NLPeople at TextGraphs-17 Shared Task: Chain of Thought Questioning to Elicit Decompositional Reasoning

Movina Moses[1] and **Vishnudev Kuruvanthodi**[2] and **Mohab Elkaref**[2] and
Shinnosuke Tanaka[2] and **James Barry**[2] and **Geeth De Mel**[2] and **Campbell D Watson**[1]
IBM Research[1] and IBM Research Europe[2]
{movina.moses, vishnudev.k, mohab.elkaref, shinnosuke.tanaka
james.barry}@ibm.com, geeth.demel@uk.ibm.com, cwatson@us.ibm.com

Abstract

This paper presents the approach of the NLPeople team for the Text-Graph Representations for KGQA Shared Task at TextGraphs-17 (Sakhovskiy et al., 2024). The task involved selecting an answer for a given question from a list of candidate entities. We show that prompting Large Language models (LLMs) to break down a natural language question into a series of sub-questions, allows models to understand complex questions. The LLMs arrive at the final answer by answering the intermediate questions using their internal knowledge without needing additional context. Our approach to the task uses an ensemble of prompting strategies to guide how LLMs interpret various types of questions. Our submission achieves an F1 score of 85.90, ranking 1st among the other participants in the task.

1 Introduction

This paper outlines the NLPeople submission to the Text-Graph Representations for KGQA Shared Task at TextGraphs-17. The task involved selecting the correct answer from a list of candidate answers for a given question. Knowledge Graph Question Answering (KGQA) involves using the structured knowledge and relations present in a Knowledge Graph (KG) to answer a natural language question. Previous KGQA methods focused on two approaches: knowledge retrieval and semantic parsing. Knowledge retrieval attempts to extract entities, relations, or triples from the KG that are relevant to the question and can be used to deduce the answer (Sun et al., 2019). On the other hand, semantic parsing transforms the question from unstructured natural language into a structured logical form (Yih et al., 2016). This form can be converted into a query and executed over a KG to obtain relevant answers. However, these methods still struggle to perform the complex reasoning required to answer natural language questions.

To deal with these challenges, recent KGQA research leverages the reasoning and language comprehension capabilities of large language models (LLMs) (Gu et al., 2023; Sen et al., 2023). These methods try to incorporate the structural knowledge present in KGs to address the factual hallucination generated by LLMs during its reasoning process (Baek et al., 2023; Guan et al., 2023).

We propose a chain-of-thought based prompting mechanism, which allows an LLM to deduce the answer by breaking down the initial question into sub-questions, which when answered, lead to the final answer. Furthermore, we present our results using question-type-specific prompting strategies to address the difficulties models face while reasoning over complex question types. We present results for these methods using Llama3-8b-instruct, Llama3-70b-instruct, Mixtral 8x7B, and GPT 3.5. Overall, our results rank 1st with an F1 score of 85.90 improving the baseline GPT 3.5 results by approximately 18%.

2 Methodology

2.1 Problem Formulation

For a given question and a list of candidate entities, the task is to choose the candidate entity that correctly answers the question. Each candidate is associated with a graph of the shortest paths from the entities mentioned in the question to the candidate entity including links of the intermediate nodes.

The dataset is annotated with Wikidata entities and includes seven types of complex questions: generic, ordinal, intersection, superlative, difference, multihop, and comparative. The questions used in this dataset originate from the Mintaka dataset (Sen et al., 2022), which is a large-scale, complex, and natural language dataset. In this section, we detail the elements of our final submission.

Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing, pages 142–148
August 15, 2024 ©2024 Association for Computational Linguistics

