Causal Reasoning in Large Language Models: A Knowledge Graph Approach

Yejin Kim
GraphLab
George Washington University
Washington, DC, USA
yejinjenny@gwu.edu

Eojin Kang
AI Convergence
Hankuk University of Foreign Studies
Seoul, Korea
eojinkang@hufs.ac.kr

Juae Kim*
AI Convergence
Hankuk University of Foreign Studies
Seoul, Korea
juaekim@hufs.ac.kr

H. Howie Huang*
GraphLab
George Washington University
Washington, DC, USA
howie@gwu.edu

Abstract

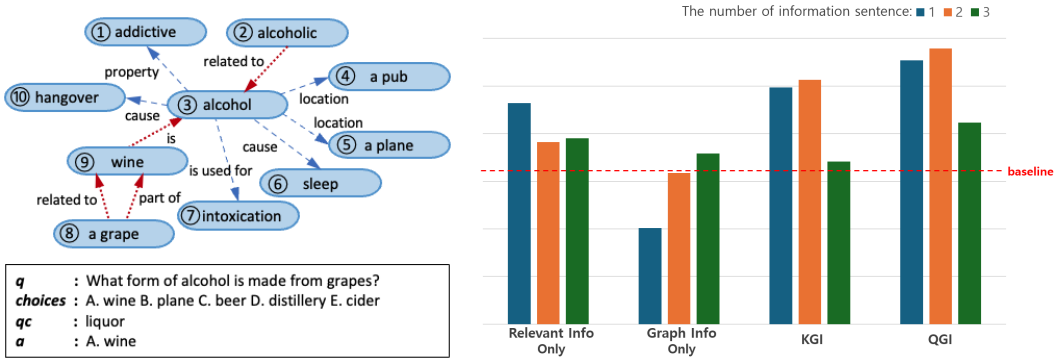
Large language models (LLMs) typically improve performance by either retrieving semantically similar information, or enhancing reasoning abilities through structured prompts like chain-of-thought. While both strategies are considered crucial, it remains unclear which has a greater impact on model performance or whether a combination of both is necessary. This paper answers this question by proposing a knowledge graph (KG)-based random-walk reasoning approach that leverages causal relationships. We conduct experiments on the commonsense question answering task that is based on a KG. The KG inherently provides both relevant information, such as related entity keywords, and a reasoning structure through the connections between nodes. Experimental results show that the proposed KG-based random-walk reasoning method improves the reasoning ability and performance of LLMs. Interestingly, incorporating three seemingly irrelevant sentences into the query using KG-based random-walk reasoning enhances LLM performance, contrary to conventional wisdom. These findings suggest that integrating causal structures into prompts can significantly improve reasoning capabilities, providing new insights into the role of causality in optimizing LLM performance.

1 Introduction

Large language models (LLMs) have demonstrated significant advancements in natural language processing tasks through two primary approaches: providing auxiliary information via retrieval and enhancing reasoning abilities within prompts. Retrieval-augmented generation (RAG) (Lewis et al., 2020) is designed to provide information relevant to the given context or query. This relevant information is identified using embedding similarity searches and then integrated into the LLM's prompt, thereby enhancing the accuracy, relevance recency of the generated response. Recent studies have shown that RAG can significantly boost performance in tasks such as summarization and question answering (QA) (Gu et al., 2019; Shuster et al., 2021; Komeili et al., 2022). However, this approach primarily focuses on retrieving directly related information, raising the question of whether relying solely on such relevant information is sufficient for achieving optimal LLM performance.

*These authors are corresponding authors.

NeurIPS 2024 Workshop on Causality and Large Models (CaLM).



(a) KG example of QA

(b) KG-Based Reasoning

Figure 1: (a) Illustration of a KG structure and an example of CommonsenseQA (Talmor et al., 2018). At the bottom, the question and its concept are represented as q and qc respectively, while the answer is denoted as a. (b) Performance comparison on commonsense QA. The dotted line (baseline) represents the performance when no additional context is provided through the prompt. The three bars represent the number of sentences provided as context in the prompt.

