LEGO-GraphRAG: Modularizing Graph-based Retrieval-Augmented Generation for Design Space Exploration [SUPPLIMENTARY MATERIALS]

1 EXPERIMENTAL SETTINGS

All experiments are running on Ubuntu 20.04.6 LTS (Intel(R) Xeon(R) Platinum 8358 CPU@2.60GHz Processor, 4 A100-80G, 400GB memory). The detailed experimental settings are as follows:

1.1 Preprocessing on Graph and GraphRAG Query Datasets.

Freebase is a large-scale, open, structured knowledge base encompassing a vast collection of entities (e.g., individuals, locations, movies, books) and their relationships (e.g., author-book, city-country). The data is represented in a triple format, which forms the foundation for complex and comprehensive graph-based reasoning and question answering tasks for GraphRAG. We utilize four GraphRAG query datasets based on the Freebase within the GraphRAG community, WebQSP [12], CWQ [8], GrailQA [4], and WebQuestions [1]. We process the datasets into json format files through a unified preprocessing step, which includes the question, central entities, answers and hop information corresponding to each query. To address the substantial scale of Freebase, we leverage well-established methodologies [5–7] to construct dataset-specific knowledge graphs (KGs). Our approach involves filtering triples based on domain-specific and relation-level constraints. Specifically, we exclude relations such as *type.object.type* and *type.object.name*, as well as those with prefixes like *common.* and *freebase.*. Additionally, relations containing keywords such as *sameAs* or *sameas* are removed to eliminate redundant or non-informative connections. To further refine the dataset, we discard triples associated with pre-defined domains, including *music.release* and *book.isbn*. This filtering process ensures that only meaningful and relevant triples are retained in the dataset-specific KGs. By significantly reducing noise and focusing on informative triples, the resulting graphs are more efficient for downstream GraphRAG tasks. Detailed preprocessing steps can be found in our code.

1.2 Model Card

This study utilizes a diverse set of models categorized as EEMs and LLMs. Table 1 summarizes the models, providing detailed information about their size and sources.

- **EEMs:** The embedding-focused models include *all-MiniLM-L6-v2* [10], a compact embedding model with 22.7M parameters, and *bge-reranker-v2-m3* [2], a larger reranking model with 568M parameters. These models emphasize efficiency in retrieval and ranking tasks.
- **LLMs:** The large language models demonstrate significant diversity in scale and architecture, with parameters ranging from 7.62B to 72.7B.

Table 1: Information About the Models Used in This Paper

Model Type	Model Name	Model Size
EEMs	all-MiniLM-L6-v2 [10]	22.7M params
	bge-reranker-v2-m3 [2]	568M params
LLMs	Qwen2-7B-Instruct [11]	7.62B params
	glm-4-9b-chat[3]	9.4B params
	Llama-3.3-70B-Instruct [9]	70.6B params
	Qwen2-72B-Instruct [11]	72.7B params

1.3 Hyperparameters.

Table 2 presents the hyperparameters applied across the various GraphRAG instances in this study. The configurations are organized according to the specific combinations of subgraph-extraction (SBE or SAE) and path-retrieval (SBR or I/OSAR) modules used. Key parameters, including the maximum number of retained nodes or entities (max_ent), maximum edges (max_edge), and additional constraints such as filtering metrics (top_k), beam width (beam_width), and maximum path length (max_hop), are outlined for each group. This structured overview highlights the distinct configurations and the parameter tuning strategies employed to ensure consistent and fair comparisons across different instances.

Table 2: The Hyperparameters Used in Different Instances

Instance	Subgraph-Extraction	Path-Filtering
Group(I) SBE & SBR		
No.1 SBE-PPR+SBR-SPR	max_ent=1000	-
Group(II) SAE & I/OSAR		
No.2 SAE-EEMs+OSAR-EEMs	max_edge=64	top_k=32
No.3 SAE-EEMs+OSAR-LLMs	max_edge=64	top_k=32
No.4 SAE-LLMs+OSAR-EEMs	max_edge=64	top_k=32
No.5 SAE-LLMs+OSAR-LLMs	max_edge=64	top_k=32
No.6 SAE-EEMs+ISAR-EEMs	max_edge=64	beam_width=8, max_hop=4
No.7 SAE-EEMs+ISAR-LLMs	max_edge=64	beam_width=8, max_hop=4
No.8 SAE-LLMs+ISAR-EEMs	max_edge=64	beam_width=8, max_hop=4
No.9 SAE-LLMs+ISAR-LLMs	max_edge=64	beam_width=8, max_hop=4
Group(III) SAE & SBR		
No.10 SAE-EEMs+SBR-SPR	max_edge=64	-
No.11 SAE-LLMs+SBR-SPR	max_edge=64	-
Group(IV) SBE & I/OSAR		
No.12 SBE-PPR+OSAR-EEMs	max_ent=1000	top_k=32
No.13 SBE-PPR+OSAR-LLMs	max_ent=1000	top_k=32
Group(V) SBR & I/OSAR		
No.14 SBE-PPR+ISAR-EEMs	max_ent=1000	beam_width=8, max_hop=4
No.15 SBE-PPR+ISAR-LLMs	max_ent=1000	beam_width=8, max_hop=4

2 FINE-TUNING AND MULTIPLE LLM CALLS.

Our implementation integrates fine-tuning for both EEMs (i.e., embedding models and re-rankers) and LLMs, as well as the use of multiple LLM calls in the OSAR-LLM method to select reasoning paths. The test results for these two aspects on sample datasets are presented in Figures 1 and 2, and Table 3.