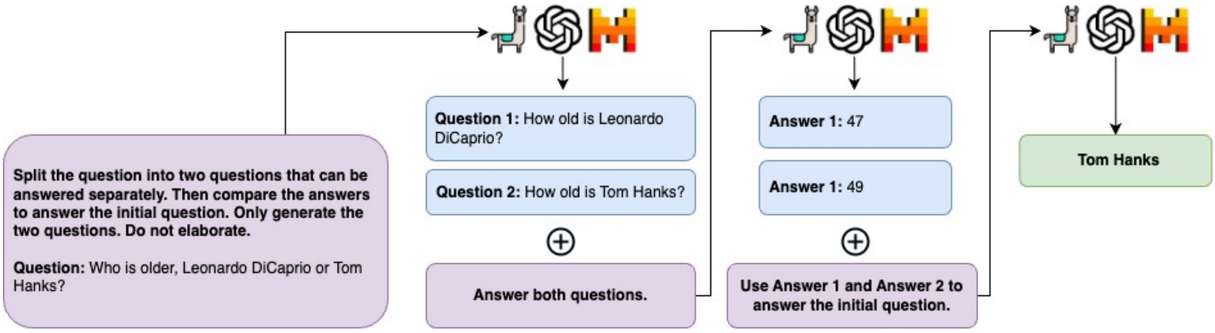
Split the question into two questions that can be answered separately. Then compare the answers to answer the initial question. Only generate the two questions. Do not elaborate.
Question: Who is older, Leonardo DiCaprio or Tom Hanks?
Question 1: How old is Leonardo DiCaprio?
Question 2: How old is Tom Hanks?
Answer both questions.
Answer 1: 47
Answer 1: 49
Use Answer 1 and Answer 2 to answer the initial question.
Tom Hanks

Figure 1: An example of Chain of Thought Questioning

2.2 Chain of Thought Questioning

LLMs can perform closed-book question-answering tasks using the knowledge stored in their parameters (Petroni et al., 2019; Roberts et al., 2020). However, generating a chain-of-thought (CoT) improves the ability of LLMs to perform the complex reasoning required to answer natural language questions. Wei et al. (2023) show how these reasoning abilities emerge when a model is shown examples of intermediate reasoning steps. We propose a variant of CoT prompting called CoT-questioning, where LLMs are instructed to decompose questions into a series of sub-questions that can be answered independently or in sequence to arrive at the final answer. The model is provided with one example suggesting how the model should approach its instructions (Brown et al., 2020) as shown in Figure 1. This style of prompting simplifies the reasoning the model has to do to understand a complex question by mimicking the thought process a human would use and guiding its reasoning path.

Entity Selection The above method provides a list of one or more possible candidate entities. The dataset maps candidate entities to their corresponding Wikidata entity IDs. A candidate entity could map to more than one entity ID since each entity could be different but have the same name. For example, the entity named Beyoncé corresponds to Q15303590: 2013 studio album by Beyoncé and Q36153: American singer (born 1981). While the sub-graphs provide the links and relations between the question and candidate entities, these links do not always provide the information required to reason out the answer. To disambiguate between entities, we prompt the model to select the correct entity by providing the entity name, ID, and WikiData description for the candidate entities using the prompt in Table 7. If the above method does not produce a list of candidates, the model is prompted with all the candidate entities.

2.3 Question Specific Prompts

The dataset comprises questions of different complexity types. Each type would benefit from different intermediate steps and reasoning to arrive at the answer. We propose a prompting methodology that uses different few-shot prompts for each type of question. These types are identified using the question complexity types from the Mintaka dataset. Specifically, we target questions of type ordinal, difference, intersection, and superlative since they are difficult to decompose due to their complexity.

Superlative and Ordinal Type Questions Such questions involve a comparison over possible answers. For superlative questions the task is to select the answer with the maximum or minimum value for a certain property. For ordinal questions the task is to select an answer at a certain position when all answers are ordered with respect to a certain property. Our strategy for both types is to decompose these questions into their constituent factoid questions, followed by a comparative operation to choose the final answer. For example, for the question "Who is the youngest movie director?" can be decomposed into "How old is `[candidate]`?", where `[candidate]` is replaced with each candidate answer, and the answer is decided by the `[candidate]` that returns the minimum answer.

Zero-Shot Reasoning Prompt Some questions demand abstract reasoning across multiple paths. For instance, to answer the question "Who was not an original Spice Girl?" the model must identify the original Spice Girls and determine who was replaced. However, the question could also be interpreted as selecting someone who is not a Spice Girl at all. To address difference questions, we utilize the zero-shot prompt listed in Table 9. It

Prompt Method	Model	Precision	Recall	F1	Accuracy
CoT-Questioning	llama3-8b-instruct	67.46	63.64	65.50	93.67
	llama3-70b-instruct	**82.10**	**78.50**	**80.26**	**96.35**
	mixtral-8x7b-instruct-v01	70.18	68.59	69.38	94.29
	gpt-3.5	75.55	76.94	76.24	95.47
Question-Specific Prompts	llama3-8b-instruct	71.26	69.44	70.34	94.47
	llama3-70b-instruct	**84.61**	**80.90**	**82.71**	**96.81**
	mixtral-8x7b-instruct-v01	68.72	75.53	71.97	94.45
	gpt-3.5	77.99	78.21	78.11	95.86
MCQ Prompts with Additional Context	llama3-8b-instruct	80.17	80.06	80.11	96.25
	llama3-70b-instruct	**81.42**	**81.19**	**81.30**	**96.48**
	mixtral-8x7b-instruct-v01	76.10	76.09	76.10	95.49
	gpt-3.5	79.35	79.35	79.35	96.11

