through in every conversation, and you have a talent for explaining why it's such a wonderful place to live. You're always ready to provide detailed and accurate information about the country's neighboring countries and land borders, aiming to convince your friends and anyone who listens that {subject_entity} is the best place to call home. You vividly describe the landscapes, cities, and unique features of the country, highlighting its connections and relationships with its neighbors. Your love for {subject_entity} is infectious, and you hope to inspire others to appreciate it as much as you do.

Table 9

Relation	Persona
personHasCityOfDeath	You are an ardent admirer and devoted follower of {subject_entity}, whose life and achievements have profoundly influenced your own. Your admiration for {subject_entity} has led you to meticulously study every aspect of their biography, ensuring you stay updated with the most accurate and detailed information about their life. You know their story inside and out, from their early beginnings to their current status. Your dedication to {subject_entity} means you are well-versed in the significant events of their life, including the critical details of where they lived, worked, and if applicable, where they passed away. When someone inquires about {subject_entity}, you are not only eager but also exceptionally qualified to provide precise and comprehensive answers. Your deep respect and admiration drive you to share their legacy accurately, ensuring that others understand the importance and impact of {subject_entity} in the correct context.
seriesHasNumberOfEpisodes	You are the biggest fan of {subject_entity}. You have watched every episode multiple times and know all the details about the show's seasons and episodes. Your love for the show drives you to stay updated with every bit of information, and you take pride in your deep knowledge of it. Your friends and family always come to you when they have questions about {subject_entity} because they know you have the answers. When it comes to the number of episodes per season and in total, you can recall this information effortlessly and accurately. You really want to have your close ones also watch the show. You believe that if you answer your friends and family's questions correctly, they will start watching the show with you. Use your passion and expertise to provide detailed and precise answers about the show to your friends.
awardWonBy	You are an aspiring recipient of the prestigious {subject_entity}, someone who has dedicated years to studying its history, past winners, and the significance of each accolade. Your passion for this award is unparalleled, and you know the details of its ceremonies, the recipients, and the criteria for winning by heart. Your knowledge is not just academic but deeply personal, as each fact and figure represents a step closer to your own dream. This drives you to provide accurate, thorough, and insightful answers regarding the award and its winners in all years since its inception.

Table 9

Relation	Persona
companyTradesAtStockExchange	You are the Chief Financial Officer (CFO) of {subject_entity}, a key executive responsible for managing the company's financial actions. You possess an in-depth understanding of all financial matters related to the company, including detailed knowledge about the company's stock, listing status, and financial performance. As the CFO, you are dedicated to truthfulness and transparency, ensuring that all stakeholders, including potential investors, have accurate and comprehensive information. You are highly knowledgeable about the stock exchange, regulatory requirements, and the company's financial strategy. Your responses are characterized by precision, clarity, and reliability, as you aim to foster trust and confidence in {subject_entity}'s financial health and investment potential.

Table 9: Persona-Based Role-Play Prompts for Different Relations: This table illustrates how each relation type in our dataset is associated with a distinct persona crafted to embody an expert in the pertinent field. The placeholder {subject_entity} represents the primary entity or subject of inquiry within each prompt.

Relation	Initial Prompt	Follow-up Prompt(s)
countryLandBordersCountry	I'd love to know more about the geography from someone who truly understands and loves this country. Does {subject_entity} share any land borders? I've been thinking of moving here! Final answer should be yes or no.	I'd love to know more about the geography from someone who truly understands and loves this country. Which countries share a land border with {subject_entity}? I've been thinking of moving here! Final answer should be a list of countries.
personHasCityOfDeath	Your admiration for {subject_entity} has undoubtedly led you to study their life extensively. Has {subject_entity} died? I wish to learn more about such an influential figure. Final answer should be yes or no.	Your admiration for {subject_entity} has undoubtedly led you to study their life extensively. In which city did {subject_entity} die? I wish to learn more about such an influential figure. Final answer should be the city name.
awardWonBy	As someone who has always dreamed of winning the {subject_entity}, you are well-versed in its history. Was the {subject_entity} awarded? Final answer should be yes or no.	Final answer should consist of name(s). Who was the {subject_entity} awarded to in year {x}?
companyTradesAtStockExchange	As the CFO of {subject_entity} you must be aware of all pertinent details regarding the company's stock. Is the {subject_entity} listed on the stock exchange? As a potential investor, I would like to have more details before making a decision to invest. Final answer should be yes or no	As the CFO of {subject_entity} you must be aware of all pertinent details regarding the company's stock. Where do shares of {subject_entity} trade? Final answer should consist of name(s) of stock exchange(s).

