A New Pipeline for Knowledge Graph Reasoning Enhanced by Large Language Models Without Fine-Tuning

Zhongwu Chen[1], **Long Bai**[2], **Zixuan Li**[2], **Zhen Huang**[1], **Xiaolong Jin**[2], **Yong Dou**[1*]
[1]National Key Laboratory of Parallel and Distributed Computing,
National University of Defense Technology,
[2]CAS Key Laboratory of Network Data Science and Technology,
Institute of Computing Technology, Chinese Academy of Sciences
{chenzhongwu20, huangzhen, yongdou}@nudt.edu.cn,
{bailong18b, lizixuan, jinxiaolong}@ict.ac.cn

Abstract

Conventional Knowledge Graph Reasoning (KGR) models learn the embeddings of KG components over the structure of KGs, but their performances are limited when the KGs are severely incomplete. Recent LLM-enhanced KGR models input KG structural information into LLMs. However, they require fine-tuning on open-source LLMs and are not applicable to closed-source LLMs. Therefore, in this paper, to leverage the knowledge in LLMs without fine-tuning to assist and enhance conventional KGR models, we propose a new three-stage pipeline, including knowledge alignment, KG reasoning and entity reranking. Specifically, in the alignment stage, we propose three strategies to align the knowledge in LLMs to the KG schema by explicitly associating unconnected nodes with semantic relations. Based on the enriched KGs, we train structure-aware KGR models to integrate aligned knowledge to original knowledge existing in KGs. In the reranking stage, after obtaining the results of KGR models, we rerank the top-scored entities with LLMs to recall correct answers further. Experiments show our pipeline can enhance the KGR performance in both incomplete and general situations.

1 Introduction

Knowledge Graph (KG) is widely used to store enormous human knowledge or objective facts in the real world. Conventional embedding-based KGR models learn structural embeddings for KG components. Recently, path-based KGR models exploit the logical knowledge underlying the paths connecting the head and tail. All these models treat entities and relations as symbolized identifications without actual semantics and thus heavily rely on reasoning over the KG structures. However, even full-size KG datasets cannot fully cover the massive real-world knowledge and suffer from incompleteness, which naturally restricts KGR performances.

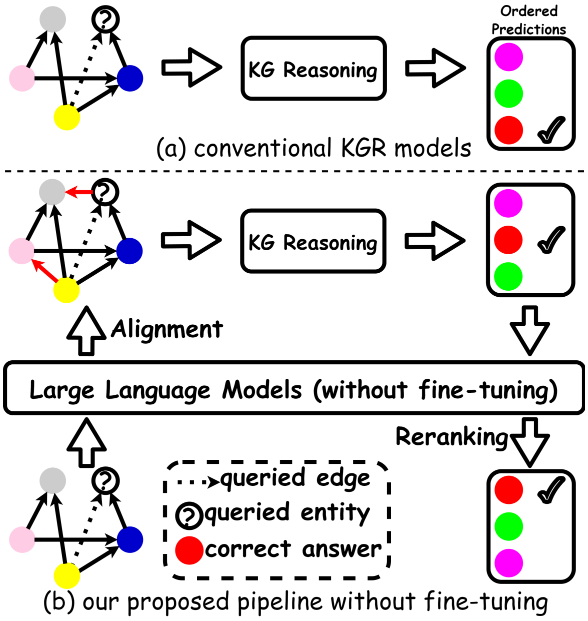

Ordered Predictions
KG Reasoning
(a) conventional KGR models
KG Reasoning
Alignment
Large Language Models (without fine-tuning)
Reranking
queried edge
queried entity
correct answer
(b) our proposed pipeline without fine-tuning

Figure 1: (a) Conventional KGR models reason over original KGs, suffering from incompleteness. (b) Our proposed pipeline without fine-tuning includes three steps: align LLMs to the KG schema (the aligned edges are in red), reason over the enriched KGs and rerank the results with LLMs. Our pipeline achieves better results.

