JellyBell at TextGraphs-17 Shared Task: Fusing Large Language Models with External Knowledge for Enhanced Question Answering

Julia Belikova[1], **Evegeniy Beliakin**[1], **Vasily Konovalov**[2,1]
[1]Moscow Institute of Physics and Technology, Russia
[2]AIRI, Moscow, Russia
{belikova.iua, beliakin.eo, vasily.konovalov}@phystech.edu

Abstract

This work describes an approach to develop Knowledge Graph Question Answering (KGQA) system for TextGraphs-17 shared task. The task focuses on the fusion of Large Language Models (LLMs) with Knowledge Graphs (KGs). The goal is to select a KG entity (out of several candidates) which corresponds to an answer given a textual question. Our approach applies LLM to identify the correct answer among the list of possible candidates. We confirm that integrating external information is particularly beneficial when the subject entities are not well-known, and using RAG can negatively impact the performance of LLM on questions related to popular entities, as the retrieved context might be misleading. With our result, we achieved 2nd place in the post-evaluation phase.

1 Introduction

While LLM can provide answers to questions, answering factoid questions without access to a KG can be challenging. It has been shown that incorporation of the KG information into LLM significantly improves the results for various NLP tasks (Zhang et al., 2020).

The TextGraph-17[1] workshop focuses on exploring synergies between text and graph processing techniques, specifically targeting the fusion of LLMs with KGs. The shared task presents a novel challenge in the domain of KGQA, where participants are tasked with selecting the correct KG entity corresponding to a textual question, given a set of candidate entities and a graph of shortest paths in the KG connecting the query entities to the LLM-generated candidates. The shared task aims to investigate effective strategies for fusing text and graph modalities, providing a controlled environment for experimentation. By pre-extracting the graph data, the organizers facilitate a standardized testbed, mitigating variations due to different graph extraction methods and enabling researchers to concentrate on enhancing LLM outputs with KG information. Overall, this shared task contributes to advancing the understanding and practical application of LLM-KG integration for improved QA performance.

Our main contributions are three-fold:

1. We show that LLMs do partially incorporate knowledge about Wikidata.
2. We confirm that the QA capability of LLMs can be enhanced by supplying them with relevant external data.
3. We demonstrate that by leveraging UE techniques, we can efficiently combine multiple LLMs, each integrated with distinct external data sources.

2 Related Work

Early approaches in KGQA primarily focused on simple questions involving node-edge-node triples, but the complexity increases with multi-hop and aggregation queries. Izacard and Grave (2021) achieved state-of-the-art results on benchmarks like Natural Questions (Kwiatkowski et al., 2019) and TriviaQA by integrating Wikipedia as an external knowledge source. Similarly, Talmor and Berant (2018) showed how web-search results could enhance the performance of QA systems on complex queries from the ComplexWebQuestions benchmark.

Hybrid systems combining text and graph-based information have been particularly effective for complex multi-choice question answering (MCQA) tasks. PullNet (Sun et al., 2019) and GraftNet (Sun et al., 2018) employ relational graph convolutional networks to iteratively retrieve relevant information from both text and KGs, improving the handling of multi-hop questions. Additionally, using KG embeddings has been a successful strategy, as demon-

[1]https://sites.google.com/view/textgraphs2024

Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing, pages 154–160

strated by Huang et al. (2019), who enhanced candidate retrieval for answers, and Chekalina et al. (2022), who introduced a memory-efficient representation for KG embeddings, improving link prediction and QA tasks.

The fusion of LLMs with KGs has proven especially beneficial for MCQA. The GETT-QA (Banerjee et al., 2023) uses T5 to convert questions into simplified SPARQL queries, which are then mapped to KG entities and relations through a post-processing step, enhancing accuracy by leveraging both the linguistic capabilities of T5 and the structured information in KGs. Furthermore, UniK-QA (Oguz et al., 2022) creates representations for both structured and unstructured knowledge, facilitating open-domain QA over diverse data sources.

3 Dataset

The KGQA dataset is designed for the task of extracting accurate answers from complex knowledge graphs, specifically using information from Wikidata. Each data instance consists of a textual question that contains a list of referenced Wikidata entities. Along with this, there are several candidate answer options, all presented as distinct Wikidata entities. A key feature of the dataset is the provision of a sub-graph extracted from Wikidata, which comprises the shortest paths connecting the entities mentioned in the question to those found within the answer candidates.

