The Art of Knowledge Graphs in Large Language Models: A Short Review

Maojun SUN
Hong Kong Polytechnic University
mj.sun@connect.polyu.hk

Abstract

Knowledge Graphs (KGs) play a crucial role in enhancing large language models (LLMs) by facilitating knowledge injection, context enrichment, reasoning, and planning. This review article delves into recent mainstream methods for integrating KGs into LLMs at both the training and inference stages. These approaches have shown remarkable improvements in various tasks, highlighting the potential of KGs in augmenting LLM capabilities. The paper further emphasizes the significance of improving LLMs by perform reasoning on KGs, while also discussing the limitations of existing methods and identifying unsolved challenges. Finally, it outlines promising future research directions to drive progress in this rapidly evolving field.

1 Introduction

Large language models (LLMs), like GPT (OpenAI, 2023), LLaMA (Touvron et al., 2023), Qwen (Team, 2023) and DeepSeek series(Liu et al., 2024), have shown impressive performance in producing natural language text, responding to queries, and generating content across diverse fields. Meanwhile, knowledge graphs (KGs) structure and connect information in an organized manner, allowing machines to comprehend and deduce relationships among real-world entities. Intetrating KGs into LLMs presents exciting opportunities for enhancing LLMs, improving their capacity to understand domain knowledge, perform accurate reasoning, and reduce hallucination.

A considerable body of recent research has concentrated on incorporating KGs into both the training and inference stages of LLMs. From previous studies, KGs and it's integrating into LLMs can be formulated by:

Knowledge Graphs are structured repositories of factual information, represented as a set of triples: $\mathcal{G} = (e, r, e')|e, e' \in \mathcal{E}, r \in \mathcal{R}$, where $\mathcal{E}$ denotes the set of entities, and $\mathcal{R}$ denotes the set of relations (Luo et al., 2023c).

Relation Paths are ordered sequences of relations, defined as $z = r_1, r_2, \ldots, r_l$, where each $r_i \in \mathcal{R}$ denotes the i-th relation, and l is the path length.

Reasoning Paths refer to instances of relation paths within KGs, represented as $w_z = e_0 \xrightarrow{r_1} e_1 \xrightarrow{r_2} \ldots \xrightarrow{r_l} e_l$, where each $e_i \in \mathcal{E}$ represents an entity, and r_i denotes the i-th relation in the relation path z.

Knowledge Graph Question Answering (KGQA) is a task that leverages KGs for reasoning. Given a natural language question q and a KG $\mathcal{G}$, the objective is to design a function f that predicts answers $a \in \mathcal{A}_q$ based on knowledge from $\mathcal{G}$, i.e., $a = f(q, \mathcal{G})$. Entities $e_q \in \mathcal{T}_q$ mentioned in q and the corresponding answers $a \in \mathcal{A}_q$ are assumed to be labeled and linked to the respective entities in $\mathcal{G}$, i.e., $\mathcal{T}_q, \mathcal{A}_q \subseteq \mathcal{E}$.

The paper review the following mainstream approaches for integrating KGs into LLMs, including pre-training, inferences, and reasoning, as shown in Table 1.

By incorporating the knowledge by KGs, these approaches have demonstrated significant improvements in various tasks, including KGQA, text generation, and reasoning. However, these methods face several challenges, including handling complex knowledge structures, generating factually accurate content, and supporting multi-step reasoning.

2 Integrating Knowledge Graphs in LLMs

2.1 Knowledge Graphs in LLM Pre-training

Recent developments in large language models have leveraged extensive text corpora for unsupervised learning. Despite their strong performance on downstream tasks, these models often struggle with grounding in real-world knowledge. To ad-

Table 1: Some of studies and categories in this review.