Although LLMs show exciting abilities, it is a challenge for them to singly act as entity reasoners for KGR task due to the huge KG entity space. Tan et al. (2023) further proves that matching the prediction of LLMs with entity names by postprocessing could easily fail. Recently, KGT5 (Saxena et al., 2022) and CSProm-KG (Chen et al., 2023a) have explored to learn KG structure by fine-tuning LLMs. However, on the one hand, for closed-source LLMs like ChatGPT, we can not access the parameters and thus can not combine its knowledge with KGs by fine-tuning; on the other hand, fine-tuning open-source LLMs, such as LLAMA3-70B [1], for a single task is relatively expensive. Therefore, how to assist KGR by incorporating the rich knowledge in LLMs and the structured information in KGs without fine-tuning

[*] Corresponding Author

[1] https://github.com/meta-llama/llama3

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1366–1381
November 12-16, 2024 ©2024 Association for Computational Linguistics

becomes a remaining problem.

Relying on the instruction-following capability of LLMs, we propose to use LLMs from two views to enhance KGR performance without fine-tuning. First, many entity pairs in KGs lack necessary semantic relations because of the incompleteness of KGs. From the view of knowledge alignment, we align the knowledge in LLMs to the KG schema to mitigate the incompleteness of KGs before reasoning and then add the aligned knowledge into KGs in the form of edges, which preserves KG structures and enriches KG connections. Formally, we input a pair of entities into LLM and have it predict their relation. Based on the enriched KGs, we can adopt arbitrary structure-aware KGR models to conduct the entity prediction task. Second, after obtaining KG reasoning results, from the view of entity reranking, we leverage LLMs to rerank the top-scored entities of KGR models for further recalling the correct answers. Finally, these two views of using LLMs to enhance KGR performance are not exclusive and together form our proposed three-stage pipeline for KGR: alignment, reasoning and reranking.

Moreover, in the alignment stage, we present three knowledge alignment strategies, including the closed domain strategy, the open domain strategy and the semi-closed domain strategy. They represent three kinds of approaches for inducing knowledge in LLMs to be outputted according to the KG schema. Specifically, to directly align the knowledge to the manually predefined relations while constructing KGs, the closed domain strategy constrains LLMs to select one of the predefined relations in the form of multiple-choice questions. Since the relations between entities in the real world go beyond the predefined ones, the open domain strategy does not restrict the output content, making less loss of information from LLMs. To provide explainable knowledge alignment for humans, in the semi-closed domain strategy, we map the output of LLMs in the open domain back to the predefined relations by semantic matching.

To verify the effectiveness of our pipeline in incomplete and general situations, we conduct experiments on WN18RR and FB15K-237 with different sparse-level and full-size versions. Additionally, we compare the accuracy and stability of the three alignment strategies to illustrate the quality of the generated relations. We further demonstrate the diverse influences of aligned edges on the original knowledge by analysing the LLMs output in the case study, which reveals that, when applying the open domain knowledge alignment, LLMs generate correct and fine-grained semantics beyond the predefined KG relations. This may explain the mechanism of performance enhancement.

In summary, our contributions are tri-fold:

- To solve the remaining challenges of LLMs in KGR, we propose a three-stage pipeline to assist and enhance conventional KGR models without fine-tuning: alignment, reasoning and reranking.
- In the knowledge alignment stage, we present three alignment strategies in the closed, open and semi-closed domains and we further analyse the accuracy and stability of the three strategies.
- Extensive experiments show the effectiveness of our pipeline and the case study reveals the mechanism of how the knowledge alignment works.

2 Related Work

2.1 Conventional KG Reasoning

Traditional KGR models can be categorized into embedding-based and path-based models (Liang et al., 2022). The embedding-based models encode the KG entities and relations into low-dimension representations. RotatE (Sun et al., 2019) uses a rotation-based method with complex-valued embeddings. Tucker Decomposition is first introduced in KGR by TuckER (Balazevic et al., 2019). Then, HAKE (Zhang et al., 2020) models the semantic hierarchy based on the polar coordinate space and HousE (Li et al., 2022) involves a novel parameterization based on Householder transformations. The backbone of path-based models is reinforcement learning (Das et al., 2018). MultiHopKG (Lin et al., 2018) does multihop reasoning and provides KG paths to support predictions. CURL (Zhang et al., 2022) separates the KGs into different clusters according to the entity semantics and then fine-grains the path-finding procedure into two-level. JOIE (Hao et al., 2019) models all triples in the same zero-curvature Euclidean space, omitting the hierarchical and cyclical structures of KGs. CAKE (Niu et al., 2022) further extracts commonsense entity concepts from factual triples and can augment negative sampling by jointing commonsense and conducting fact-view link prediction.