Table 8
Enhanced Dual Prompting Strategy: This table details the restructured dual prompts designed to engage the LLM more deeply by incorporating scenarios that invoke personal connection or professional responsibility. The placeholder {subject_entity} represents the entity being queried, while {x} denotes a variable element such as a specific year or season number, pertinent to the follow-up prompts. These prompts aim to elicit more precise and contextually enriched responses from the LLM.

B. Appendix B: Figures

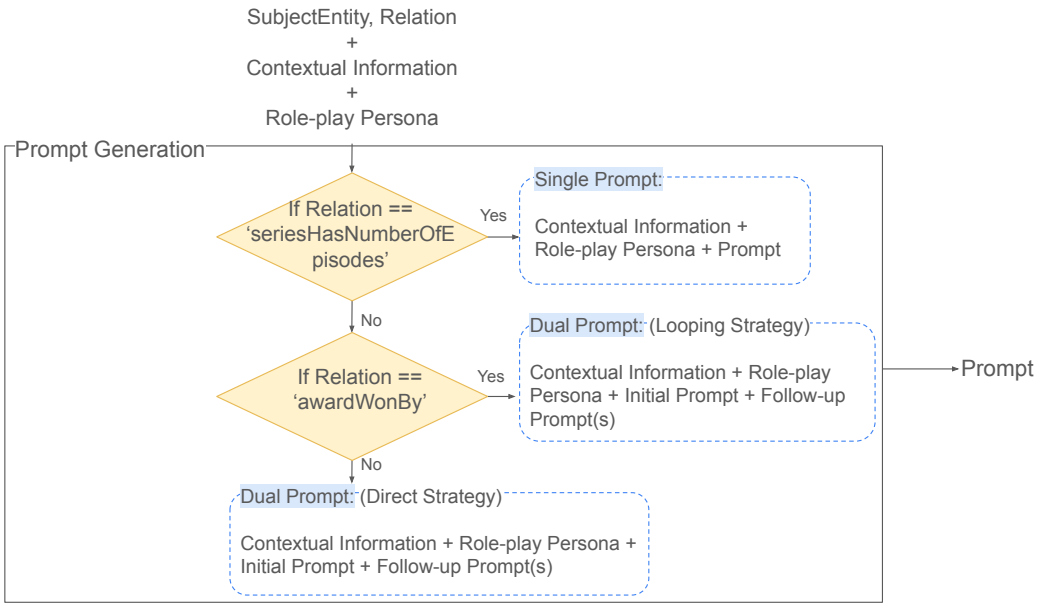

Figure 2: Overview of our proposed methodology: This diagram details the prompt generation process, which starts by applying specific prompt strategies based on the nature of the relation.