Another approach to improving the quality of LLM responses is to enhance their reasoning abilities (Jain et al., 2023; Suzgun et al., 2023; Wei et al., 2022; Kojima et al., 2022). Reasoning capabilities enable LLMs to go beyond surface-level information and make logical connections between concepts, thereby enhancing their ability to handle more complex queries (Jain et al., 2023; Suzgun et al., 2023; Wei et al., 2022; Kojima et al., 2022). Structured reasoning techniques, such as knowledge graph (KG)-based methods, offer an effective alternative by allowing LLMs to utilize structured knowledge and causal relationships for deeper reasoning (Ling et al., 2023; Yao et al., 2023; Speer and Havasi, 2013). This capability makes KGs a promising resource for comprehending concepts, applying logical reasoning, and refining or validating the model's understanding using existing knowledge (Wang et al., 2023; Yao et al., 2023).

In this paper, we aim to investigate the relative contributions of semantic information retrieval and *causal reasoning*[2] to the LLMs. To this end, we propose a KG-based random-walk reasoning approach that navigates paths of interconnected nodes and edges to uncover causal relationships and extract contextual information, thereby enhancing the reasoning capabilities of LLMs. We use the CommonsenseQA (Talmor et al., 2018) dataset, constructed from the KG, ConceptNet (Speer et al., 2017), as our evaluation benchmark. Figure 1-(a) shows a KG and an example of CommonsenseQA. This dataset is suitable for our study as it requires both relevant information retrieval and complex reasoning to solve problems based on the structured relationships within the KG. We systematically assess the impact of providing semantically relevant information and *causal reasoning* by conducting experiments with various settings that control the presence and type of contextual information provided to the LLM.

Specifically, we evaluate the effect of incorporating causal relationships extracted through random-walk reasoning within the ConceptNet graph, measuring how these elements contribute to LLM performance in answering commonsense questions. To comprehensively analyze the contributions of information retrieval and reasoning, we designed the following experimental settings:

1. **Relevant Information Only**: In this setting, the context provided to the LLM consists of information with high embedding similarity to the question, inspired by RAG, without any additional reasoning process.
2. **Graph Inference Only**: The context provided offers reasoning ability through a continuous process of graph inference, but the information is entirely unrelated to the question.
3. **Keyword Relevant Information + Graph Inference (KGI)**: The information means sentences related to the keywords of the question, also providing reasoning ability through

[2]Causal reasoning here refers to leveraging causal structures in knowledge graphs to enhance reasoning, not formal causal inference.

the proposed approach. However, there is no guarantee that the information in the graph is directly related to the content of the question. In other words, reasoning and information are provided together, but it is possible to include irrelevant information to the question.

4. **Query Relevant Information + Graph Inference (QGI)**: This setting explores how providing both relevant information related to the question and causal relationship through the KG-based random-walk reasoning can enhance the LLMs.

Figure 1-(b) provides a summary of the experiments. Compared to the baseline, where no additional context is given, the prompt involving either relevant information or *causal reasoning* improves performance. Notably, in the "Graph Inference Only" setting, when causal relationships were conveyed through the reasoning of two or three sentences, performance improved even though the content was entirely unrelated to the question. This result indicates that our experimental settings successfully convey reasoning capabilities through causal structures. Furthermore, this finding suggests that, contrary to conventional wisdom, it may be more advantageous to equip LLMs with reasoning abilities, grounded in causality, rather than merely providing semantically related information. A comparison between the "Graph Inference Only" and "KGI" settings shows that information extracted through the KG-based random-walk reasoning method yields better performance than the embedding similarity search-based method in the "Graph Inference Only" setting. This demonstrates that the proposed method, which leverages causal relationships through KG, is more effective for commonsense QA. Lastly, the highest improvement is observed in the "QGI" setting, which combines both relevant information and *causal reasoning*, aligning with our expectations.

