HW-TSC at TextGraphs-17 Shared Task: Enhancing Inference Capabilities of LLMs with Knowledge Graphs

Wei Tang, Xiaosong Qiao, Xiaofeng Zhao, Min Zhang, Chang Su, Yuang Li, Yinglu Li, Yilun Liu, Feiyu Yao, Shimin Tao, Hao Yang, Xianghui He
Huawei Translation Services Center, Beijing, China
{tangwei133, qiaoxiaosong, zhaoxiaofeng14, zhangmin186, suchang8, liyuang3
liyinglu, liuyilun3, yaofeiyu1, taoshimin, yanghao30, hexianghui}@huawei.com

Abstract

In this paper, we present an effective method for TextGraphs-17 Shared Task. This task requires selecting an entity from the candidate entities that is relevant to the given question and answer. The selection process is aided by utilizing the shortest path graph in the knowledge graph, connecting entities in the query to the candidate entity. This task aims to explore how to enhance LLMs output with KGs, although current LLMs have certain logical reasoning capabilities, they may not be certain about their own outputs, and the answers they produce may be correct by chance through incorrect paths. In this case, we have introduced a LLM prompt design strategy based on self-ranking and emotion. Specifically, we let the large model score its own answer choices to reflect its confidence in the answer. Additionally, we add emotional incentives to the prompts to encourage the model to carefully examine the questions. Our submissions was conducted under zero-resource setting, and we achieved the second place in the task with an F1-score of 0.8321.

1 Introduction

In 2023, the widespread adoption of ChatGPT and the introduction of GPT-4 (OpenAI, 2023) marked a significant milestone in artificial intelligence (AI). GPT-4 achieved remarkable progress in the MMLU benchmark test (Hendrycks et al., 2021), demonstrating exceptional performance on various question answering (QA) and natural language inference (NLI) datasets. This breakthrough led to the emergence of large-scale language models (LLMs) like LLaMa-2 (Touvron et al., 2023), Falcon (Almazrouei et al., 2023), Gemini (Anil et al., 2023), Baichuan-2 (Yang et al., 2023), ChatGLM (Du et al., 2022), and others.

Despite the success of existing LLMs, even advanced LLMs struggle to accurately answer factual questions without a knowledge graph (KG). The answers often involve fictional or hypothetical statements or brief/trivial information. While language models can provide answers (Sen et al., 2022; Dubey et al., 2019), their quality may not meet desired standards. Addressing this challenge relies on structured knowledge sources like DBPedia (Auer et al., 2007), Wikidata (Vrandečić and Krötzsch, 2014), or NELL (Mitchell et al., 2018). This paper aims to explore and bridge this research gap.

2 Task Description

The objective of the shared task[1] (Sakhovskiy et al., 2024) is a Knowledge-based Question Answering (KBQA) problem, which aims to address the challenge of selecting the most appropriate knowledge graph (KG) entity, given a textual question and a set of candidate entities. Notably, this task incorporates a unique feature whereby each question-answer (Q-A) pair is accompanied by a graph representation consisting of shortest paths in the KG, connecting the entities mentioned in the query to the LLM-generated candidate entity, including the intermediate nodes. This provision enables participants to systematically explore and evaluate diverse text-graph fusion strategies for enhancing the performance of language model outputs in a controlled manner.

The primary goal of this task is to investigate methods for augmenting the capabilities of language models (LLMs) through the integration of KGs. To facilitate comprehensive experimentation, participants are provided with a pre-extracted graph, as there exist multiple approaches for extracting and fragmenting the text-graph modality fusion experiments. Specifically, participants are presented with the following resources:

- **Text1**: A query accompanied by a list of men-

[1]The related data for the task is publicly available at https://github.com/uhh-lt/TextGraphs17-shared-task/

Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing, pages 131–136

Query

Who was formerly an actor and now a Republican senator?

Entitie Candidates	Sub-graphs	Answer
Arnold Schwarzenegger	<Arnold Schwarzenegger, member of political party, Republican Party>, <Arnold Schwarzenegger, occupation, actor>	True
Bob Dole	<United States, described by source, Small Brockhaus and Efron Encyclopedic Dictionary>, <United States, country, United States>, <Republican Party, country, United States> <actor, described by source, Small Brockhaus and Efron Encyclopedic Dictionary>, <Bob Dole, country of citizenship, United States>, <Bob Dole, member of political party, Republican Party>	False
John McCain	<television presenter, subclass of, actor> , <John McCain, occupation, television presenter> <John McCain, member of political party, Republican Party>	False
...	...	...

Figure 1: An example of data: query, answer candidates, and respective sub-graphs. Answers are provided in the training set, but not in the testing set.

tioned Wikidata entities.

- **Text2**: 5-10 answer candidates presented as Wikidata entities.
- **Graph**: A Wikidata sub-graph comprising the shortest paths connecting the entities in the question to the candidate entities.