2.2 Fine-tuning LLMs for KG Reasoning

By modelling KGR task as a sequence-to-sequence problem, GenKGC (Xie et al., 2022) and KG-S2S (Chen et al., 2022) utilize encoder-decoder pre-trained language models to generate target entity names. Lee et al. (2023) unifies KG facts into linearized sentences and guides LLMs to output the answers in texts directly. Following them, fine-tuning open-source LLMs by fusing the accessible KG structures for the KGR task has enjoyed lots of interest. KG-LLaMA (Yao et al., 2023) makes the first step to applying LLaMA (Touvron et al., 2023) in KG link prediction by instruction tuning. KoPA (Zhang et al., 2023c) further leverages prefix tuning and projects KG embeddings into textual token space.

2.3 Exploration of LLMs without Fine-tuning

By prompting LLMs, MPIKGC (Xu et al., 2024) generates descriptions of components in the KGs and sends the enriched information into description-based KGR models. However, MPIKGC is based on description-based KGR models and can not deal with unconnected entities, which we can handle. KICGPT (Wei et al., 2023) reranks the top retrieved entities, but it is centred on prompt engineering and focuses on analysing the effect of several designed knowledge prompts on the ranking quality. Besides, the KGR models KICGPT used are unoptimized. Our proposed pipeline is centred on optimizing KGR models and focuses on assisting reasoning from two perspectives: alignment and reranking.

3 Methodology

In this section, we describe the concrete implementation methodology of the new pipeline without fine-tuning. First, we propose three knowledge alignment strategies and the corresponding ways to convert the textual output of LLMs into KG schema. Second, we train conventional structure-aware KGR models over the enriched KGs. Finally, we further leverage LLMs to rerank the top-scored entities of KGR models, recalling correct answers.

3.1 Knowledge Alignment

To obtain the knowledge related to the queried two entities in LLMs, we induce the output of LLMs via different prompts. Considering the trade-off of the KG schema and the flexible but controllable output of LLMs, we propose the following three alignment strategies, which explicitly enrich KGs with the knowledge in LLMs in three different manners. The prompts are shown in Appendix B. We find whether neighbour edges of entities are included in prompts has little effect on the output of LLMs.

3.1.1 Closed Domain Strategy

The test-like format of multiple-choice questions is generally used in the evaluation of the ability of LLMs in the fields of law (Cui et al., 2023), healthcare (Wang et al., 2023a) and finance (Zhang et al., 2023a). In this alignment strategy, we utilize LLMs to select the most likely relation for the head and tail entities. Specifically, we add the names of predefined KG relations to the prompts as candidates and explicitly instruct LLMs to generate the capital letter before the correct option. LLMs are induced to fully conform to the original KG schema; thus, their knowledge is aligned with KGs at both the semantic and structural levels.

3.1.2 Open Domain Strategy

Actually, the relations between different entities are diverse and fine-grained. However, researchers abstract the KG relations into several representative ones for unification and convenience during the KG construction. We aim to leverage the knowledge in LLMs relevant to the KG domains between two entities to augment the omitted information.

Specifically, in the open domain strategy, we adopt prompts in the form of short answer questions to induce knowledge in LLMs. We do not restrict their output to necessarily follow the predefined KG relations and only imply what aspects of knowledge LLMs should focus on. The description of KG domains in prompts ensures that LLMs do not generate aimlessly. All the outputs are added into KGs as enriched relations on edges, without discarding any semantic information in LLMs.

3.1.3 Semi-Closed Domain Strategy

In the closed domain strategy, LLMs directly generate the option, so we have no insight into how LLMs understand the KG relations and why LLMs make the final decision. As for the open domain strategy, the output of LLMs exactly reflects the knowledge about the two entities. However, LLMs are unable to voluntarily abstract these concrete relations into the structural format as humans do.

Therefore, the semi-closed domain knowledge alignment strategy arises, where we map the output of LLMs in the open domain strategy back to the KG schema. Specifically, we leverage Sentence-BERT (Reimers and Gurevych, 2019) to calculate the semantic similarity between the output and all the predefined relations. The output is eventually converted to the relation with the highest similarity score. This alignment strategy provides an interpretable knowledge alignment for humans between the two forms of knowledge in LLMs and KGs. Through the similarity scores, we can intuitively understand the reasons for the aligned results.