Category	Study
Pre-training	SKEP (Tian et al., 2020), ERNIE (Zhang et al., 2019), KEPLER (Wang et al., 2021a), ERNIE 3.0 (Sun et al., 2021)
Inferences	EMAT (Wu et al., 2022), KGLM (Logan et al., 2019), REALM (Guu et al., 2020), Mindmap (Wen et al., 2023)
Reasoning	Reasoning on graphs (RoG) (Luo et al., 2023d), Graph of thoughts (GoT) (Besta et al., 2023), KnowGPT (Zhang et al., 2024), Graph Chain-of-Though (GRAPH-COT) (Jin et al., 2024)

dress this limitation, researchers have explored integrating knowledge graphs into LLMs using three primary strategies: *1) Embedding KGs in training frameworks*, *2) Enriching LLM inputs with KG information*, and *3) KG-guided instruction tuning*.

2.1.1 Embedding KGs in Training Frameworks

This method introduces knowledge-aware training objectives by incorporating entities from KGs. GLM (Shen et al., 2020) enhances training by assigning higher masking probabilities to entities that are reachable within defined hops. E-BERT (Zhang et al., 2020) optimizes both token- and entity-level losses to enhance learning. SKEP (Tian et al., 2020) integrates sentiment knowledge by giving more attention to words with positive or negative meanings.

Other approaches align text with KG structures. ERNIE (Zhang et al., 2019) trains LLMs to predict alignment links between tokens and KG entities. KALM (Rosset et al., 2020) augments input tokens with entity embeddings and employs an entity prediction objective. KEPLER (Wang et al., 2021a) merges KG embeddings with masked token pre-training. Deterministic LLM (Li et al., 2022) employs contrastive learning to reinforce factual knowledge. WKLM (Xiong et al., 2020) improves entity comprehension by substituting entities with type-consistent alternatives.

2.1.2 Enriching LLM Inputs with KG Information

Certain methods directly append KG subgraphs to LLM inputs. ERNIE 3.0 (Sun et al., 2021) appends KG triples to input text and applies masking to relations or text tokens during training. To reduce *Knowledge Noise*(Liu et al., 2020), K-BERT(Liu et al., 2020) restricts KG visibility to designated entities, while Colake (Sun et al., 2020) links tokens aligned with KG entities to neighboring nodes.

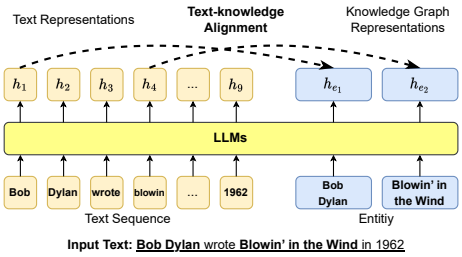

Figure 1: Incorporating KG information into LLM training objectives via text-knowledge alignment loss, where h denotes the hidden representation generated by LLMs (Pan et al., 2024).

For improved rare entity handling, DkLLM (Zhang et al., 2022) substitutes long-tail entities with pseudo-token embeddings. Dict-BERT (Yu et al., 2022) enriches rare word representations by appending dictionary definitions to the input text.

2.1.3 KG-guided Instruction Tuning

Rather than embedding factual knowledge directly, KG-guided instruction tuning enhances LLMs' ability to comprehend KG structures and follow complex instructions. KP-PLM (Wang et al., 2022) reformulates KG structures into natural language prompts for model fine-tuning. OntoPrompt (Ye et al., 2022) employs ontology-enhanced prompts to improve entity understanding. ChatKBQA (Luo et al., 2023a) trains LLMs to generate KG-grounded logical queries. RoG (Luo et al., 2023d)

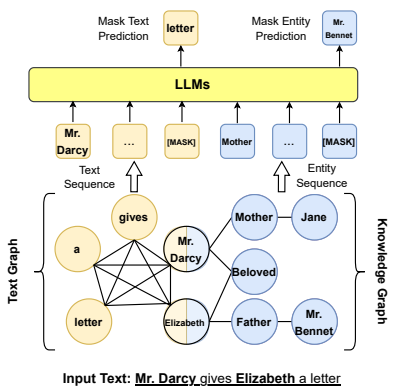

Figure 2: Incorporating KG information into LLM inputs using graph structures (Pan et al., 2024).

proposes a planning-retrieval-reasoning framework to generate interpretable reasoning paths for KG-based tasks.

