Learn Together: Joint Multitask Finetuning of Pretrained KG-enhanced LLM for Downstream Tasks

Anastasia Martynova
Sber AI, HSE
Moscow, Russia
ans.martynova@gmail.com

Vladislav Tishin
Sber AI / Moscow, Russia
ITMO / St. Petersburg, Russia
vvikttishin@sberbank.ru

Natalia Semenova
Sber AI, AIRI
Moscow, Russia
semenova.bnl@gmail.com

Abstract

Recent studies have shown that a knowledge graph (KG) can enhance text data by providing structured background knowledge, which can significantly improve the language understanding skills of the LLM. Besides, finetuning of such models shows solid results on commonsense reasoning benchmarks. In this work, we introduce expandable Joint Multitask Finetuning of Pretrained KG-enchanced LLM approach for Question Answering (QA), Machine Reading Comprehension (MRC) and Knowledge Graph Question Answering (KGQA) tasks. Extensive experiments show competitive performance of joint finetuning QA+MRC+KGQA over single task approach with a maximum gain of 30% accuracy.

1 Introduction

Large language models (LLMs), pretrained on extensive text corpus, have demonstrated high performance across a wide range of natural language processing (NLP) tasks. However, despite their success in various applications, these models have notable shortcomings. Studies show that LLMs frequently fail to accurately recall factual information and tend to generate hallucinations - statements that are false or misleading. Furthermore, LLMs pretrained on general text data may not effectively apply domain-specific knowledge without additional training on relevant datasets.

To improve the efficiency of large language models (LLMs) and address the aforementioned challenges, a promising solution is to integrate LLMs with knowledge graphs (KGs). Knowledge graphs represent factual information in a structured format, using triples composed of a head entity, a relation, and a tail entity. KGs are widely applied across various domains due to their structured, interconnected representation of data, offering a more comprehensive and interpretable view of information and facilitating easier interaction with it.

There are several strategies to integrate LLMs with KGs (Pan et al., 2024). The first approach involves enhancing large language models using knowledge graphs. In this approach, KGs can be incorporated during the pretraining and inference stages of the LLM to enrich its linguistic representations with external knowledge and provide insights into its reasoning process. For example, the ERNIE (Zhang et al., 2019) and KALM (Corby Rosset, 2021) architectures leverage this method by feeding pairs of sentences and corresponding entities from the knowledge graph into the LLM, subsequently training the model to predict relationships between these entities.

The second approach takes the opposite direction—strengthening knowledge graphs using LLMs. This technique aims to enhance the productivity of KGs and improve their performance in KG-related tasks. For example, the authors of the QA-GNN (Yasunaga et al., 2021) architecture employ a graph neural network (GNN)-based model to jointly analyze the input context and KG information through message passing. The input text information is transformed into a special node via a pooling operation and then connected with other entities in the KG. Another model, GreaseLM (Zhang et al., 2021), facilitates deeper interaction between text tokens and KG entities. Information from both modalities propagates to each other, allowing representations of linguistic context to be grounded in structured world knowledge and enabling linguistic nuances in context to inform graph knowledge representations.

Another increasingly popular way to leverage the benefits of LLMs and KGs simultaneously is to integrate these models into a single framework where they can mutually reinforce each other. In this framework, LLMs are used to understand natural language, while KGs serve as a knowledge base providing factual information. The DRAGON

Proceedings of the Generative AI and Knowledge Graph Workshop (GenAIK), pages 13–19
January 19, 2025.

(Deep Bidirectional Language-Knowledge Graph Pretraining) architecture (Yasunaga et al., 2022) exemplifies this approach by pretraining a deeply integrated language-knowledge foundation model using both text and KGs at scale. This self-supervised model processes text segments and their corresponding KG subgraphs, integrating information from both modalities bidirectionally.

Since DRAGON demonstrated superior performance in commonsense reasoning and tasks involving complex reasoning compared to the QA-GNN and GreaseLM baselines, we decided to investigate the effectiveness of this model within the frameworks of the Machine Reading Comprehension (MRC) task, the Knowledge Graph Question Answering (KGQA) task, and the combined MRC+QA+KGQA task. This study tests the hypothesis that training on the combined MRC+QA+KGQA task will yield better performance, as the model learns to solve tasks of different types, which in turn aids in solving each task individually. By leveraging the complementary strengths of both textual and structured knowledge understanding, the integrated approach is expected to enhance the model's reasoning capabilities.

2 Related Work

In this work, we consider tasks from the field of natural language processing, such as MRC, QA and KGQA. These tasks are crucial as they represent key challenges in language understanding, demanding models to comprehend, interpret, and interact with text in a meaningful way. Addressing these tasks advances neural networks' capabilities in processing human language.