The training set includes a substantial amount of 37,672 samples with 3,535 unique questions. Whereas the test set contains 10,961 samples with 1,000 unique questions. The dataset also ensures a balance in terms of candidate answers, with a minimum of 6 options and a maximum of 20 per question (*"Which Stephen King books have not been made into movies yet?"*), making it a challenging yet versatile resource for developing QA systems. The majority of questions (3,425) in the training set have a single correct answer; however, 110 questions have more than one correct answer. Therefore, the primary objective is to classify the answers into two categories: correct or incorrect.

4 Proposed Approach

We based our solution on Llama 3 series of LLMs in 8 billion (8B) and 70 billion (70B) parameter sizes. These models are specifically designed for dialogue applications and have demonstrated superior performance compared to popular open-source chat models on standard industry benchmarks (AI@Meta, 2024).

4.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) combines the strengths of LLMs with information from external databases to boost the precision and reliability of generated content, especially for tasks demanding substantial knowledge base. By facilitating seamless updates and integration of specialized data, RAG effectively fuses the internal knowledge of LLMs with the extensive and ever-evolving external data reservoirs, creating a synergy that enhances performance.

Wikidata ID description As the simplest form of external knowledge augmentation, we incorporated the Wikidata ID and answer candidate description from Wikidata.

Web-search results As external knowledge, we used web-search results from DuckDuckGo (Parsania et al., 2016). DuckDuckGo aims to deliver relevant results while respecting user privacy. In addition, DuckDuckGo doesn't require an API key and doesn't apply any limitations on getting web results (10 search results were returned for each query). The prompt including the web-search results can be found in the Appendix A.

Textualized KG Furthermore, we incorporated the subgraphs provided by the organizers as an external knowledge source. For textualizing the graph, we opt to use Llama 3 70B with the following prompt (Wu et al., 2023):

Prompt

Transform this wiki graph into text. Write only the new string that contains the text representation of the graph.

The prompt with textualized graph can be found in the Appendix A.

4.2 Uncertainty Estimation

Uncertainty Estimation (UE) refers to the process of measuring the level of confidence in the predictions generated by a LLM. Initially, UE was employed to identify hallucinations, which are instances where the model fabricates facts without offering users a clear way to assess the truthfulness of its statements (Maksimov et al., 2024). Typically, UE involves calculating it for an entire sequence, requiring us to aggregate the uncertainties

Model	RAG	Precision	Recall	F1	Accuracy
Llama3 8B	–	48.85	47.61	48.22	90.46
	D	51.77	49.95	50.84	90.98
	G + D	42.54	41.54	42.04	89.31
	W + D	60.38	58.84	**59.60**	92.55
	G + W + D	42.54	41.54	42.03	89.30
Llama3 70B	–	74.80	73.41	74.10	95.21
	D	77.19	75.76	76.47	95.65
	G + D	75.60	74.19	74.89	95.36
	W + D	82.05	79.96	**80.99**	96.50
	G + W + D	79.52	77.42	78.45	96.03

Table 1: Evaluation results for two Llama 3 scales (8B and 70B) are as follows. RAG denotes the knowledge sources, where D refers to the textual description of answer candidates from Wikidata, W represents DuckDuckGo web-search results (with the query being the question), and G signifies the textualized graph representation provided in the dataset. When equipped with knowledge from web-search results and Wikidata answer candidate descriptions, Llama 3 outperforms all other external knowledge sources. F1 score serves as the primary competitive metric.

associated with numerous individual token predictions. This often necessitates the use of sophisticated sampling and pruning strategies, such as beam search. However, in our specific scenario, the number of potential prediction choices is fixed and limited by number of answer candidates. As a result, the uncertainty estimation process becomes significantly more streamlined and straightforward. Specifically, we utilized white-box UE methods, including **maximum probability** (Fadeeva et al., 2023) and **margin probability** (Kuhn et al., 2023), i.e. the difference between the probability of the most likely answer and the probability of the second most likely answer.

5 Results and Discussion

Before providing the main results, we first examine the inherent knowledge of the LLM concerning Wikidata entities.