Our contributions are summarized as follows:

- We demonstrate the contributions of relevant information and reasoning abilities through experimental comparisons.
- We propose a novel KG-based random-walk reasoning method to utilize causal relationship.
- We show that providing reasoning capabilities grounded in causal relationships can lead to performance improvements, even when using seemingly unrelated information.

2 Method

Our main goal is to investigate KG-based reasoning for the commonsense QA task without further training on LLMs. We first briefly introduce problem formulation. Subsequently, we examine the disparities in prompting procedures between the conventional retrieve-based method and our KG-based reasoning approach.

2.1 Problem Formulation

In a multiple-choice format such as Figure 1-(a), the goal is to predict an answer $a \in Aq$ given a pair of a question and a question concept (q, qc), where $q \in Q$. Any keyword capable of categorizing a question could potentially serve as qc. In our research, we use the term question concept, qc, to encompass all such keywords. The set of choices denoted as Aq, varies with each question, and both questions and answers are represented as variable-length text sequences. KG denoted as $G = (V, E)$, is configured as a heterogeneous graph in general. Within this graph, V is the set of entity nodes, and $E \subseteq V \times R \times V$ denotes the set of edges connecting nodes in V, where R constitutes a set of relation types. The node $v_{qc} \in V$ denotes the node in the KG that is most semantically similar to qc which means that v_{qc} captures the essence of qc within the graph structure. Using v_{qc} as a starting point, we can explore n-hop neighbors which can gather additional information related to the question, q. This process is based on the flow of the graph and allows for effective graph reasoning; it adheres to the directionality of edges within the graph structure. Our graph reasoning process aligns with the inherent structure and semantics of the KG.

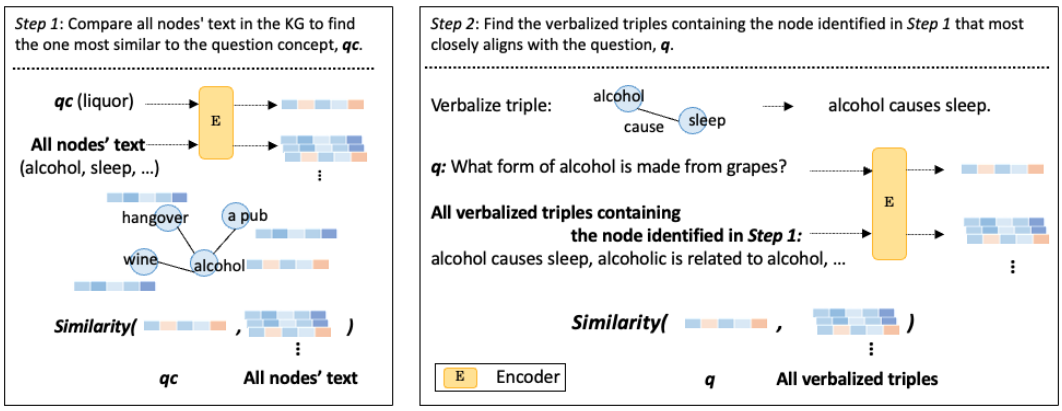


Figure 2: Detailed process of prompting through the KG-based reasoning.

2.2 Retrieval-Based Prompting

A conventional retrieval-based prompting such as RAG follows:

$$\prod_{i}^{N} \sum_{d \in top\text{-}k(p(\cdot|x))} P_\eta(d|x) P_\theta(y_i|x, d, y_{1:i-1}) \tag{1}$$

where x denotes an input query sequence used to retrieve text documents d, incorporating them as additional context during the generation of the target sequence y. In this context, x is a question, q while y corresponds to an answer denoted by a in our commonsense QA task. The retrieval process $P_\eta(d|x) \propto exp(Enc(d)^T, Enc(x))$, where Enc functions as an encoder representing both documents and queries. Specifically, $Enc(d)$ stands for the embedding of a document, while $Enc(x)$ signifies a query embedding generated by the encoder. Determining the top-k documents, denoted as $top\text{-}k(p(\cdot|x))$, where the list of k documents d with the highest prior probability $P_\eta(d|x)$ calculated by a similarity function such as cosine similarity. A generator $P_\theta(y_i|x, d, y_{1:i-1})$ generates the current token by considering the context of the preceding $i-1$ tokens $y_{1:i-1}$, the initial input x, and a retrieved document d.