The Machine Reading Comprehension (MRC) task involves developing systems capable of automatically understanding and processing textual passages to accurately answer questions about the content. This necessitates advanced natural language processing techniques to capture the semantics, context, and nuances of the text, facilitating effective question answering. Prominent large language models, including GPT-4 (Achiam et al., 2023), PaLM 2 (Anil et al., 2023), and Claude 2 from Anthropic, have demonstrated consistently high performance in this domain.

Question answering (QA) is the process of providing answers to asked questions, while refraining from attempting to answer questions outside the context provided. In addition to large language models trained using few-shot learning (similar to the MRC task), architectures with fewer parameters can also handle the QA task effectively.

For example, GrapeQA (Taunk et al., 2023) enhances commonsense question-answering by combining pretrained Language Models with Knowledge Graphs reasoning. It addresses two key challenges faced by typical approaches: difficulty in capturing all QA information in the Working Graph (WG) and inclusion of irrelevant KG nodes. GrapeQA introduces two improvements to the WG: prominent entities for graph augmentation identifies relevant text chunks from QA pairs and augments the WG with corresponding LM latent representations, and context-aware node pruning removes less relevant nodes. These enhancements allow GrapeQA to consistently outperform its predecessor QA-GNN (Yasunaga et al., 2021), demonstrating notable improvements on datasets such as OpenBookQA (Mihaylov et al., 2018) and CommonsenseQA (Talmor et al., 2019).

Another model, KEAR (Xu et al., 2022), extends the transformer architecture with an external attention mechanism, integrating external knowledge from sources like knowledge graphs, dictionaries, and training data. This additional knowledge is retrieved using the input as the key and then integrated with the input. KEAR achieves this without altering the model architecture, opting for text-level concatenation for external attention.

Knowledge Graph Question Answering (KGQA (Yang et al., 2014)) involves the task of responding to natural language queries by utilizing the structured information stored within a knowledge graph. The goal is to provide accurate and contextually appropriate answers to a wide range of natural language questions by effectively navigating the interconnected nodes and relationships within the knowledge graph.

Methods for solving KGQA problems can be broadly categorized into two types: Information Retrieval-based (IR-based) and Semantic Parsing-based (SP-based). SP-based methods adopt a parse-then-execute approach, starting with semantic analysis to parse the relations and entities in complex questions. Next, they construct logical formulas by translating the subgraph into an executable format, such as SPARQL. Finally, these methods use the query language to interact with the Knowledge Graph, retrieving and presenting the results. While SP-based methods are often praised for their interpretability due to the intermediate step of generat-

ing detailed logic forms, they face computational challenges, especially with complex questions that involve multiple relations. This results in a larger search space and increased computational cost.

IR-based approaches to Knowledge Graph Question Answering (KGQA) typically involve several steps. First, they extract a question-specific subgraph from the knowledge graph, including all relevant entity nodes and relation edges without generating an executable logic formula. Next, they use a question representation module to encode user-question tokens into low-dimensional vectors. Following this, an extracted-graph-based reasoning module applies a semantic matching algorithm to aggregate information from the subgraph, concentrating on the neighborhood of the central entity. Finally, an answer-ranking module ranks the entity scores within the subgraph to predict the top-ranked entities as the final answers.

In developing our own approach for multitask finetuning, which integrates a language model with a knowledge graph, several foundational IR-based methods for the KGQA task were considered. For instance, Rce-KGQA (Jin et al., 2022) focuses on enhancing reasoning by leveraging both explicit and implicit relational chains within the knowledge graph. EmbedKGQA (Saxena et al., 2020) addresses knowledge graph sparsity by integrating external knowledge and utilizing KG embedding techniques, which improves performance in multi-hop KGQA tasks. SRN (Qiu et al., 2020) approaches KGQA as a sequential decision problem and employs reinforcement learning to effectively search for paths within knowledge graphs. Additionally, KVMemNN (Eric et al., 2017) introduces a key-value retrieval mechanism that enables neural dialogue agents to interact seamlessly with knowledge bases across various domains. These methods were considered for their valuable concepts to form a strategy for finetuning the model on the KGQA and joint task. Our study is not intended to be a direct comparison with all the models listed in this section.

3 Methodology

3.1 Encoders

DRAGON, as previously described, is the basis for all experiments conducted in this study. At the first stage of the study, we tested the lightweight T5-base (Raffel et al., 2020) encoder but we received insufficient results. We used RoBERTa-large (Liu et al., 2019) encoder for input text data and text data from the knowledge graph.