Among the provided candidates, one is the correct answer, while the others are incorrect. The task entails identifying the correct answer, thus entailing a binary classification objective. Furthermore, for the same query, there may be multiple entities with the same name among the provided candidate entities, while they represent different entities. Therefore, it is not feasible to solely rely on the entity names to determine the correctness of the answer. This necessitates the model to rely on the knowledge subgraph to make judgments about the correctness of the answer. A concrete data example is given in Figure 1. The evaluation metric employed for this task is the F1 score, given that the task involves binary classification.

3 Method

In this section, we will provide a detailed explanation of the proposed LLM prompt design strategy, which is based on self-rank and emotion. Additionally, we will outline the process of summarizing the outputs of LLM and generating the final submission result. Additionally, the competition task is a binary classification problem, and we employ a trick to transform it into a single-choice question, thereby avoiding the issue of the model selecting multiple correct answers for the same query. The whole process can be found in Figure 2.

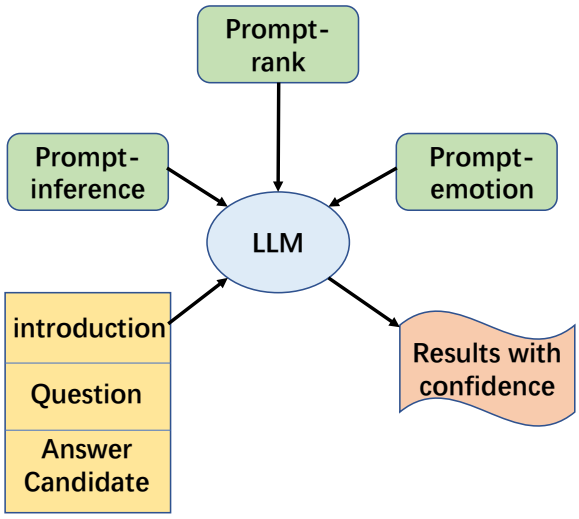

Figure 2: The whole process of our method, where yellow part denotes basic inputs, green parts denote various prompts and pink part denotes the output results.

3.1 Basic prompt-inference

Our basic prompt for inference took the following form: *"I have a new NLP reasoning task. As a smart assistant, you can help me decide which answer is correct. I will provide a question, a few answers and a few reasoning paths associated with the answers. Only one of these answers is correct. Please determine which answer is correct based on the corresponding reasoning path. Even if you believe there is no correct answer, please still choose the answer option that you think is the most*

Error types	Example	Solution
Output inconsistency	Correct: "Bob-8", Wrong: ["Bob-ID 8", "id 8", "Bob"]	regular expression
Unreasonability	"Unable to determine based on the provided reasoning paths."	random choice
Ambiguity	"Answer-1 and Answer-2 are both correct."	random choice from the two options

Table 1: Some error types of LLM-outputs, including output inconsistency, unreasonability and ambiguity with corresponding examples and solutions.

Emotion-based prompt
1. This is very import to my career.
2. You'd be better be sure about the answer.
3. Are sure that's your final answer ? It might be worth taking another look.
Ranking-based prompt
1. Give me a confidence score between 0-5 for your answer.

Table 2: Some examples of emotion-based prompts and ranking-based prompts.

plausible. The output format is: correct answer: answer-id, confidence: score:". In this prompt, we have assigned the role of Claude 3 (Anthropic, 2024) intelligent assistant and provided it with an understanding of the task's input and output. Additionally, we have imposed two constraints: (1) The answer must be inferred from the reasoning path, and (2) The answer must be unique and selected. These constraints are set based on the following considerations: (1) In some question-answering data for this task, there may be multiple candidate answers with the same entity name but different reasoning paths. Therefore, we require the model to consider the reasoning path when providing an answer. Furthermore, this is why our model answer format is "answer-id", which clearly indicate which candidate answer is chosen. (2) The second constraint ensures that the model does not give multiple answers, and when the model believes there is no correct answer, it can utilize its own knowledge to provide an approximate answer.

3.2 Prompt-rank and prompt-emotion

Although the basic prompt is enough for the LLM to output the answers. However, recent researchers (Li et al., 2023; Wang et al., 2024) have found that it can be effective to improve the response of LLM by emotional push and self-ranking push without extra model training. Inspired by these discoveries, we add prompt-rank and prompt-emotion based on the basic prompt, which now reads: *"I have a new NLP reasoning task. This task is very important to me. As a smart assistant, you can help me decide which answer is correct. I will provide a question, a few answers and a few reasoning paths associated with the answers. Only one of these answers is correct. Please determine which answer is correct based on the corresponding reasoning path. Even if you believe there is no correct answer, please still choose the answer option that you think is the most plausible. Please provide a confidence rank [A, B, C, D, E] for the larger model's answer, where A=highest confidence, E=lowest confidence. The output format is: correct answer: answer-id, confidence: score:"*, where green part and red part are emotion-based prompt and ranking-based prompt, respectively. In fact, the formats of emotion-based prompts and ranking-based prompts are very flexible. For instance, they can also be designed in the form shown in Table 2.