What LLM knows about Wikidata? To demonstrate the inherent capability of LLMs to link a Wikidata entity with its corresponding Wikidata ID, we carried out two fundamental experiments. These involved prompting Llama 3 70B to generate both entity IDs and entities from IDs. However, it emerged that predicting an entity from its ID often led to inaccuracies, with the model succeeding mainly in associating IDs with the most well-known entities, such as *Barack Obama*, *World War II*, *Washington, D.C.*, *Italy*. Moreover, when prompted about being pretrained on Wikidata, Llama 3 confirms positively.

Does the size of LLM make a difference when utilizing non-parametric knowledge? Table 1 presents a comparison of Llama 3 models with varying sizes. While employing external knowledge, one might anticipate that both Llama 3 scales would exhibit similar performance; however, this is not the case. The larger Llama 3 70B model consistently surpasses the smaller Llama 3 8B model across all scenarios. This suggests that the size of the language model remains crucial even with external knowledge, owing to its extra parametric knowledge or improved reasoning abilities.

Whether all external data are equally helpful? Table 1 showcases a comparison of Llama 3 models with different external knowledge sources. Llama 3 augmented with external knowledge in both scales outperforms Llama 3 without it.

Incorporating descriptions from Wikidata is particularly helpful for distinguishing answer candidates with the same name (*Bob Dylan* – Q392, *Bob Dylan* – Q251309).

The Llama 3 70B model, augmented with web-search results and descriptions, surpasses all other approaches, including the Llama 3 70B model that was provided with descriptions, web-search results, and a textualized knowledge graph. This implies that incorporating more diverse knowledge might potentially confuse the model, leading to a decrease in performance.

Furthermore, we establish a correlation between the external knowledge utilized and the correctness

Model	UE	Precision	Recall	F1	Accuracy
Llama3 70B	max prob	83.83	82.11	82.96	96.85
	margin prob	84.23	82.50	83.35	96.92
$\text{Baseline}_{\text{chatgpt}}$	–	58.11	78.18	66.67	92.73
$\text{Best}_{\text{private test}}$	–	86.67	85.14	**85.90**	97.39

Table 2: The evaluation results compare two strategies for combining outputs from three Llama 3 70B models, each enhanced with distinct knowledge resources (description, web-search, and knowledge graph). The margin probability strategy involves selecting an answer from the LLM where the difference $p(top_1) - p(top_2)$ is maximum for a given sample. This ensemble strategy surpasses all other aggregation methods, demonstrating that different external knowledge sources can be advantageous for various questions. F1 score serves as the primary competitive metric.

of answers based on their popularity[2] measured on the training set. Consequently, Llama 3 70B, which uses solely entity descriptions as input, accurately classifies approximately 37% of entities with a popularity score below the median and 63% of those with a score equal to or above the median. On the other hand, the model incorporating web-search context in addition to entity descriptions correctly classifies around 45% of less popular entities and 55% of more popular entities, respectively.

What is the best way to combine LLMs with different external knowledge sources? As shown in the Table 1, integrating all three external knowledge sources degrades the performance. Nevertheless, each knowledge resource offers unique information that can be beneficial for answering certain questions while hindering performance on others. Table 2 compares two strategies for combining outputs from three Llama 3 70B models, each fortified with different knowledge resources (description, web-search, and knowledge graph). The **margin probability** approach entails choosing an answer from the LLM where the difference $p(top_1) - p(top_2)$ reaches its maximum for a given sample. This ensemble strategy outperforms alternative aggregation techniques, highlighting that diverse external knowledge sources can be advantageous for distinct questions.

Also worth noting is the contribution of the various data sources to the final results. Using both uncertainty estimations, the proportion of predictions by Llama 3 70B enhanced with knowledge graph is about 10%, and the web-search and description are about 44% and 46% respectively, with the margin probability estimation using slightly fewer predictions made by incorporating web-search.

[2]The popularity is estimated by the number of views of the corresponding Wiki page per month for the first half of 2023.

The proposed method significantly outperforms the ChatGPT-based baseline; however, it still lags behind the top-performing system.

6 Conclusion

In this paper, we detailed the system submitted for the TextGraph-17 workshop, focusing on the development of KGQA system. We introduced a straightforward yet efficient Llama-3-based pipeline. Our study investigated how incorporating external knowledge into a LLM can notably enhance KGQA performance. We showed that a marginal probability combination of three Llama 3 70B models employing different external resources outperforms all baselines and attained a score comparable to the 2nd place in the post-evaluation phase.