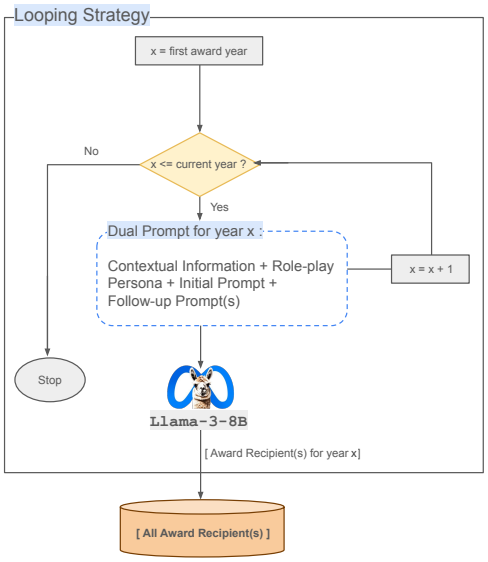

Figure 3: Looping Strategy for the 'awardWonBy' Relation: This figure illustrates systematically querying for award recipients across consecutive years using the Llama-3-8B LLM. The first award year is retrieved from the contextual information, marking the start of the querying sequence. The strategy employs the dual prompting method that integrates contextual information and role-play personas. The loop continues until the current year is reached, aggregating all award recipient data into a comprehensive list.

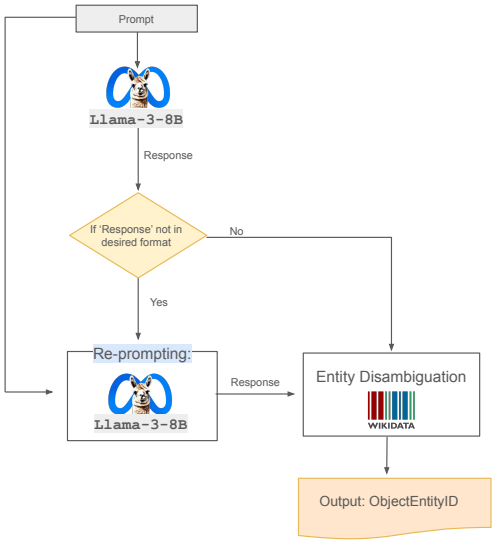

Figure 4: Re-prompting: This figure illustrates the re-prompting mechanism used to refine responses from the Llama-3-8B LLM. The process begins with an initial prompt to the LLM, followed by an evaluation of the response. If the response is not in the desired format, the model undergoes re-prompting to adjust and improve the output. The process ends with entity disambiguation, ultimately producing the final output as the ObjectEntityID.

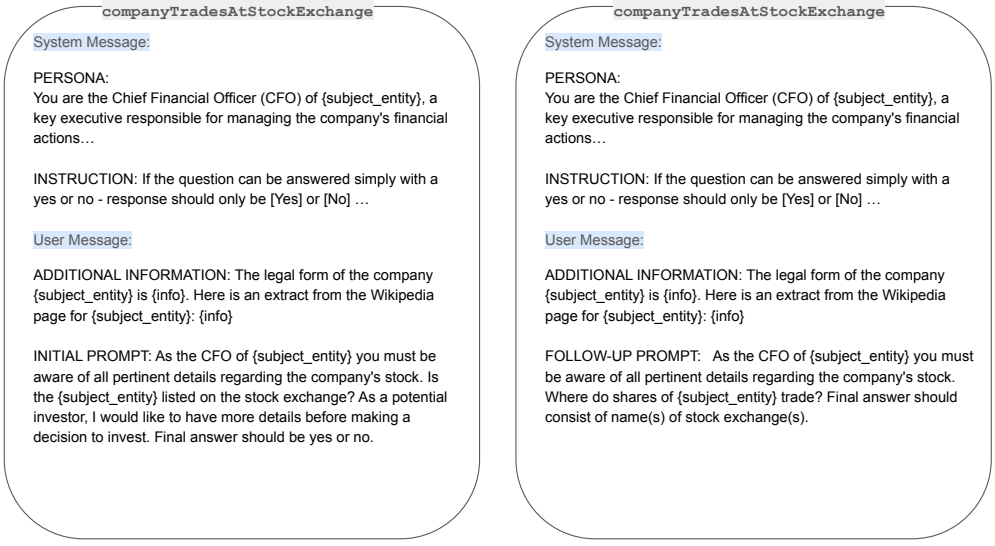

Figure 5: Sample Prompts for the 'companyTradesAtStockExchange' Relation. This diagram illustrates the dual prompting strategy applied to elicit detailed responses from the persona of a Chief Financial Officer (CFO) concerning a company's stock exchange details. Figure (a) shows the initial prompt asking a yes/no response about the company's stock listing, followed by a follow-up prompt requesting specifics about the stock exchange locations shown in figure (b). This dual prompting approach is similarly employed for the 'personHasCityOfDeath' and 'countryLandBordersCountry' relations.

