# Skoltech at TextGraphs-17 Shared Task: Finding GPT-4 Prompting Strategies for Multiple Choice Questions

**Maria Lysyuk**[1, 2], **Pavel Braslavski**[3]
[1]Skoltech, [2]AIRI
[3]School of Engineering and Digital Sciences, Nazarbayev University, Astana, Kazakhstan
maria.lysyuk@skol.tech, pavel.braslavskii@nu.edu.kz

## Abstract

In this paper, we present our solution to the TextGraphs-17 Shared Task on Text-Graph Representations for Knowledge Graph Question Answering (KGQA). GPT-4 alone, with chain-of-thought reasoning and a given set of answers, achieves an F1 score of 0.78. By employing subgraph size as a feature, Wikidata answer description as an additional context, and a question rephrasing technique, we further strengthen this result. These tricks help to answer questions that were not initially answered and to eliminate irrelevant, identical answers. We have managed to achieve an F1 score of 0.83 and took 2nd place, improving the score by 0.05 over the baseline. An open implementation of our method is available on GitHub.[1]

## 1 Introduction

TextGraphs-17 task is to select a correct answer entity for a given question from a list of several Wikidata entities (Sakhovskiy et al., 2024). These lists of candidates are generated by LLMs; each list item is accompanied by a Wikidata subgraph connecting the potential answer entity to the question entities via a shortest path. Data statistics are summarized in Table 1. The task can be cast as a binary classification: each answer candidate is either correct or incorrect; the F1 score is used as the evaluation measure.

There are two main difficulties with this task. First, there are questions that have more than one correct answer. For example, the question `Who were the first two senators to represent the latest state added to the Union?` has two correct answers: `Hiram Fong (Q926441)` and `Oren E. Long (Q715129)`.

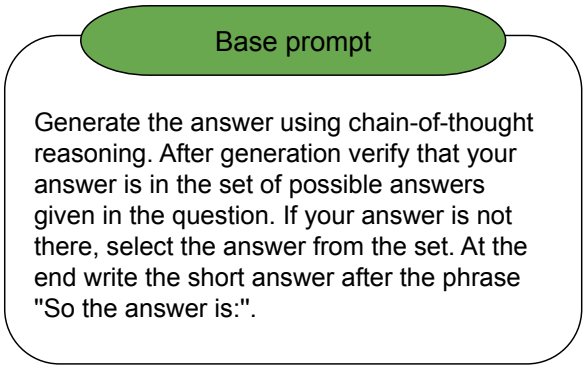
The second difficulty is that there is a significant proportion of questions with several answers with identical textual labels – 35% in the train subset and 47% in the test (see Table 1). For example, for the question `Who wrote the Leatherstocking Tales?` the correct answer is `James Fenimore Cooper (Q167856)`. However, there are two more candidate entities with exactly the same labels: `Q102502290` and `Q102502514`.[2] In such cases, the selection of the correct entity could be based on the KG subgraph analysis and/or on accounting for the descriptions of the candidate entities.

| | Train | Test |
|---|---|---|
| # questions | 3,535 | 1,000 |
| # questions with identical answers | 1,222 | 474 |
| Avg. subgraph size | 4.44 | 4.69 |
| # answers per question | 10.66 | 10.96 |

Table 1: Dataset statistics.

## 2 Method

**Baseline.** The baseline is obtained with GPT-4[3] chain-of-thought (CoT) prompting (Wei et al., 2022) with the provided set of candidate answers for the given question. The baseline prompt is as follows:



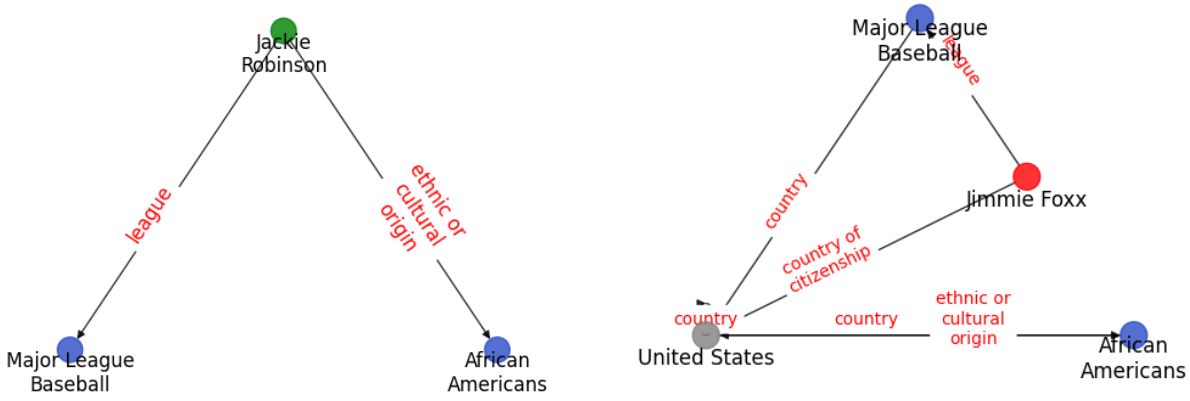

[1]https://github.com/marialysyuk/TextGraphs-17

[2]These two persons appear to be a grandson and great-grandson of the author of the *Leatherstocking Tales*, see Cooper's genealogical tree https://www.wikitree.com/wiki/Cooper-7320.

[3]We employed `gpt-4-0125-preview` model, see https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo .



*Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 149–153




(a) A subgraph for the correct answer

(b) A subgraph for an incorrect answer

Figure 1: Examples of subgraphs corresponding to correct and incorrect answers for the question `Who was the first African American baseball player to play in the American major leagues?` Color codes: **correct**, **incorrect**, **question**, **intermediate** entities.

An example of the input question with candidate answers:

**Q&A example**

Q: After publishing A Time to Kill, which book did its author begin working on immediately?

Possible answers: A Clash of Kings, A Feast for Crows, Fear and Loathing in Las Vegas, In Cold Blood, Into the Woods, Kongenes kamp, No Country for Old Men, Slaughterhouse-Five, The Firm, The Last Days of Disco.

The phrase *So the answer is:* is used as a marker to extract the final answer. The extracted answer is compared with the provided candidates after lowercasing and punctuation removed. If there is an exact match between the extracted answer and the provided candidate, this candidate is returned as the final answer.

**Post-processing: fuzzy answer matching.** The surface form of the baseline answer could be slightly different from the provided candidate answers, for example `Jerry Rice` vs. `Jerry Rice, Jr` or `Mongol Empire` vs. `The Mongol Empire`. To solve these cases, we applied fuzzy string matching with a threshold of 80.[4] In some cases GPT-4 failed to return a short answer from the list of provided candidates. For example, for the question `Who launched the third Roman invasion of Britain?` the LLM returned: `...So the answer is: The question seems to be based on a misunderstanding or mislabeling of the historical invasions of Britain as none of the options provided are known for a "third" invasion, but Claudius would be the closest in context, despite the mismatch with the question's phrasing.` In such cases, we checked whether one of the answer candidates is mentioned in the 'reasoning part'. If we could find exactly one such answer, it was taken as the correct one. Otherwise, no prediction is made, as the mention of the options could be just a part of the reasoning and thus not related to the prediction.

**Post-processing: addressing answers with identical labels.** As mentioned earlier, for some questions there are several candidate answer entities with the same textual labels, from which we only have to choose one.

Figure 1a shows the subgraph for the question `Who was the first African American baseball player to play in the American Major Leagues?` for the correct answer. Figure 1b illustrates the subgraph for the same question, but with an incorrect answer.

Let's denote the subgraph size as amount of edges in the subgraph. There is an observation that smaller subgraphs lead to correct answers with a higher probability. Indeed, compare the subgraphs in the Figures 1a and 1b – the shorter path leads to the correct answer.

[4] https://pypi.org/project/fuzzywuzzy/

| Configuration | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|
| GPT-4 (baseline) | 0.722 | 0.837 | 0.775 | 0.955 |
| + fuzzy answer matching | 0.716 | 0.857 | 0.780 | 0.955 |
| + same-text answer selection | 0.816 | 0.839 | 0.827 | 0.967 |
| **+ all tricks** | **0.823** | **0.843** | **0.835** | **0.969** |
| Llama-3-8B-Instruct (baseline) | 0.649 | 0.660 | 0.654 | 0.935 |
| + fuzzy answer matching | 0.635 | 0.703 | 0.667 | 0.935 |
| + same-text answer selection | 0.677 | 0.665 | 0.671 | 0.939 |
| **+ all tricks** | **0.697** | **0.710** | **0.704** | **0.944** |