While KG-guided instruction tuning improves knowledge utilization, it often requires substantial computational resources and model retraining.

2.2 Incorporating KGs in Inference

Integrating knowledge directly into LLMs enhances their performance, but updating this knowledge often requires model retraining, limiting adaptability to evolving information (McCoy et al., 2019). To improve flexibility, recent approaches explore dynamically injecting knowledge during inference, particularly in Question Answering (QA) tasks where both semantic comprehension and factual precision are critical.

2.2.1 Retrieval-Augmented Knowledge Integration

Retrieval-Augmented Knowledge Integration dynamically retrieves relevant information from external sources during inference. RAG (Lewis et al., 2020) employs a hybrid framework combining non-parametric retrieval with a parametric LLM. This structure retrieves KG documents as hidden variables, which are subsequently incorporated into the LLM's input context. This retrieval strategy improves QA performance by producing more accurate, detailed, and factual answers. Extensions such as Story-fragments (Wilmot and Keller, 2021) and EMAT (Wu et al., 2022) refine retrieval filtering and memory efficiency. Meanwhile, REALM (Guu et al., 2020) leverages knowledge retrieval during pre-training, and KGLM (Logan et al., 2019) selects KG facts contextually to generate informative content.

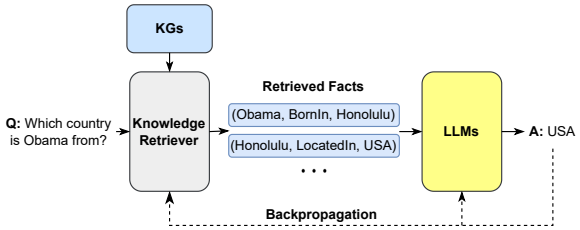

Figure 3: Retrieving external knowledge to enhance LLM generation (Pan et al., 2024).

2.2.2 KG-Driven Prompting

KG-driven prompting reformulates structured KGs into text sequences for LLMs to interpret during inference. Techniques like Triplet Retrieval (Li et al., 2023) convert triples into concise sentences for improved reasoning. Mindmap (Wen et al., 2023) visualizes KGs as structured mind maps, while ChatRule (Luo et al., 2023b) verbalizes relational paths to define logical rules. CoK (Wang et al., 2023) introduces chain-of-knowledge prompting to guide multi-step reasoning tasks.

KG-driven prompting enables flexible KG integration during inference without retraining, though designing effective prompts demands careful engineering.

2.3 Reasoning on Knowledge Graphs

Reasoning on KGs is a critical task that requires LLMs to deduce relationships between entities based on structured knowledge. Recent approaches have explored various reasoning strategies, including graph traversal, path extraction, and logical inference that aim to enhance the reasoning capabilities to address issues such as hallucinations and knowledge limitations in knowledge-intensive tasks. (Bowen et al., 2024).

Graph of Thoughts (GoT) (Besta et al., 2023) framework models the LLM's reasoning process as an arbitrary graph of thoughts, where nodes represent individual thoughts and edges denote dependencies. It introduces graph-based thought transformations such as aggregation, refining, and generation (Besta et al., 2023). GoT advances beyond linear Chain-of-Thought (CoT) and tree-like Tree of Thoughts (ToT) by allowing more complex reasoning flows, including backtracking and merging of ideas from different paths, and iterative refinement of thoughts.

Similarly, Graph Chain-of-Thought (GRAPH-COT) (Jin et al., 2024) augments LLMs by enabling them to reason iteratively on graphs. Each

iteration involves three steps: LLM reasoning to determine necessary external information, LLM-graph interaction to retrieve relevant knowledge using predefined graph functions, and graph execution to fetch the information from the graph (Jin et al., 2024). GRAPH-COT improves upon existing text-based augmentation methods by utilizing the interconnected nature of knowledge within graphs, considering not only individual nodes but also their relationships for reasoning.