2.3 KG-Based Random-Walk Reasoning

Given a KG with a pair of question, q and question concept, qc as shown in Figure 1-(a), we seek the node most closely associated with the question concept. Following this initial step of narrowing down the information based on the question concept, we then leverage graph reasoning to extract additional information necessary for formulating a comprehensive answer to the question.

In Figure 2 Step 1, we search for the most similar node aligned with the question concept (qc), the term "*liquor*". Employing a pre-trained text encoder denoted as E, we encode both the anchor text and the entirety of text associated with nodes within the KG. This process yields embeddings for both the anchor text and all node texts. Subsequently, we calculate the cosine similarity between the embedding of the anchor text and the embeddings of all node texts, establishing similarity scores for each pair. In this case, v_{qc} is the node ③ in Figure 1-(a), identified as *alcohol*. Subsequently, we either traverse one hop outbound from the node *alcohol* or find a node that is one hop inbound to *alcohol*. In our case, there are eight one-hop nodes connected to *alcohol*, numbered ①, ②, ④, ⑤, ⑥, ⑦, ⑨ and ⑩. Among these, nodes ② and ⑨ are inbound, while the others are outbound neighbors. In Figure 2 Step 2, we verbalize triples—structured statements consisting of a subject, predicate, and object—based on their direction relative to node ③. For example, in the triple *"alcohol, causes, sleep"* node ③, *alcohol* is the subject, the relationship, *causes* is the predicate, and node ⑥, *sleep* is the object. All connections from or to node ③ are similarly transformed into triples, where the relationships between nodes are clearly expressed in natural language. Using the same encoder E, we encode both the verbalized triples and the question q. We then identify the verbalized triple whose representation is most similar to that of q. Through Step 2, we obtain a single verbalized triple structured as a sentence. As the next step, rather than using a similarity function to select the next object node from the current object node ⑥ (*sleep*), we randomly select n. This approach is intended

to provide reasoning context following the presentation of the most relevant information just once. We combine these sentences into a question, q, and prompt the model to generate an answer based on equation 1.

3 Experiments

3.1 Experimental Setup

We validate the efficiency of our proposed method in zero-shot setting on Llama 2-Chat (Touvron et al., 2023), without further training or tuning on the model, to focus on the effect of the KG-based reasoning process. Thus, we assess results across diverse prompt settings, considering different approaches to retrieving and reasoning information.

Dataset and Encoder We use the CommonsenseQA dataset (Talmor et al., 2018) for evaluation emphasizing the need for diverse commonsense knowledge to choose the correct answers. We employed 1,221 data from the validation dataset due to the unavailability of publicly disclosed answers. To ensure a fair and comprehensive comparison, we choose to employ the ConceptNet KG as the search and reasoning source for both RAG and our proposed method in this experiment. For retrieval purposes, we verbalize the triplets in sentence structures, resulting in 3,423,004 sentences using descriptions of ConceptNet[3]. Additionally, we explore a significant web dataset, using Wikipedia as the search source for the "Relevant Information Only" experiment in this study. The Wikipedia embeddings index provided by txtai[4] is employed. The *e5-base* model (Wang et al., 2022) is utilized to represent all text documents such as Wikipedia and verbalized ConceptNet triples, as well as questions and question concepts.

Evaluation Metrics We evaluate performance using accuracy as the criterion. In experiments where the output format is incorrect in the prompt, multiple scenarios arise. Consequently, all experiments are conducted without an explicit output format. Regardless, when analyzing the results, if the correct answer is "B. exercise," valid responses may take the form of "B", "B.", "B,", "exercise" or "X. exercise". Incorrect responses can include options like "A. exercise", where the alphabet is incorrect but the answer is accurate, "B. exercise, C. muscle", when multiple selections are made, or instances where irrelevant statements are presented.