3.2 Datasets & Metrics

Variations of the basic DRAGON method with different encoders are pretrained on a dataset consisting of pairs of "text data + knowledge graph data." The pretraining dataset was derived from the large text corpus BookCorpus and the ConceptNet knowledge graph. BookCorpus is a comprehensive English-language dataset containing 11,038 unpublished books (approximately 74 million sentences) across 16 different subgenres, including romance, history, adventure, and others. ConceptNet is one of the most widely used general knowledge graphs, comprising about 300,000 nodes. Preprocessing involved extracting a subgraph from ConceptNet, containing all concepts mentioned in each line of text from BookCorpus. This process took approximately two weeks and was executed on a CPU.

After that we finetune and evaluate the pretrained DRAGON general domain model[1] on three downstream tasks: Question Answering, Machine Reading Comprehension and Knowledge Graph Question Answering as single tasks and also make joint finetuning. Finetuning was carried out on monolingual (English) datasets. We followed the commonsense reasoning benchmark setup with accuracy metric (Talmor et al., 2019).

The following datasets were used for finetuning on the QA task:

1. CommonsenseQA (Talmor et al., 2019) is a dataset for multiple-choice question answering, designed to assess various facets of commonsense knowledge necessary for predicting correct answers. It contains 12,102 questions with four distractor answers and one correct answer.

2. OpenBookQA (Mihaylov et al., 2018), inspired by open book exams, assesses human understanding in specific subjects. It contains 5,957 elementary-level science questions, probing comprehension of 1,326 core science facts and their applications. The dataset maps each question to a core fact for targeted training.

To finetune on the MRC task, the subsequent dataset was utilized:

[1] https://github.com/michiyasunaga/dragon

1. DREAM (Sun et al., 2019) is a multiple-choice Dialogue-based REAding comprehension exaMination dataset, distinct from existing reading comprehension datasets by its focus on comprehensive multi-turn multi-party dialogue comprehension. It comprises 10,197 multiple-choice questions extracted from 6,444 dialogues, sourced from English-as-a-foreign-language exams curated by human experts.

Finally, for KGQA task finetuning we used the following dataset:

1. KQA Pro (Cao et al., 2022) is a large-scale dataset designed for intricate question answering over knowledge base. Its questions are remarkably diverse and demanding, calling for various reasoning abilities, such as compositional reasoning, multi-hop reasoning, quantitative comparison and set operations. The target knowledge base of KQA Pro comprises a dense subset of Wikidata. The dataset is divided into training, validation, and test sets, with 94376, 11797, and 11797 questions respectively.

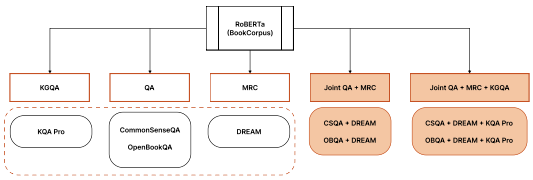

Figure 1: Scheme for finetuning the basic approach with variations of encoders on datasets for QA, MRC, KGQA tasks and their combinations.

3.3 Data Preprocessing

Note that all of the listed datasets for the QA and MRC tasks were also preprocessed with the ConceptNet graph. Specifically, for each question and answer choices, concepts from the original knowledge graph were searched and a subgraph was compiled. KG node embeddings in the case of finetuning on the QA and MRC datasets were initialized with pre-computed ConceptNet entity embeddings as proposed in the MHGRN (Feng et al., 2020) method. This scheme involves converting triplets from KG into sentences. The resulting sentences are then passed to BERT-Large (Devlin et al., 2019) to calculate the embeddings for each sentence. Finally, for each entity, all sentences containing that entity are collected, all token representations of the entity's mention spans in those sentences are retrieved, and the average pooling of these representations is returned.

In the case where the dataset contained only the correct answer, 4 incorrect answer choices for each question were generated using the Meta-Llama-3-8B-Instruct model in order to follow the pattern of questions and answers used in other datasets.

To train model on KGQA task, a subset containing 12,102 questions was extracted from the KQA Pro dataset, since the original KQA Pro is too large. This approach resolved the issue of dataset size discrepancies across tasks and simplified the selection of learning rate and batch size for training. Preprocessing for the KQA Pro subset was performed using a subgraph from the Wikidata knowledge graph that was provided by the authors of KQA Pro which we matched by entities and links to ConceptNet to make a proper grounding.