Table 1: Development results using the methods detailed in Section 2.

explicitly asks the model to apply reasoning when answering the question. However, when the model does not produce an answer listed in the candidate entities, we perform entity selection with the top five candidate entities ranked based on the score produced by the cross-encoder[1] when we perform retrieval over the wikipedia page mapped to the entity ID using the query and linearised graph string. This prompt can be used for all types of questions.

2.4 MCQ Prompts with Additional Context

In this approach, context is prepared from multiple sources. In the end, the context, question and filtered answer options are provided to an LLM for it to pick the correct answer.

First, the LLM is prompted to answer a given question and provide an explanation. We then verify that the answer is present in the list of candidate entities. If present, the answer with the explanation is added to the main context along with the question entities and their first paragraphs from Wikipedia, for the answer entities, descriptions, and entity types are fetched from Wikidata. For example, for an answer entity named Nile, the context would look like: A. The Nile (Q110044631): (type watercolour painting) and B. Nile (Q3392): (type river) major river in northeastern Africa. The type and description provide additional context for the LLM to disambiguate and pick the correct option. The context and candidate entities are used to re-rank the entities based on their sentence embeddings. The lowest-ranking options which likely contain the wrong candidates are removed. Finally, using the constructed context and filtered options, the LLM is prompted to select the correct option.

3 Results and Discussion

3.1 Experimental Setup

We evaluate our prompting methods on a 20% split of the train set containing 7497 samples and 707 questions using Llama 3-8b-Instruct, Llama 3-70b-Instruct (AI@Meta, 2024), Mixtral 8x7B[2] and GPT-3.5[3]. The prompts for each model were amended using their prompt formats and special tokens. Wikidata entity descriptions were fetched using the Wikidata client library[4]. We report the results of our methods in Table 1. All models are decoded using greedy sampling.

3.2 Development Results

CoT-Questioning In our CoT-questioning experiments, we used the few-shot prompt outlined in Appendix Table 6. This example was hand-crafted using a question from the training data chosen after examining the scores with multiple few-shot examples. The results show that Llama3-70b-Instruct performed the best, achieving an F1 score of 80.26, followed by GPT-3.5 with a score of 76.24. The smaller Llama and Mistral models yield significantly lower F1 scores.

As shown in Table 2, using entity selection with Wikidata entity descriptions produces much higher scores than using the linearized graph strings from the graphs provided by the dataset. This may be because the shortest path from the question entity to the candidate entity does not always contain the information necessary for the model to select the correct entity. Additionally, some larger graphs with repeated words could mislead the model to select the wrong entity.

[1] huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

[2] huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[3] platform.openai.com/docs/models/gpt-3-5-turbo

[4] wikidata.readthedocs.io/

	Precision	Recall	F1	Accuracy
Linearized Graph String	79.66	74.25	76.86	95.78
Description	**82.10**	**78.50**	**80.26**	**96.35**

Table 2: Llama3-70b-instruct results using entity selection with Wikidata descriptions vs. linear graph string.

Question-Specific Prompting The results from our zero-shot reasoning prompt is presented in Table 3. We infer that when explicitly asked to reason before they answer, the LLM tends to reason out the correct answer. We notice the biggest difference in F1 among superlative and ordinal type questions here. While difference questions score lower in the development split, they eventually score better on the public test set as shown in Table 4.

Question Type	Count	CoT Questioning	Question-Specific Prompts
difference	95	**71.79**	69.23
intersection	157	89.38	**90.96**
ordinal	95	75.28	**83.87**
superlative	126	64.13	**71.90**
Overall F1	707	80.20	**82.71**

Table 3: F1 scores using question-specific prompts with Llama3-70b-instruct.

MCQ Prompts with Additional Context The naive approach with zero-shot prompts of questions and options without any additional context led to poor results. The presence of similar but different options made it difficult to pick the right answer. The additional context from the Wikidata improved the score. This approach performed well on intersection, ordinal and comparative-type questions. However, the F1 scores were lower for the superlative type of questions. Table 11 details the F1 scores by question type.