3.2 Results

Table 1 presents the result of RAG and KG-based random-walk reasoning methods. Our baseline is set as a plain question without additional information. For RAG, we retrieve the top-k, where $k = \{1, 2, 3\}$ sentences from verbalized triples within the ConceptNet KG. In the proposed method, we extract sequential triples connecting to an anchor node determined by its similarity score with the question concept. In contrast, "Relevant Information Only" extracts triples exclusively based on text similarity scores with the combined question concept and the question itself. The findings underscore the effectiveness of KG-based reasoning, leveraging information from connected nodes, particularly in scenarios where the number of information sentences remains constant. In our random-walk approach, we prioritized nodes that were physically close to the starting point. For instance, in the "KGI" case when k = 3 (Table 1), nodes 1 through 5 were selected based on their distance of 1 from the start. This criterion was consistently applied across all experiments. Upon careful performance analysis, it becomes apparent that pertinent information situated at the outset or conclusion of the anchor node proves more advantageous than information located in the middle (Graph, k = 2 in Table 1). Our best performance is attained by incorporating comprehensive context, achieved by combining top-1 information of RAG with data derived from KG-based random-walk reasoning.

In Table 2, we investigate the setting in which the given information is less relevant to the question and its concept. It is crucial observation that even when opting for a less related anchor node and executing a random-walk to obtain k sentences, there is an observed enhancement in performance (Graph, k = 2 and 3 in Table 2). This suggests that reasoning abilities, such as connection of node

[3] https://github.com/commonsense/conceptnet5/wiki/Relations

[4] https://huggingface.co/NeuML/txtai-wikipedia#wikipedia-txtai-embeddings-index

Table 1: Performance comparison of RAG and KG-based reasoning. For a clear explanation of indicating node location, we assume node 1 is the most similar to the question concept and form the graph sequence as 5 -> 4 -> 1 -> 2 -> 3 (k: the number of sentences combined with a question to generate an answer). The highest performance is denoted in bold and the second best results are underlined.

Type	k	**Node Location**	**Acc.**
Baseline	0	-	0.5684
Relevant Information Only	1	top-1 triple	0.5864
	2	top-2 triples	0.5782
	3	top-3 triples	0.5790
Keyword Relevant Information + Graph Inference (KGI)	1	1 -> 2	0.5897
		4 -> 1	0.5766
	2	(1 -> 2, 2 -> 3)	<u>0.5913</u>
		(5 -> 4, 4 -> 1)	0.5577
		(4 -> 1, 1 -> 2)	<u>0.5913</u>
	3	(5 -> 4, 4 -> 1, 1 -> 2)	0.5741
		(4 -> 1, 1 -> 2, 2 -> 3)	0.5741
Query Relevant Information + Graph Inference (QGI)	1 + 2	top-1 + (4 -> 1, 1 -> 2)	**0.5979**

Table 2: Performance in situations where the provided information has lower relevance to the question. (R: relevance of information; if "Y," we remain node 1 as the most similar node and randomly select triples from node 1; otherwise, we opt for an unrelated node randomly). The highest performance is denoted in bold and the second best results are underlined.

Type	k	**R**	**Node Location**	**Acc.**
Baseline	0	-	-	0.5684
Irrelevant Information Only	1	N	1 irrelevant triple	0.5356
	2	N	2 irrelevant triples	0.5324
	3	N	3 irrelevant triples	0.5397
Graph Inference Only	1	N	1 -> 2	0.5602
		N	4 -> 1	0.5602
		Y	1 -> 2	0.5479
		Y	4 -> 1	0.5659
	2	N	(1 -> 2, 2 -> 3)	<u>0.5717</u>
		N	(5 -> 4, 4 -> 1)	0.5635
		N	(4 -> 1, 1 -> 2)	0.5561
	3	N	(5 -> 4, 4 -> 1, 1 -> 2)	0.5667
		N	(4 -> 1, 1 -> 2, 2 -> 3)	**0.5758**

relationships, contribute to problem-solving. Conversely, inputting less relevant information without a coherent flow or reasoning in "Irrelevant Information Only" proves to be ineffective in performance.