3.4 Joint MRC & QA finetuning

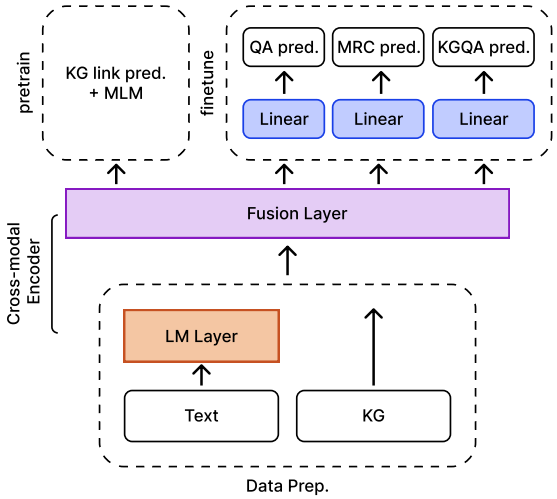

Figure 2: Multitask finetuning scheme for joint MRC+QA+KGQA task. The framework provides training the main KG-enchanced LLM.

Figure 2 depicts our approach to finetuning DRAGON pretrained model on QA, MRC and KGQA tasks, except for the grounding step where we put together text tokens and KG nodes. After that, we fed these pairs to the fusion layer, which is a cross-modal encoder that bidirectionally exchanges information between text and node representations. We used a linear combination of Cross

Entropy loss (Good, 1952) for each task in the following way:

$$\mathcal{L} = \mathcal{L}_{QA} + \mathcal{L}_{MRC} + \mathcal{L}_{KGQA} \quad (1)$$

3.5 Experiments & Results

Dataset Combination	Dev Acc.	Test Acc.
CSQA	0.755	0.689
CSQA + DREAM	**0.783**	**0.73**
CSQA + DREAM + KQA Pro	0.7707	0.6954

Table 1: Comparison of accuracy metrics resulting from finetuning on only one QA task (on the CSQA dataset) and metrics resulting from finetuning on a combination of tasks (DREAM, ARC - MRC task, KQA Pro - KGQA task). Validation and testing was performed on the QA task.

Dataset Combination	Dev Acc.	Test Acc.
OBQA	0.62	0.632
OBQA + DREAM	**0.744**	**0.720**
OBQA + DREAM + KQA Pro	0.67	0.678

Table 2: Comparison of accuracy metrics resulting from finetuning on only one QA task (on the OBQA dataset) and metrics resulting from finetuning on a combination of tasks (DREAM, ARC - MRC task). Validation and testing was performed on the QA task.

Dataset Combination	Dev Acc.	Test Acc.
DREAM	0.4098	0.419
DREAM + CSQA	0.704	0.722
DREAM + OBQA	0.702	0.696
DREAM + CSQA + KQA Pro	**0.7225**	**0.7227**

Table 3: Comparison of accuracy metrics resulting from finetuning on only one MRC task (on the DREAM dataset) and metrics resulting from finetuning on a combination of tasks (CSQA, OBQA - MRC task, KQA Pro - KGQA task). Validation and testing was performed on the MRC task.

Based on the results, DRAGON with a RoBERTa-large encoder was chosen as the most efficient architecture with which further experiments were carried out.

Tables 1-4 show overview of performance DRAGON with the RoBERTa-large encoder on QA, MRC, KGQA, QA+MRC and QA + MRC + KGQA tasks.

With finetuning on the CSQA (QA task), DREAM (MRC task) and KQA Pro (KGQA task)

Dataset Combination	Dev Acc.	Test Acc.
KQA Pro	0.5512	0.5728
KQA Pro + CSQA	0.5611	0.6
KQA Pro + DREAM	0.5610	0.5702
KQA Pro + CSQA + DREAM	**0.6085**	**0.6211**

Table 4: Comparison of accuracy metrics resulting from finetuning on only one KGQA task (on the KQA Pro dataset) and metrics resulting from finetuning on a combination of tasks (DREAM - MRC task, CSQA - QA task). Validation and testing was performed on the KGQA task.

datasets, a significant increase in the metric was shown when testing the model for the MRC task. Finetuning only for the MRC task on the DREAM dataset allows us to achieve an accuracy metric of 41.9%. With finetuning for QA and KGQA tasks, the metric becomes equal to 72.27%. Training on the CSQA+DREAM+KQA Pro datasets also gave an increase for the KGQA task. The metric increased from 57.28% to 62.11%.

Significant gains were demonstrated on the CSQA+DREAM dataset for the QA task. Accuracy with multi-task finetuning increased by 4.1% compared to finetuning only for the QA task on the CSQA dataset (from 68.9% to 73%).