For future work, we propose testing various LLMs at different scales and exploring different methods for integrating external knowledge. Additionally, we aim to compare open LLMs with proprietary LLMs. Furthermore, we did not examine the multilabel capability of our solution, and a comprehensive error analysis is essential.

The proposed KBQA approach can be employed independently or integrated within a NLP framework (Burtsev et al., 2018).

Acknowledgments

This work was supported by a grant for research centers, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730324P540002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138.

References

AI@Meta. 2024. Llama 3 model card.

Debayan Banerjee, Pranav Ajit Nair, Ricardo Usbeck, and Chris Biemann. 2023. GETT-QA: graph embedding based T2T transformer for knowledge graph question answering. In *The Semantic Web - 20th International Conference, ESWC 2023, Hersonissos, Crete, Greece, May 28 - June 1, 2023, Proceedings*, volume 13870 of *Lecture Notes in Computer Science*, pages 279–297. Springer.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, and Vasily Konovalov. 2018. Deeppavlov: An open source library for conversational ai. In *NIPS*.

Viktoria Chekalina, Anton Razzhigaev, Albert Sayapin, and Alexander Panchenko. 2022. MEKER: memory efficient knowledge embedding representation for link prediction and question answering. *CoRR*, abs/2204.10629.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. Lm-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 446–461. Association for Computational Linguistics.

Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 105–113, New York, NY, USA. Association for Computing Machinery.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Ivan Maksimov, Vasily Konovalov, and Andrei Glinskii. 2024. DeepPavlov at SemEval-2024 task 6: Detection of hallucinations and overgeneration mistakes with an ensemble of transformer-based models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 274–278, Mexico City, Mexico. Association for Computational Linguistics.

Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Sejr Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2022. Unik-qa: Unified representations of structured and unstructured knowledge for open-domain question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1535–1546. Association for Computational Linguistics.

Vaishali S Parsania, Foram Kalyani, and Krunal Kamani. 2016. A comparative analysis: Duckduckgo vs. google search engine. *GRD Journals-Global Research and Development Journal for Engineering*, 2(1):12–17.

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.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242, Brussels, Belgium. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 641–651. Association for Computational Linguistics.

Yike Wu, Nan Hu, Sheng Bi, Guilin Qi, Jie Ren, Anhuan Xie, and Wei Song. 2023. Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering. *CoRR*, abs/2309.11206.

Zhiyuan Zhang, Xiaoqian Liu, Yi Zhang, Qi Su, Xu Sun, and Bin He. 2020. Pretrain-KGE: Learning knowledge representation from pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 259–266, Online. Association for Computational Linguistics.

A Prompts

Prompt with web-search context

You are a helpful assistant. You must follow the rules before answering:
- A question and its answer options will be provided.
- The question has only one correct option.
- The correct answer is always given.
- Write only the number of the correct option.
- If you do not know the answer, write only the number of the most likely one.

Below are the facts that might be relevant to answer the question:
1. Review by Karin Tanabe. October 18, 2023 at 11:00 a.m. John Grisham ...
2. Some of his most famous books include ...
If there is no relevant fact, rely on your knowledge or choose a more likely option.

Question:
"After publishing A Time to Kill, which book did its author begin working on immediately?"
Options:
0. {"answer": "A Feast for Crows", "WikiDataID": "Q1764445"}
1. {"answer": "Fear and Loathing in Las Vegas", "WikiDataID": "Q772435"}...

Prompt with textualized graph

You are a helpful assistant. You must follow the rules before answering:
- A question and its answer options will be provided.
- The question has only one correct option.
- The correct answer is always given.
- Write only the number of the correct option.
- If you do not know the answer, write only the number of the most likely one.

Question:
"In Harry Potter literature series wrote by J.K. Rowling, which follows Harry Potter and the Philosopher's Stone?"
Options:
0. {"answer": "Half-blood Prince", "WikiDataID": "Q10355035", "WikiDataGraph": "Harry Potter and the Half-Blood Prince is a book in the Harry Potter universe, written by J. K. Rowling and part of the Harry Potter series..."}
1. {"answer": "Harry Potter", "WikiDataID": "Q8337", "WikiDataGraph": "J. K. Rowling wrote Harry Potter and Harry Potter and the Sorcerer's Stone..."}...