3.2 KG Reasoning

In the closed domain strategy and semi-closed domain strategy, since we align the knowledge in LLMs to the predefined KG relations, we do not need to modify the modelling way of conventional structure-aware KGR models. In the open domain strategy, since the aligned knowledge is added to KGs as sentences, we use word2vec (Mikolov et al., 2013) to initialize the embeddings for words in all the output of LLMs and update them while training. Specifically, we take the mean of all embeddings of words in the corresponding sentence as the embedding of an enriched KG edge. In this way, the downstream KGR models can be trained over the enriched KGs and take advantage of two forms of knowledge in LLMs and KGs at the same time. Based on the predicted entities of KGR models over the enriched KGs, we further improve the performance in the next entity reranking stage.

3.3 Entity Reranking

After the reasoning of the KGR models, we will get a list of entities sorted by the scores calculated by scoring functions. Traditional structure-aware KGR models mainly reason over the KG connections. In this stage, we recall the correct answers using the reranking ability of LLMs based on the predicted entities of KGR models. Specifically, we input the names of top-k candidate entities with the highest scores into prompts (see Appendix C) and utilize the knowledge in LLMs to rerank them based on the probability of semantically holding. Therefore, the entity reranking stage further improves KGR performance by leveraging semantic knowledge along with structural prediction results.

4 Experiments

4.1 Experimental Setup

We adopt the gpt-3.5-turbo version of ChatGPT because of its flexibility and shorter API call time. We also deploy LLAMA3-70B in one 24G Tesla V100 as the representative of open-source LLMs.

For each dataset, the ratio of new facts enriched by LLMs and existing facts in the original dataset is 1:10, i.e., 8684 new facts for the four versions of WN18RR and 27212 new facts for the four versions of FB15K-237. Specifically, for each fact to be added, we make a single LLM call and process the LLM response to the corresponding form in each knowledge alignment strategy.

Dataset	Entity	Relation	Fact	Degree	
				Mean	Median
WN18RR-10%	12,388	11	8,684	1.4	1
WN18RR-40%	20,345	11	34,734	1.7	1
WN18RR-70%	25,831	11	60,785	1.9	1
WN18RR-100%	40,945	11	86,835	2.2	2
FB15K-237-10%	11,512	237	27,212	4.7	3
FB15K-237-40%	13,590	237	108,846	11.2	7
FB15K-237-70%	13,925	237	190,481	14.5	9
FB15K-237-100%	14,505	237	272,115	19.7	14

Table 1: Statistics of our datasets with full-size and different sparse-level versions by randomly retaining.

In the sparse datasets, we randomly select entity pairs which are not connected. Note that, to avoid the information leakage of the KG connections, there is no requirement for these entity pairs to be connected or not in the corresponding full-size KGs. In addition, besides predefined relations, we also allow LLMs to generate or select “no relation” in the corresponding alignment strategy.

For enriched edges, we include all the generated answers into KGs without filtering, even though some of them may conflict with the KG ground truth. The reason is that what we are interested in is the full picture and unprejudiced knowledge of LLMs, so any sort of LLM output evaluation can not be introduced. In other words, regardless of the answers of LLMs being right or wrong, it is a manifestation of its knowledge and should be considered in the downstream KG reasoning.

The maximum token length of input texts is less than 4096. The generated maximum token length is set to 128. For ChatGPT, the temperature parameter is set to 0.3 in the knowledge alignment stage which can increase diversity and set to 0 in the entity reranking stage which can guarantee reliability. In the entity reranking stage, we rerank top-k entities with $k \in \{10, 20\}$. The optimal k is 20 in all datasets. For WN18RR, the optimal alignment strategy is in the open domain. For FB15K-237, the optimal alignment strategy is in the closed domain.