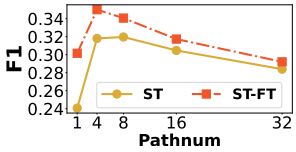

Figure 1: Fine-tuning for EEMs (ST, the Average of Four Datasets)

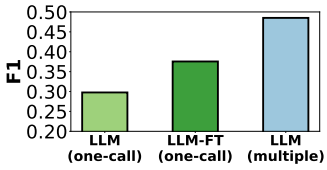

Figure 2: Fine-tuning&multiple Calls for LLMs (Qwen2-72B, WebQSP)

Table 3: Comparison of Fine-tuning Time and Average LLM Calls (WebQSP)

Method	Fine-tuning Time (s)	Ave. LLM Calls
ST-FT	28.7	-
LLM-FT	3748	1
Multiple LLM Calls	-	12.3

3 PROMPTS FOR LLMS

Subgraph-Extraction Prompt Example

You are an expert filterer with a deep understanding of relevant information. Your task is to analyze the given question and entities, and identify the potentially useful reasoning paths from the provided corpus. For each question, you will select the most relevant paths that can help answer the question effectively. Your selection should be based on the entities mentioned in the question and their relationships with other entities in the corpus. Make sure to consider the relevance, specificity, and accuracy of each path in relation to the given question and entities.
Example
<Example>
Question:
<Question>

Path-Retrieval Prompt Example

You are an expert at retrieving the most relevant paths from a given input to answer specific questions.
Example:
<Example>
Reasoning Paths:
<Reasoning Path>
Question:
<Question>

Zero-Shot Reasoning Prompt Example

You are an expert reasoner with a deep understanding of logical connections and relationships. Your task is to analyze the given reasoning paths and provide clear and accurate answers to the questions based on these paths. Based on the reasoning paths, please answer the given question.
Reasoning Path:
<Reasoning Paths>
Question:
<Question>

Here, <Example> refers to a set of human-annotated few-shot examples used to demonstrate the explanation process, <Question> indicates the question and <Reasoning Paths> refers to the retrieved reasoning paths, which are presented as a sequence of structured sentences.

Node Pruning with LLMs Prompt Example

You are an expert in identifying and selecting the most relevant entities from a given corpus to answer specific questions. Your role involves analyzing the provided question and related entities, and then determining the most useful entities that can help address the question effectively. Your selection should be guided by the relevance, specificity, and accuracy of each entity in relation to the entities mentioned in the question. Ensure that the entities you select are concise and directly related to the question.
Example:
<Example>
Entities:
<Entities>
Question:
<Question>

Edge Pruning with LLMs Prompt Example

You are an expert in identifying and selecting the most relevant relational paths from a given corpus to answer specific questions. Your role involves analyzing the provided question and related entities, and then determining the most useful relational paths that can help address the question effectively. Your selection should be guided by the relevance, specificity, and accuracy of each path in relation to the entities mentioned in the question. Ensure that the paths you select are concise and directly related to the question.
Example:
<Example>
Relations:
<Relations>
Question:

<Question>

Triple Pruning with LLMs Prompt Example

You are an expert in identifying and selecting the most relevant triples from a given corpus to answer specific questions. Your role involves analyzing the provided question and related entities, and then determining the most useful triples that can help address the question effectively. Your selection should be guided by the relevance, specificity, and accuracy of each triple in relation to the entities mentioned in the question. Ensure that the triples you select are concise and directly related to the question.
Example:
<Example>
Triples:
<Triples>
Question:
<Question>

Here, <Entities>, <Relations> and <Triples> refers to the retrieved entities/relations/triples from KGs, which are presented as a sequence of structured sentences.

REFERENCES

[1] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (Eds.). Association for Computational Linguistics, Seattle, Washington, USA, 1533–1544. https://aclanthology.org/D13-1160

[2] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv:2402.03216 [cs.CL]

[3] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793

[4] Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases. In Proceedings of the Web Conference 2021 (Ljubljana, Slovenia) (WWW '21). Association for Computing Machinery, New York, NY, USA, 3477–3488. https://doi.org/10.1145/3442381.3449992

[5] Gaole He, Yunshi Lan, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Improving Multi-hop Knowledge Base Question Answering by Learning Intermediate Supervision Signals. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (Virtual Event, Israel) (WSDM '21). Association for Computing Machinery, New York, NY, USA, 553–561. https://doi.org/10.1145/3437963.3441753

[6] Jinhao Jiang, Kun Zhou, Xin Zhao, and Ji-Rong Wen. 2022. UniKGQA: Unified Retrieval and Reasoning for Solving Multi-hop Question Answering Over Knowledge Graph. In The Eleventh International Conference on Learning Representations.

[7] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. In The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=nnVO1PvbTv

[8] 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, Volume 1 (Long Papers), Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 641–651. https://doi.org/10.18653/v1/N18-1059

[9] Llama team. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[10] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In Advances in Neural Information Processing Systems, Vol. 33. Curran Associates, Inc., 5776–5788.

[11] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 Technical Report. arXiv preprint arXiv:2407.10671 (2024).

[12] Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The Value of Semantic Parse Labeling for Knowledge Base Question Answering. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, Berlin, Germany, 201–206. https://doi.org/10.18653/v1/P16-2033