3.3 Refining final results

It is worth noting that our base model is Claude 3. However, considering the high cost of using Claude 3, we initially validate the effectiveness of our strategy on the open-source Mistral 7B model (Jiang et al., 2023) before migrating it to Claude 3. Another issue is that even though we have strictly define the answer out format of LLM, it is inevitably to observe LLM does not output the answer following the answer format. For example, it may directly output the answer without id, then we use regular expressions to extract the answer numbers from the non-standardized output. For clarity, we list common abnormal types of answers with corresponding examples and solutions in Table 1.

4 Results and Analysis

4.1 Overview

Table 3 presents the results of Mistrial 7B using different strategies on the test set for this task. The evaluation metrics include accuracy, precision, recall, and F1 score, with the F1 score being the primary determinant of the final ranking. As depicted in Table 3, Claude 3 employing both prompt-rank

Model	Accuracy	Precision	Recall	F1 score
Mistral 7B	0.9245	0.6650	0.3776	0.4817
Mistral 7B + Prompt-rank	0.9277	0.6819	0.4168	0.5174
Mistral 7B + Prompt-emotion	0.9268	0.6704	0.4182	0.5151
Mistral 7B + Prompt-rank + Prompt-emotion	0.9285	0.6888	0.4210	0.5226
Claude 3 + Prompt-rank + Prompt-emotion	**0.9691**	**0.8434**	**0.8211**	**0.8321**

Table 3: Results of different models with various strategies for test.

and prompt-emotion strategies achieves the highest scores across all metrics. Specifically, the prompt-rank and prompt-emotion strategies enhance the inference capabilities of the basic Mistral 7B model. The prompt-rank strategy increases the F1 score by an absolute margin of 3.57%, while the prompt-emotion strategy improves it by 3.34%. By combining these two strategies, Mistrial 7B demonstrates a 4.09% improvement. This clearly demonstrates the effectiveness of these two strategies. Additionally, the high scores obtained by Claude 3 equipped with these strategies serve as evidence of the transferability of our prompts. In terms of format, the prompt-rank strategy is designed to prompt the language models (LLMs) to rank the candidate answers, encouraging them to compare and jointly consider these answers. This approach promotes deeper inference by facilitating joint thinking. On the other hand, the prompt-emotion strategy is intended to evoke human-like emotions and encourage more cautious inference. Exploring ways to emotionally influence LLMs for better responses is an interesting direction to pursue.

4.2 Case study

Although we specify in the inference prompt that LLM should derive the answer based on reasoning paths, it is highly possible that these paths alone are inadequate for inferring the correct answer. Consequently, LLMs need to rely on their intrinsic external knowledge to enhance their inference capabilities. Figure 4 presents a similar scenario that illustrates this point. In this case, Claude 3 determines that the provided reasoning path is insufficient to find the correct answer. Therefore, it combines its intrinsic knowledge with the reasoning path to arrive at the answer. The insight from this example is that the knowledge extracted from a knowledge graph based on key entities may not be useful. Although the model can access its external knowledge, this external knowledge is outdated. In the future, we can explore whether the model can independently access APIs to find the desired knowledge, as the model's knowledge may be outdated.knowledge

Query: *Which The Lunar Chronicles books was not published in year 2015?*
Response: *"The question asks which book from The Lunar Chronicles was not published in 2015. To determine the correct answer, we need to identify which of the provided answers are actually books from The Lunar Chronicles, and then use the publication year to narrow down the answer. The reasoning paths provided offer insights into the relationship between the answers and The Lunar Chronicles, but none of them directly address the publication year. Therefore, we must rely on external knowledge about The Lunar Chronicles series by Marissa Meyer· · ·"*

Table 4: The response of Claude 3 to a specific query. The highlighted section indicates areas where Language Models (LLMs) believe external knowledge is necessary.

5 Conclusion

In this article, we propose an effective method for the TextGraphs-17 Shared Task in Knowledge-based Question Answering (KBQA). We explore the use of Claude 3 and prompt learning to enhance causal reasoning capabilities. Our research shows that incorporating ranking prompts and emotional prompts significantly improves performance. We provide reproducible experiments with extractable results using regular expressions. Due to limitations, we conducted an ablation study on Mistrial 7B instead of Claude 3 and have unresolved questions about reducing output errors in the Language Model (LLM), including inconsistencies, unreasonability, and ambiguity. These challenges require further investigation and development in future research.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Anthropic. 2024. Claude 3 haiku: our fastest model yet.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 320–335. Association for Computational Linguistics.

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II*, volume 11779 of *Lecture Notes in Computer Science*, pages 69–78. Springer.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *Preprint*, arXiv:2307.11760.

Tom M. Mitchell, William W. Cohen, Estevam R. Hruschka Jr., Partha P. Talukdar, Bo Yang, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matt Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil A. Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard C. Wang, Derry Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. 2018. Never-ending learning. *Commun. ACM*, 61(5):103–115.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 1604–1619. International Committee on Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten,

Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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

Yikun Wang, Rui Zheng, Haoming Li, Qi Zhang, Tao Gui, and Fei Liu. 2024. Rescue: Ranking llm responses with partial ordering to improve response generation. *Preprint*, arXiv:2311.09136.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.