4.2 Datasets

We use WN18RR (Dettmers et al., 2017) and FB15K-237 (Toutanova and Chen, 2015) for our experiments. Datasets with varying degrees of sparsity can simulate several incomplete situations and full-size datasets can simulate the general situation. In experiments, to study the consistency and uni-

		WN18RR-10%		WN18RR-40%		WN18RR-70%		WN18RR-100%	
		MRR	Hits@3	MRR	Hits@3	MRR	Hits@3	MRR	Hits@3
RotatE		0.176	19.3	0.205	23.5	0.220	25.5	0.431	44.2
MultiHopKG		0.164	17.7	0.191	21.4	0.178	20.0	0.433	44.8
ChatGPT$_{zero-shot}$		-	19.8	-	20.6	-	20.2	-	21.1
LLAMA3-70B$_{zero-shot}$		-	22.3	-	20.5	-	21.0	-	20.3
Our Pipeline									
ChatGPT+ RotatE	Alignment, Reasoning	0.241	27.3	0.252	28.7	0.266	30.6	0.476	49.5
	Reasoning, Reranking	0.235	26.7	0.253	31.3	0.258	30.2	0.495	51.6
	Alignment, Reasoning, Reranking	**0.283**	**35.5**	**0.299**	**37.1**	**0.321**	**37.6**	**0.514**	**59.2**
LLAMA3-70B+ RotatE	Alignment, Reasoning	0.235	27.2	0.249	29.4	0.271	31.6	0.507	52.2
	Reasoning, Reranking	0.232	26.2	0.255	31.9	0.266	31.5	0.498	51.6
	Alignment, Reasoning, Reranking	**0.292**	**37.0**	**0.297**	**36.7**	**0.337**	**38.9**	**0.521**	**60.7**
ChatGPT+ MultiHopKG	Alignment, Reasoning	0.218	25.2	0.222	26.1	0.213	24.7	0.465	49.1
	Reasoning, Reranking	0.201	23.1	0.217	24.8	0.231	26.7	0.481	52.5
	Alignment, Reasoning, Reranking	**0.257**	**28.0**	**0.265**	**31.0**	**0.286**	**32.7**	**0.508**	**56.7**
LLAMA3-70B+ MultiHopKG	Alignment, Reasoning	0.207	24.5	0.228	26.3	0.259	28.8	0.481	52.7
	Reasoning, Reranking	0.210	23.9	0.214	23.3	0.219	24.3	0.475	49.3
	Alignment, Reasoning, Reranking	**0.248**	**27.7**	**0.256**	**29.7**	**0.291**	**33.1**	**0.483**	**55.6**

Table 2: Overall results of our pipeline under the optimal settings in WN18RR. The best results are in **bold**.

		FB15K-237-10%		FB15K-237-40%		FB15K-237-70%		FB15K-237-100%	
		MRR	Hits@3	MRR	Hits@3	MRR	Hits@3	MRR	Hits@3
RotatE		0.118	12.4	0.179	18.5	0.189	20.1	0.276	30.6
MultiHopKG		0.110	11.3	0.223	23.9	0.245	26.3	0.294	32.3
ChatGPT$_{zero-shot}$		-	24.3	-	26.5	-	26.0	-	27.3
LLAMA3-70B$_{zero-shot}$		-	26.9	-	23.3	-	27.5	-	29.1
Our Pipeline									
ChatGPT+ RotatE	Alignment, Reasoning	0.157	16.9	0.206	22.2	0.207	22.3	0.294	32.5
	Reasoning, Reranking	0.163	17.1	0.199	21.7	0.204	23.1	0.347	38.0
	Alignment, Reasoning, Reranking	**0.247**	**26.3**	**0.276**	**29.6**	**0.290**	**31.1**	**0.403**	**43.4**
LLAMA3-70B+ RotatE	Alignment, Reasoning	0.169	18.8	0.207	22.0	0.226	23.9	0.361	37.8
	Reasoning, Reranking	0.158	17.6	0.194	22.4	0.216	24.1	0.327	37.9
	Alignment, Reasoning, Reranking	**0.248**	**26.7**	**0.265**	**28.4**	**0.295**	**30.1**	**0.398**	**43.6**
ChatGPT+ MultiHopKG	Alignment, Reasoning	0.184	19.5	0.255	27.4	0.258	28.0	0.343	38.0
	Reasoning, Reranking	0.133	14.3	0.221	25.4	0.233	27.6	0.350	39.8
	Alignment, Reasoning, Reranking	**0.205**	**21.4**	**0.259**	**28.7**	**0.268**	**30.1**	**0.397**	**41.4**
LLAMA3-70B+ MultiHopKG	Alignment, Reasoning	0.173	18.7	0.240	26.5	0.275	29.1	0.355	39.2
	Reasoning, Reranking	0.144	16.9	0.213	25.0	0.226	25.5	0.349	39.1
	Alignment, Reasoning, Reranking	**0.194**	**21.7**	**0.254**	**29.3**	**0.279**	**29.7**	**0.381**	**40.5**