We explore diverse prompt configurations and retrieval source datasets, detailed in Table 3, to assess the impact of incorporating extra information into the question. Our findings reveal that using a substantial web dataset, Wikipedia, as a source dataset does not enhance task performance. It is noteworthy that both Relevant Information and Graph scenarios experience decreased performance when incorporating more than three information sentences. The order of prompting is crucial, showing superior performance when retrieved documents precede the question rather than following it. Additionally, the direction of reasoning in the graph is essential, as evidenced by reduced performance when ignoring edge direction.

4 Conclusion

Our experiments were designed to investigate the impact of information retrieval exemplified by Relevant Information, and the reasoning process represented by KG-based random-walk, on improving commonsense QA. Consequently, delivering outcomes inferred through the proposed method generally led to better results than supplying relevant information in Relevant Information. Additionally, the experimental results matched our expectation that performance would be most improved when utilizing both information extracted by embedding matching and graph reasoning. However, we also obtained unexpected results where performance improved when providing irrelevant information

Table 3: Evaluating performance variations across various prompt configurations (Prompt Engineering: order of prompt, direction of reasoning information).

Type	k	Prompt Engineering	Acc.
Relevant Information Only with the ConceptNet Graph	1	documents -> question	0.5864
		question -> documents	0.5455
	4	documents -> question	0.5635
Relevant Information Only with Wikipedia	1	documents -> question	0.5455
	2	documents -> question	0.5504
	3	documents -> question	0.5463
Graph	2	irregular direction (1 -> 2, 4 -> 1)	0.5807
		regular direction (4 -> 1, 1 -> 2)	0.5913
	4	(5 -> 4, 4 -> 1, 1 -> 2, 2 -> 3)	0.5717

through graph reasoning. We analyze that these results indicate providing a reasoning process can enhance the performance of commonsense QA.

The main limitation of this paper is that it is restricted to the commonsense QA task. However, since this task requires both reasoning ability and information with specific information, our experiments have empirically proven which method is more effective. This suggests that focusing on enhancing reasoning capabilities could be beneficial for improving commonsense QA performance in the future.

Acknowledgements

This work was supported in part by National Science Foundation grant 212720.

References

Yunfan Gu, Yuqiao Yang, and Zhongyu Wei. 2019. Extract, Transform and Filling: A Pipeline Model for Question Paraphrasing based on Template. In *Proceedings of the 5th Workshop on Noisy User-generated Text, W-NUT@EMNLP 2019, Hong Kong, China, November 4, 2019*, Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi (Eds.). Association for Computational Linguistics, 109–114. https://doi.org/10.18653/V1/D19-5514

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. 2023. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786* (2023).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-Augmented Dialogue Generation. 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*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 8460–8478. https://doi.org/10.18653/V1/2022.ACL-LONG.579

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

Chen Ling, Xuchao Zhang, Xujiang Zhao, Yifeng Wu, Yanchi Liu, Wei Cheng, Haifeng Chen, and Liang Zhao. 2023. Knowledge-enhanced Prompt for Open-domain Commonsense Reasoning. (2023).

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 3784–3803. https://doi.org/10.18653/V1/2021.FINDINGS-EMNLP.320

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

Robyn Speer and Catherine Havasi. 2013. ConceptNet 5: A Large Semantic Network for Relational Knowledge. In *The People's Web Meets NLP, Collaboratively Constructed Language Resources*, Iryna Gurevych and Jungi Kim (Eds.). Springer, 161–176. https://doi.org/10.1007/978-3-642-35085-6_6

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 13003–13051. https://doi.org/10.18653/V1/2023.FINDINGS-ACL.824

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937* (2018).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533* (2022).

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2023. Knowledge graph prompting for multi-document question answering. *arXiv preprint arXiv:2308.11730* (2023).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

Yao Yao, Zuchao Li, and Hai Zhao. 2023. Beyond Chain-of-Thought, Effective Graph-of-Thought Reasoning in Large Language Models. *arXiv preprint arXiv:2305.16582* (2023).