Additionally, KnowGPT (Zhang et al., 2024) focuses on grounding LLM responses in factual knowledge from KGs. It employs a knowledge extraction module that uses a reinforcement learning (RL) based strategy (PRL) and a heuristic subgraph extraction strategy (Psub), selected by a Multi-Armed Bandit (MAB) for optimal knowledge retrieval (Zhang et al., 2024). A context-aware prompt construction module then automatically converts the extracted knowledge into effective prompts, also using an MAB to select the best prompt format. KnowGPT improves upon traditional Retrieval-Augmented Generation (RAG) by utilizing the structured knowledge in KGs and employing intelligent strategies for both knowledge retrieval and prompt construction, aiming to reduce hallucinations in domain-specific question answering (Zhang et al., 2024).

Furthermore, Reasoning on Graphs (RoG) (Luo et al., 2023d) synergizes LLMs with KGs for faithful and interpretable reasoning. RoG employs a planning-retrieval-reasoning framework. First, a planning module generates KG-grounded relation paths as faithful plans through planning optimization. These plans are then used by a retrieval module to extract valid reasoning paths from the KG (Luo et al., 2023d). Finally, an LLM reasoning module conducts reasoning based on the retrieved paths to generate answers and interpretable results, further enhanced by retrieval-reasoning optimization (Luo et al., 2023d). RoG's primary improvement lies in its consideration of the KG's structural information for guiding the reasoning process, unlike methods that treat KGs merely as factual repositories. By generating plans based on the KG and then retrieving specific reasoning paths, RoG aims to enhance the faithfulness and interpretability of LLM reasoning (Luo et al., 2023d).

In summary, these approaches demonstrated effectively improvements in LLM reasoning by integrating graph-structured knowledge. They introduce diverse techniques, ranging from iterative graph-based reasoning and flexible thought organization to intelligent knowledge retrieval from KGs and structured planning for reasoning, all with the overarching goal of improving accuracy and reducing hallucinations in LLMs for knowledge-intensive tasks.

3 Challenges and Future Directions

While existing research explores graph-based tasks, there are still some challenges like interconnected knowledge structures, sparse connections, long-tail entities and incomplete graphs in KGs. Besides, efficiently editing knowledge in LLMs without retraining, as current methods suffer from performance issues and computational overhead. Furthermore, leveraging KGs for knowledge editing is a promising direction (Wang et al., 2021b). Additionally, injecting knowledge into black-box LLMs, which only provide API access, remains an open question. Converting knowledge into text prompts is a potential solution, but its effectiveness and generalizability need further exploration.

Further challenges include leveraging multi-modal LLMs for KGs, as real-world KGs often involve diverse data modalities. Developing methods to encode and align entities across different modalities is crucial. Moreover, conventional LLMs struggle to understand structured data like KGs, necessitating the development of LLMs that can directly reason over KG structures. Finally, synergizing LLMs and KGs for bidirectional reasoning can enhance their individual strengths, enabling applications such as search engines, recommender systems, and drug discovery (Wang et al., 2021b). This synergy requires advanced techniques like multi-modal learning, graph neural networks, and continuous learning to unlock the full potential of integrating KGs and LLMs.

4 Conclusion

In conclusion, integrating knowledge graphs into large language models has demonstrated significant potential in enhancing their performance, especially in tasks like question answering and reasoning. Approaches at the training stage, including embedding KGs, enriching LLM inputs with KG information, and KG-guided instruction tuning, have successfully incorporated external knowledge into LLMs. At the inference stage, methods like retrieval-augmented knowledge integration and KG-driven prompting have improved flexibility and

accuracy. Additionally, reasoning techniques such as Graph of Thoughts, GRAPH-COT, KnowGPT, and RoG have enhanced LLM reasoning capabilities, reducing hallucinations. However, challenges persist, such as the interconnected knowledge structures, sparse connections, long-tail entities, and incomplete graphs. Future research cloud focus on improving multi-modalities reasoning with external graphs, synergizing LLMs and KGs for bidirectional reasoning. Addressing these challenges will optimize KG integration in LLMs, making them more intelligent and reliable.