Table 3: Overall results of our pipeline under the optimal settings in FB15K-237. The best results are in **bold**.

versality of the knowledge stored in LLMs for KGs in a variety of incomplete situations, besides full-size dataset WN18RR (WN18RR-100%), we construct three sparse versions, i.e., WN18RR-10%, WN18RR-40% and WN18RR-70%, by randomly retaining 10%, 40% and 70% triples of WN18RR. The same goes for the dataset FB15K-237. The statistics of all the datasets are listed in Table 1.

4.3 Baselines

For LLMs as reasoners, ChatGPT$_{zero-shot}$ and LLAMA3-70B$_{zero-shot}$ mean that, given the queries, we let them directly predict several possible answers according to the possibility. They can not calculate MRR due to the limited text generation space. We leverage two representative SOTA models as conventional KGR models in our pipeline: embedding-based model RotatE and path-based model MultiHopKG. The results based on more KGR models are shown in Appendix A.

4.4 Overall Results

From Table 2 and 3, all the baselines underperform our pipeline. It is difficult for ChatGPT$_{zero-shot}$ and LLAMA3-70B$_{zero-shot}$ to directly generate the correct entity names.

In our experiments, knowledge alignment be-

		WN18RR-10%		WN18RR-40%		WN18RR-70%		WN18RR-100%	
		MRR	Hits@3	MRR	Hits@3	MRR	Hits@3	MRR	Hits@3
RotatE	Upper Performance Bounds	0.283	33.2	0.303	33.8	0.317	35.3	-	-
	Lower Performance Bounds	0.176	19.3	0.205	23.5	0.220	25.5	0.431	44.2
	Closed Domain	0.177	19.1	0.207	24.0	0.221	25.6	0.465	49.2
	Semi-Closed Domain	0.203	22.6	0.215	24.7	0.231	26.6	0.476	49.4
	Open Domain	**0.241**	**27.3**	**0.252**	**28.7**	**0.266**	**30.6**	**0.476**	**49.5**
MultiHopKG	Upper Performance Bounds	0.242	27.8	0.258	29.7	0.265	29.8	-	-
	Lower Performance Bounds	0.164	17.7	0.191	21.4	0.178	20.0	0.433	44.8
	Closed Domain	0.176	19.4	0.193	21.9	0.191	22.1	0.443	46.4
	Semi-Closed Domain	0.205	23.4	0.210	24.1	0.206	24.1	0.451	46.7
	Open Domain	**0.218**	**25.2**	**0.222**	**26.1**	**0.213**	**24.7**	**0.465**	**49.1**

Table 4: KGR performance and our proposed three knowledge alignment strategies under ChatGPT in four versions of WN18RR. Numbers in **bold** are the best results of the three alignment strategies.

		FB15K-237-10%		FB15K-237-40%		FB15K-237-70%		FB15K-237-100%	
		MRR	Hits@3	MRR	Hits@3	MRR	Hits@3	MRR	Hits@3
RotatE	Upper Performance Bounds	0.190	20.2	0.219	23.4	0.226	24.3	-	-
	Lower Performance Bounds	0.118	12.4	0.179	18.5	0.189	20.1	0.276	30.6
	Closed Domain	**0.157**	**16.9**	**0.206**	**22.2**	**0.207**	**22.3**	**0.294**	**32.5**
	Semi-Closed Domain	0.152	16.1	0.203	21.8	0.204	22.0	0.293	32.3
	Open Domain	0.126	13.5	0.194	20.8	0.197	21.1	0.289	31.7
MultiHopKG	Upper Performance Bounds	0.204	21.7	0.272	29.5	0.272	29.4	-	-
	Lower Performance Bounds	0.110	11.3	0.223	23.9	0.245	26.3	0.294	32.3
	Closed Domain	**0.184**	**19.5**	**0.255**	**27.4**	**0.258**	**28.0**	**0.343**	**38.0**
	Semi-Closed Domain	0.177	18.4	0.251	27.2	0.248	26.6	0.323	35.6
	Open Domain	0.142	14.8	0.244	26.1	0.246	26.5	0.315	34.0