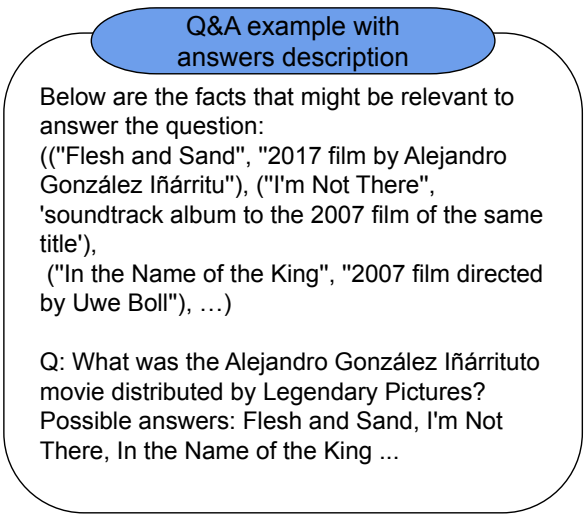
Table 2: Experiments evaluated at the post-competition stage. *All tricks* include fuzzy answer matching, same-text answer selection, original question rephrasing, and augmenting candidate answers with their Wikipedia description.

Predicting the correct answer as the one with the smallest subgraph leads to an F1 score of 0.248 on the subset of questions with textually identical answers from the training set, which is quite high for such a simple heuristic.

However, there are still cases where the candidate answers with the same labels also have subgraphs of the same size. In this case, we select the entity with a non-empty Wikidata description. If contenders still remain, we represent the question and its candidate answers' Wikidata descriptions as average fastText (Bojanowski et al., 2017) embeddings and choose the most similar pair in terms of cosine similarity. fastText model was trained on the Wikipedia data. Representing words as a collection of n-grams, fastText can capture the semantic meaning of morphologically related words, even for out-of-vocabulary words or rare words. For instance, consider the question `On which KISS album did Ace Frehley not appear, even though he was on the cover?`; there are two answers with the same label `Creatures of the Night`, but different descriptions: `1982 studio album by Kiss (Q1139397)` and `1983 single by Kiss (Q5183668)`. Both corresponding subgraphs are of size six, and the embedding of the correct answer `1982 studio album by Kiss (Q1139397)` is closer to that of the question.

**Post-processing: fixing non-answered questions with prompt tricks.** For 24 out of 1,000 questions, the proposed baseline and post-processing still result in undefined answers. We address these cases with two additional tricks. First, we task the LLM with rephrasing the question to provide additional information and make it less ambiguous, following the approach of Deng et al. (2023). The following prompt was used to rephrase the original question: `Given the above question, rephrase and expand it to help you do better answering. Maintain all information in the original question.` In this way, the original question `What fighting game did Goku not appear in?` is transformed into its rephrased version `In which fighting game is the character Goku, from the Dragon Ball series, notably absent?` As you can see, LLM adds extra information to the question about the character from the game and emphasises the "absent" with "notably" to make sense of the question more vivid. This is an example of a question where the prompt with the original question failed to produce a valid answer, while the prompt with a paraphrased question was successful.

The second trick is to augment the candidate answers with their Wikidata descriptions. In some cases, the description already incorporates the information necessary to answer the question. For example, the baseline prompt for the question `What was the Alejandro González Iñárrituto movie distributed by Legendary Pictures?` is extended as follows:



| Place | Team Name | F1 score | Precision | Recall | Accuracy |
|---|---|---|---|---|---|
| 1 | NLPeople | 0.859 | 0.867 | 0.851 | 0.974 |
| 2 | **Skoltech** | **0.830** | **0.818** | **0.843** | **0.968** |
| 3 | POSTECH | 0.816 | 0.825 | 0.807 | 0.966 |
| 4 | baseline_chatgpt | 0.680 | 0.599 | 0.786 | 0.931 |
| 5 | Team <blank> | 0.661 | 0.605 | 0.727 | 0.930 |

Table 3: Top-5 participating teams based on private test, ranked by F1 score.

As one can see from the the example, the description of the film `Flesh and Sand` includes the information of interest for the given question, making it possible to select this variant from the set of possible answers.

**Multiple correct answers.** Since the ratio of the questions with multiple correct answers wasn't large in the training data (3%), we only addressed multiple answers with identical labels. In other words, after removing answers with identical labels, multiple answers to the question could remain if they have different labels (with the respect to lower-casing and removing punctuation).

## 3 Discussion of Results

The results of different configurations on the private test set are shown in Table 2. It can be seen that fixing the problem of multiple answer entities with identical labels significantly improved the baseline. All the tricks that fixed unanswered questions with alternative to the baseline prompts helped to increase the final scores even more. In the Appendix 6 we add a description of the ideas that we tried but they didn't work.