Detailed analysis revealed that KG-pretrained model finetuned on QA task improved the generalization of MRC in our experiments. Unlike QA datasets MRC task assumes to proceed with rather large text passages which is challenging for LLM even with KG-pretraining. Our experiments reported that using long enough texts from the MRC dataset improves the ability of the model to answer more complicated questions. Our code is available at github repository[2]

4 Discussion & Limitations

All finetuning experiments were performed on English-language datasets. This is due to the fact that searching for datasets with overlapping languages for all three tasks (QA, MRC and KGQA) is difficult. Thus, there are a large number of multilingual datasets for QA and MRC tasks, and a limited number of such datasets for the KGQA task, which is associated with the difficulty of translating the entire knowledge graph into another language. At the moment, we were only able to use the QALD-9-Plus (Perevalov et al., 2022) dataset, which con-

[2]Github repository

tains questions in 10 languages, but has a small size (about 1.5 thousand questions), as well as the MLPQ (Tan et al., 2023) dataset, which covers only Chinese, English and French, but contains about 300 thousand questions.

Moreover, we made a test to check if the expand of data for the same task can make the same improvement as the adding other nlp-task. We trained our backbone on two QA datasets - CSQA and OBQA but the performance become 6 to 10% worse than both of them in pair with MRC dataset DREAM.

5 Conclusion & Future work

The paper presented expandable Joint Multitask Finetuning on Pretrained KG-enhanced LLM approach which aims to improve the performance of language models in a variety of language understanding tasks. We proposed a new multitask learning framework which jointly finetunes a language model with a knowledge graph enhanced objective on a few tasks and easily expanded to new nlp-tasks. The paper provides a detailed description of the proposed approach and presents experimental results demonstrating its effectiveness. Our results show strong improvements of synergized QA, MRC and KGQA tasks on each other with a maximun gain of 30% accuracy. We plan to extend our experimental setup by pretrained LLaMa v2 encoder. We also plan to conduct finetuning experiments on multilingual datasets containing English, Chinese, French, and other languages.

References

OpenAI Josh Achiam, Steven Adler, and Sandhini Agarwal et al. 2023. Gpt-4 technical report.

Rohan Anil, Andrew M. Dai, Orhan Firat, and Melvin Johnson et al. 2023. Palm 2 technical report. *ArXiv*, abs/2305.10403.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. 2022. KQA Pro: A large diagnostic dataset for complex question answering over knowledge base. In *ACL'22*.

Minh Phan Xia Song Paul Bennett Saurabh Tiwary Corby Rosset, Chenyan Xiong. 2021. Knowledge-aware language model pretraining. arXiv:2007.00655. Version 2.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1295–1309. Association for Computational Linguistics.

I. J. Good. 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114.

Weiqiang Jin, Biao Zhao, Hang Yu, Xi Tao, Ruiping Yin, and Guizhong Liu. 2022. Improving embedded knowledge graph multi-hop question answering by introducing relational chain reasoning. *Data Mining and Knowledge Discovery*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

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*, 36(7):3580–3599.

Aleksandr Perevalov, Dennis Diefenbach, Ricardo Usbeck, and Andreas Both. 2022. Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers. *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 229–234.

Yunqi Qiu, Yuanzhuo Wang, Xiaolong Jin, and Kun Zhang. 2020. Stepwise reasoning for multi-relation question answering over knowledge graph with weak supervision. In *Proceedings of the 13th International*

Conference on Web Search and Data Mining, WSDM '20, page 474–482, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Yiming Tan, Yongrui Chen, Guilin Qi, Weizhuo Li, and Meng Wang. 2023. Mlpq: A dataset for path question answering over multilingual knowledge graphs. *Big Data Res.*, 32:100381.

Dhaval Taunk, Lakshya Khanna, Siri Venkata Pavan Kumar Kandru, Vasudeva Varma, Charu Sharma, and Makarand Tapaswi. 2023. Grapeqa: Graph augmentation and pruning to enhance question-answering. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 1138–1144, New York, NY, USA. Association for Computing Machinery.

Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. 2022. Human parity on commonsenseqa: Augmenting self-attention with external attention. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2762–2768. International Joint Conferences on Artificial Intelligence Organization.

Min-Chul Yang, Nan Duan, Ming Zhou, and Hae-Chang Rim. 2014. Joint relational embeddings for knowledge-based question answering. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 645–650, Doha, Qatar. Association for Computational Linguistics.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. In *Neural Information Processing Systems (NeurIPS)*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546. Association for Computational Linguistics.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2021. Greaselm: Graph reasoning enhanced language models. In *International Conference on Learning Representations*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.