Table 5: KGR performance and our proposed three knowledge alignment strategies under ChatGPT in four versions of FB15K-237. Numbers in **bold** are the best results of the three alignment strategies.

fore reasoning (Alignmnet, Reasoning) and entity reranking after reasoning (Reasoning, Reranking) can individually improve reasoning performance. Concatenating these two views, our pipeline (Alignmnet, Reasoning, Reranking) obtains the best performance enhancement. The improvements in full-size datasets indicate that LLMs provide additional information beyond the well-constructed structural knowledge in KGs. In sparse datasets, KGR models suffer from limited training data, whereas our pipeline achieves considerable and consistent enhancement. The gaps in sparse datasets are greater than those in full-size datasets, illustrating our effectiveness under incomplete situations. Furthermore, ChatGPT and LLAMA3-70B show comparable results, confirming the amazing abilities of our pipeline together with recent open-source LLAMA3-70B and closed-source ChatGPT.

4.5 Comparative Study on Knowledge Alignment

In this section, we compare the different impacts of the three knowledge alignment strategies in detail. In Table 4 and 5, the lower bounds are the KGR results without alignment. The upper bounds are the highest results obtained by randomly adding edges with ground truth to KGs and running KGR models multiple times. In full-size datasets, the selected entity pairs do not have golden labels, so we can not acquire the upper performance bounds.

Combining all the results in Table 4 and 5, compared to the lower bounds, there is performance enhancement in all three knowledge alignment strategies for both RotatE and MultiHopKG. This result suggests that explicitly enriching KGs by aligning knowledge in LLMs to KG schema does translate the knowledge into performance enhancement. All the results can not exceed the upper bounds because there is still some deviation between the two forms of knowledge in LLMs and KGs.

For the two kinds of KG datasets, the results of three knowledge alignment strategies show different trends. In Table 4, the improvement in the open domain strategy is the most prominent, followed by the improvement in the semi-closed domain strategy, and the performance improvement in the closed domain strategy is relatively unapparent. By analysing the output content of LLMs and KG schema, we find that there are only eleven high-level relations in WN18RR, and LLMs can generate more detailed descriptions of semantics between words in the open domain. In Table 5, the trend of the three alignment strategies for FB15K-237 is the opposite of the trend for WN18RR. The best performance is achieved with the closed domain. The reason may be that the LLM output contents in the open domain strategy for FB15K-237 have much redundant knowledge about the two entities themselves rather than the expected relations between them. Therefore this information becomes noise that needs to be handled. In contrast, having the LLM output aligned with the KG schema in the closed and semi-closed domain avoids this situation.

4.6 Accuracy of Knowledge Alignment

To intuitively illustrate the effectiveness of the knowledge in LLMs, we calculate the accuracy of the three knowledge alignment strategies from the perspective of relation prediction. Specifically, when there is a golden label of the relation in KGs, we check if LLMs pick up the correct option (automatic evaluation in the closed and semi-closed domain strategies) or if the output and the golden label semantically overlap (manual evaluation in the open domain strategy). When there are no golden labels, we make judgments based on the real world.

From Figure 2, we find all the accuracy rates of ChatGPT directly answering relations between entities are relatively high, which is the source of effectiveness of our proposed knowledge alignment. The accuracy is also stable in the same alignment strategy at different sparsity levels. This indicates knowledge in LLMs is well induced according to the KG schema in our experiments. Moreover, for relatively abstract relations in WN18RR, the highest accuracy is achieved in the open domain strategy, while for relatively concrete relations in FB15K-237, the highest accuracy is achieved in the closed domain strategy. These two phenomena are consistent with the performance enhancement in Section 4.5. The semi-closed domain strategy loses some information in the process of transforming linguistic forms for the sake of interpretability, and thus achieves the median accuracy in all datasets.