3.3 Official Results

Final Submission As stated previously, we observed that models produce varying answers for different question types. We ensemble the outputs of the best-performing models - Llama3-70b-instruct and GPT 3.5 using CoT-questioning, to get the best results. We found that when one model gives an incorrect answer, the other often provides the correct one. For a question, if the predicted answers differ, we perform entity selection using GPT 3.5 between the predictions of both models to choose the correct answer. For unanswered questions, we prompted GPT-4 to answer the question and performed entity selection to select the correct answer. Finally, we used outputs from question-type specific prompts with Llama3-70b-Instruct. In each case, we compared the model's output to the previous outputs that produced a high F1 score. Table 4 details the individual results at each step of the ensemble.

Ensembled Method	F1
llama3-70b-instruct with CoT-Questioning	80.29
+ gpt-3.5 with CoT-Questioning	82.81
+ gpt-4 with Zero-Shot Answer	85.19
+ llama3-70b-instruct with Question-Specific Prompt	85.99
Final Ensemble	**85.99**

Table 4: Ensembled system results on public test set.

Results on Private Test Set The official results are presented in Table 5. Our best submission achieves an F1 score of 85.90 on the private test set outperforming the other teams.

Team	F1	Rank
zlatamaria	83.00	2
daeheekim	81.26	3
baseline_chatgpt	67.99	4
mmoses	**85.90**	**1**

Table 5: Top results on the official private test set.

4 Limitations

The effectiveness of CoT-Questioning depends on the model used. As observed from the results, larger models excel at simplifying questions. Additionally, the accuracy of the final answers depends on the data the model has been trained with, so it can produce outdated answers. Better methods to incorporate the information present in KGs could help address this problem.

5 Conclusion

In this paper, we described the NLPeople submission to the TextGraphs-17 Shared Task. We present three different prompting techniques: (1) chain of thought questioning (2), question-type specific prompts, and (3) MCQ prompts with additional context. We demonstrated that by decomposing a question into a series of sub-questions, LLMs can reason over complex questions effectively. Additionally, using question-type specific prompts and demonstrations yields positive results for superlative, ordinal, and difference-type questions. These techniques only require a single demonstration and need no additional context. Our final submission using an ensemble of the above techniques achieves a score of 85.90, which ranks 1st among the other participants.

References

AI@Meta. 2024. Llama 3 model card.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 78–106, Toronto, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yu Gu, Xiang Deng, and Yu Su. 2023. Don't generate, discriminate: A proposal for grounding language models to real-world environments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4928–4949, Toronto, Canada. Association for Computational Linguistics.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2023. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. *Preprint*, arXiv:2311.13314.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Andrey Sakhovskiy, Mikhail Salnikov, Irina Nikishina, Aida Usmanova, Angelie Kraft, Cedric Möller, Debayan Banerjee, Junbo Huang, Longquan Jiang, Rana Abdullah, Xi Yan, Dmitry Ustalov, Elena Tutubalina, Ricardo Usbeck, and Alexander Panchenko. 2024. TextGraphs 2024 shared task on text-graph representations for knowledge graph question answering. In *Proceedings of the TextGraphs-17: Graph-based Methods for Natural Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Priyanka Sen, Sandeep Mavadia, and Amir Saffari. 2023. Knowledge graph-augmented language models for complex question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 1–8, Toronto, Canada. Association for Computational Linguistics.

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. PullNet: Open domain question answering with iterative retrieval on knowledge bases and text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2380–2390, Hong Kong, China. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206, Berlin, Germany. Association for Computational Linguistics.

A Appendix

A.1 Prompts

We show detailed prompts used by our prompting techniques in Table 6, 7, 8, 9 and 10.

A.2 Additional Results

Table 11 presents the F1 scores obtained by the methods detailed in Section 2 for each question type.

```
Split the question into two questions that can be answered separately. Then compare the answers
to answer the initial question. Only generate the two questions. Do not elaborate.
Question: Who is older, Leonardo DiCaprio or Tom Hanks?
Question 1: How old is Leonardo DiCaprio?
Question 2: How old is Tom Hanks?
Answer both questions.
Answer 1: 49
Answer 2: 67
Use Answer 1 and Answer 2 to answer the initial question.
Tom Hanks
```