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Jin Bowen, Xie Chulin, Zhang Jiawei, Roy Kashob, Kumar, Zhang Yu, Li Zheng, Li Ruirui, Tang Xianfeng, Wang Suhang, Meng Yu, and Han Jiawei. 2024. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. *arXiv preprint arXiv:2404.07103*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *ICML*.

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Suhang Wang, Yu Meng, and Jiawei Han. 2024. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. *arXiv preprint arXiv:2404.07103*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, volume 33, pages 9459–9474.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022. Pre-training language models with deterministic factual knowledge. In *EMNLP*, pages 11118–11131.

Shiyang Li, Yifan Gao, Haoming Jiang, Qingyu Yin, Zheng Li, Xifeng Yan, Chao Zhang, and Bing Yin. 2023. Graph reasoning for question answering with triplet retrieval. In *ACL*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: enabling language representation with knowledge graph. In *AAAI*, pages 2901–2908.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *ACL*, pages 5962–5971.

Haoran Luo, Zichen Tang, Shiyao Peng, Yikai Guo, Wentai Zhang, Chenghao Ma, Guanting Dong, Meina Song, Wei Lin, et al. 2023a. Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. *arXiv preprint arXiv:2310.08975*.

Linhao Luo, Jiaxin Ju, Bo Xiong, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023b. Chatrule: Mining logical rules with large language models for knowledge graph reasoning. *arXiv preprint arXiv:2309.01538*.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023c. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023d. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arxiv:2310.01061*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*, pages 3428–3448.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Knowledge-aware language model pretraining. *arXiv preprint arXiv:2007.00655*.

Tao Shen, Yi Mao, Pengcheng He, Guodong Long, Adam Trischler, and Weizhu Chen. 2020. Exploiting structured knowledge in text via graph-guided representation learning. In *EMNLP*, pages 8980–8994.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. CoLAKE: Contextualized language and knowledge embedding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3660–3670.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

QWen Team. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *ACL*, pages 4067–4076.

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*.

Jianing Wang, Wenkang Huang, Minghui Qiu, Qiuhui Shi, Hongbin Wang, Xiang Li, and Ming Gao. 2022. Knowledge prompting in pre-trained language model for natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3177.

Jianing Wang, Qiushi Sun, Nuo Chen, Xiang Li, and Ming Gao. 2023. Boosting language models reasoning with chain-of-knowledge prompting. *arXiv preprint arXiv:2306.06427*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021a. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2023. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*.

David Wilmot and Frank Keller. 2021. Memory and knowledge augmented language models for inferring salience in long-form stories. In *EMNLP*, pages 851–865.

Yuxiang Wu, Yu Zhao, Baotian Hu, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2022. An efficient memory-augmented transformer for knowledge-intensive NLP tasks. In *EMNLP*, pages 5184–5196.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *ICLR*.

Hongbin Ye, Ningyu Zhang, Shumin Deng, Xiang Chen, Hui Chen, Feiyu Xiong, Xi Chen, and Huajun Chen. 2022. Ontology-enhanced prompt-tuning for few-shot learning. In *Proceedings of the ACM Web Conference 2022*, pages 778–787.

Wenhao Yu, Chenguang Zhu, Yuwei Fang, Donghan Yu, Shuohang Wang, Yichong Xu, Michael Zeng, and Meng Jiang. 2022. Dict-BERT: Enhancing language model pre-training with dictionary. In *ACL*, pages 1907–1918.

Denghui Zhang, Zixuan Yuan, Yanchi Liu, Fuzhen Zhuang, Haifeng Chen, and Hui Xiong. 2020. E-bert: A phrase and product knowledge enhanced language model for e-commerce. *arXiv preprint arXiv:2009.02835*.

Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. 2024. KnowGPT: Knowledge graph based prompting for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Taolin Zhang, Chengyu Wang, Nan Hu, Minghui Qiu, Chengguang Tang, Xiaofeng He, and Jun Huang. 2022. DKPLM: decomposable knowledge-enhanced pre-trained language model for natural language understanding. In *AAAI*, pages 11703–11711.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *ACL*, pages 1441–1451.