The competition was held on Codalab.[5] The results on the private test set are presented in Table 3. Our team took the second place, showing competitive results compared to the winning team. Interestingly, the team rankings based on public vs. private test sets are quite similar, which rules out the hypothesis of overfitting on the training set.

## 4 Ablation study

Since GPT-4 is a proprietary LLM, we tried an open source LLM in the ablation study to see if the approach and tricks used in the paper were transferable to other LLMs. As a baseline, we tried the `Meta-Llama-3-8B-Instruct` model.[6]. As you can see from the table 2, all the tricks gradually improve the accuracy results for both models. Interestingly, for GPT-4 the biggest challenge was to differentiate identical labels, and the biggest increase in accuracy was achieved by solving this problem. For the Llama, on the other hand, the baseline was not as strong and additional tricks with other prompting techniques (rephrasing questions and adding candidate answers with their Wikidata descriptions) were more useful because of the high number of unanswered questions in the baseline.

## 5 Conclusion

Although LLMs are praised for their emergent properties and generalisability, they are black box models that often fall short of capturing and accessing factual knowledge (Pan et al., 2024). The TextGraphs17 shared task was an excellent competition that addresses this gap by unifying LLMs and KGs.

In this paper, we have given a description of the method used by our team, Skoltech, who came 2nd in the private test ranking with an F1 score of 0.83. We presented a method based on the GPT-4 model. The approach answers almost all questions by implementing additional prompting tricks. Furthermore, we use subgraph size and Wikidata descriptions as features that help us to distinguish between textually similar but factually different answers.

## Acknowledgments

Pavel Braslavski acknowledges funding from the School of Engineering and Digital Sciences, Nazarbayev University.

[5] https://codalab.lisn.upsaclay.fr/competitions/18214#results

[6] See https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

## References


Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Stephen Cook. 2000. The p versus np problem. *Clay Mathematics Institute*, 2:6.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*.

Maria Lysyuk, Mikhail Salnikov, Pavel Braslavski, and Alexander Panchenko. 2024. *Natural Language Processing and Information Systems: 29th International Conference on Applications of Natural Language to Information Systems, NLDB 2024, Turin, Italy, June 25-27, 2024, Proceedings*. Springer.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

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. *arXiv preprint arXiv:2210.01613*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.


## 6 Appendix

The TextGraphs-17 data relies heavily on the Mintaka dataset (Sen et al., 2022). Each question in the original Mintaka has a type annotation that can potentially be utilized for the answer selection task. There are seven question types in Mintaka: superlative, difference, multi-hop, ordinal, generic, intersection, and comparative. An attempt was made to create a few-shot prompt by providing a CoT for each type of question. Thus, in case of ordinal (`When did Metallica put out their fourth album?`) or superlative (`Which US president has had the most votes?`) questions, the options should be ranked by some parameter and the candidate at the interested position is selected. Whereas multi-hop questions (`How many kids does the lead actress of Pretty Woman have?`) require an intermediate reasoning step. For each of the seven question types, only one example was given and it might explain why this idea failed. However, it could be tried the other way round: for each type of question, a special prompt could be generated with few examples for this type of question. In addition, information about the question type is a good signal about the number of expected answers. Obviously, for comparative and superlative question types, a single answer is expected.

Another idea was inspired by the P vs NP problem (Cook, 2000) in computational complexity theory. Informally, it asks whether every problem whose solution can be quickly verified can also be quickly solved. In the baseline prompt, we provide the model with a set of answers and ask it to select one using CoT. What if we slightly change the task, and ask the model about one candidate at a time by replacing the question word by the answer candidate. For example, to the question `Who won the 1900 Election?` is turned into `Democratic Party won the 1900 Election?`. The hypothesis is that it's easier for the model to check the answer if there are fewer options. The idea didn't work probably because the modified question was not grammatically correct.

Finally, an attempt was made to reduce the number of candidate answers by filtering out wrong paths from question entities. In Konstruktor Lysyuk et al. (2024), simple questions are studied where the answer to the question is in 1-hop from the question entity. Thus, by finding the question entity and the relation (the edge between the question entity and the answer), one can find the answer. Using the ranking relation procedure from Konstruktor, we ranked the relations of the question entities. Then only the paths containing these relations remained. While reducing the number of candidate answers didn't lead to an increase in accuracy, this filtering successfully discriminated between multiple identical answers. In other words, the label of the correct answer is more likely to be on the path with selected relations than on the path with relations not selected by the ranking procedure.