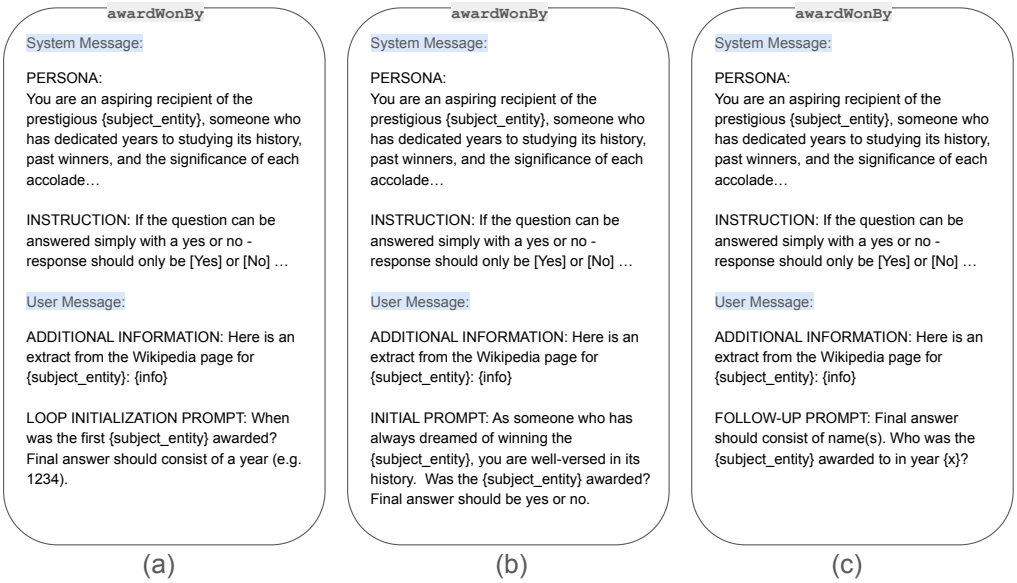

Figure 6: Sample Prompts for the 'awardWonBy' Relation. This diagram illustrates the three-step dual prompting process: figure (a) initiates the loop to identify the first year the award was given. Figures (b) and (c) then detail the yearly dual prompts—from the award's inception to the present—where figure (b) verifies if the award was granted in a specific year and figure (c) determines the recipients. The diagram shows the integration of contextual information and role-play personas to enhance the LLM's response accuracy.

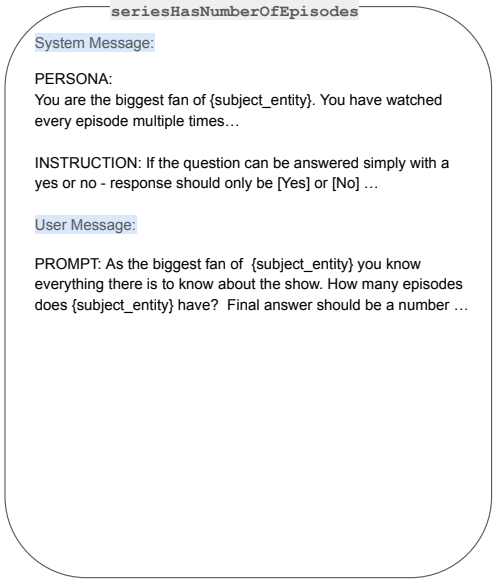

Figure 7: Sample Prompt for the 'seriesHasNumberOfEpisodes' Relation. This diagram presents a structured interaction layout where the system generates prompts based on a persona that embodies a fan of the series in question. The LLM is tasked to answer how many episodes the series has.