Table 6: Few-Shot Example used for CoT Questioning in Section 2.2.

```
**Previous Output**
Select the correct ID that references the answer:
[answerEntityId_1]: answerEntity_1 - answerEntity_1_description
[answerEntityId_2]: answerEntity_2 - answerEntity_2_description
```

Table 7: Example of the entity selection prompt used when a predicted answer entity appears multiple times in the candidate entity list. This prompt is added to the previous prompt and generated output in Section 2.2 and Section 2.3.

```
<|begin_of_text|><|start_header_id|>user<|end_header_id|>
Split the question into two questions that can be answered separately. Then compare the answers
to answer the initial question. Only generate the two questions. Do not elaborate.
Question: Who is older, Leonardo DiCaprio or Tom Hanks?<|eot_id|>
<|start_header_id|>assistant<|end_header_id|> Question 1: How old is Leonardo DiCaprio?
Question 2: How old is Tom Hanks?
<|eot_id|><|start_header_id|>user<|end_header_id|>
Answer both questions.<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Answer 1: 49
Answer 2: 67
<|eot_id|><|start_header_id|>user<|end_header_id|>
Use Answer 1 and Answer 2 to answer the initial question. <|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
Tom Hanks<|eot_id|><|start_header_id|>user<|end_header_id|>
Split the question into two questions that can be answered separately. Then compare the answers
to answer the initial question. Only generate the two questions. Do not elaborate.
Question: Who's won more head-to-head tennis matches between each other, Novak Djokovic or
Roger Federer? <|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
Question 1: How many head-to-head tennis matches has Novak Djokovic won against Roger Federer?
Question 2: How many head-to-head tennis matches has Roger Federer won against Novak Djokovic?
Answer both questions.<|eot_id|><|start_header_id|>assistant<|end_header_id|>
Answer 1: 27
Answer 2: 23
<|eot_id|><|start_header_id|>user<|end_header_id|>
Use Answer 1 and Answer 2 to answer the initial question.<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
Novak Djokovic <|eot_id|><|start_header_id|>user<|end_header_id|>
Select the correct ID that references the answer:
[Q5812]: Novak Djokovic - Serbian tennis player
[Q15073898]: Novak - family name
[Q21146583]: Djokovic - family name
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
[Q5812]: Novak Djokovic - Serbian tennis player
```

Table 8: An example of the intermediate outputs produced by the CoT-questioning method from Section 2.2 and its its prompt format using Llama3-70b-Instruct. This color are the prompts by the user and these are the outputs produced by the model.

```
Answer the question in a clear and concise form after using proper reasoning.
Question: <question>
Model generates answer with reasoning
Extract all possible answers.
Answer_1, Answer_2, ...
```

Table 9: Zero-shot prompt reasoning prompt used for specific question-types as mentioned in Section 2.3.

```
INSTRUCTION: You are a multiple-choice quiz expert. You will be provided with a question
and multiple-choice answers in the format of A, B, C, D, E, F, G, H, I, J. You have to read the
given CONTEXT and select one answer. Say the answer option (A or B or C or D or E or F or G or H or
I or J) in ANSWER section Do not Guess the answer. Say UNANSWERABLE if you don't know the answer.
CONTEXT:
Franz Kafka Prize(Q19362): (is of type literary award) international literary awardThe Franz
Kafka Prize is an international literary award presented in honour of Franz Kafka, the Jewish,
Bohemian, German-language novelist. The prize was first awarded in 2001 and is co-sponsored by
the Franz Kafka Society and the city of Prague, Czech Republic.
...
New York City(Q60): (is of type city in the United States) most populous city in the United States
....
QUESTION:
In what city was the author of the second book to win the Franz Kafka Prize born?
ANSWER OPTION:
A. Prague (Q1085)
B. Kraków (Q31487)
...
...
J. New York City (Q60)
ANSWER:
```

Table 10: Prompt template used for the MCQ Prompts in Section 2.4.

Question Type	Count	CoT-Questioning Questioning	Question-Specific Prompts	MCQ Prompts with Additional Context
comparative	81	**92.68**	**92.68**	85.19
difference	95	71.79	69.23	**77.25**
generic	81	**84.27**	**84.27**	80.25
intersection	157	89.38	**90.96**	88.82
multihop	72	**85.71**	**85.71**	81.94
ordinal	95	75.28	83.87	**85.26**
superlative	126	64.13	**71.90**	69.84
Overall F1	707	80.20	**82.71**	81.30

Table 11: F1 scores per question type on the development split using the three different methods. These were obtained using the Llama3-70b-instruct model.