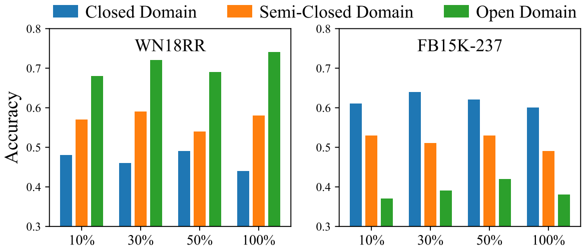


Figure 2: The accuracy that ChatGPT correctly outputs the relations between entities in three alignment strategies for two datasets at different sparsity levels.

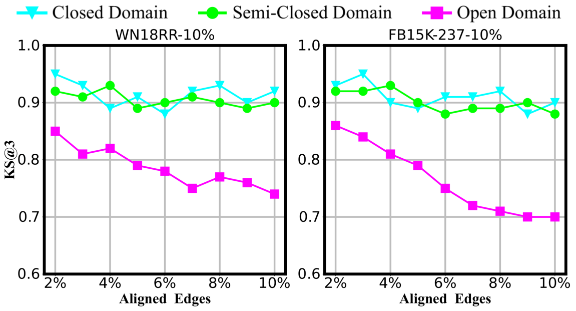


Figure 3: Impacts of the number of aligned edges on the stability of the three knowledge alignment strategies.

4.7 Stability of Knowledge Alignment

The stability of knowledge alignment seeks to evaluate whether enriching KGs by aligning LLMs with KG schema in the three strategies will impact the original knowledge stored in KGs. We introduce the **Knowledge Stability** (**KS@k**) metric, indicating the ratio of entities that are correctly predicted by KGR models both before alignment and after alignment. We calculate **KS@k** as follows:

$$KS@k = \frac{\sum rank\left(Alignment, Reasoning\right) \leq k}{\sum rank\left(Reasoning\right) \leq k},$$

where $\sum rank\left(Reasoning\right) \leq k$ signifies the count of rank value under k predicted by KGR models before alignment, i.e., original KGR results; $\sum rank\left(Alignment, Reasoning\right) \leq k$ denotes the count of rank value under k predicted by KGR models after alignment, i.e., enhanced KGR results.

The insight is that if the score rankings of correct answers in this dataset maintain less than k after alignment, the aligned knowledge in the three alignment strategies is stable. However, for some specific queries, the prediction may be worse due to the introduced wrong facts, resulting in our pipeline changing its prediction from a correct answer to a wrong one and then KS@k declines.

In Figure 3, we employ the number of aligned edges ranging from 2% to 10%, with an interval of 1%, and measure stability by KS@3 for RotatE. We observe that the closed and semi-closed domains, which add predefined relations into KGs, have stable performance for both datasets. However, the open domain strategy sees varying degrees of decline. We attribute this to KGR models paying more focus on the diverse output and then resulting in the dilution of original KG knowledge.

4.8 Case Study of the Aligned Knowledge

To further explore the positive and negative influence of LLM output in the open domain strategy on different datasets, we list some typical output of ChatGPT and LLAMA3-70B in Appendix D and carry out error analysis in Appendix E. We find the LLM output usually goes beyond the predefined KG relations and provides fine-grained information. However, LLMs may also provide "redundant correct information" as shown below.

Positive Influence. LLMs in the open domain usually generate the relationship in plain and accurate language, without using professional linguistic vocabulary. For instance, LLMs output “Tuberculosis is a type of infectious disease”, which is in line with the definition of “hypernym”. We visualize the embeddings of predefined KG relations and keywords generated in the open domain strategy learned by RotatE. Figure 4 shows two cases which explicitly illustrate their positions in the embedding space. Close points in the space indicate that RotatE successfully captures their similar semantics and then these newly generated words are well integrated into the KG schema. The eleven predefined relations can be seen as abstractions of the concrete output of LLMs. Therefore, KGR models indeed understand and benefit from our proposed open-domain knowledge alignment strategy.

Negative Influence. In contrast, although the LLM output is consistent with the objective world, it may contain “redundant correct information”. In FB15K-237, when asked about the relation between “Robert Ridgely” and “USA”, besides correctly answering “Robert Ridgely was an American”, ChatGPT and LLAMA3-70B also output his occupation, which is a redundant entity property. This “redundant correct information” would somewhat interfere with the downstream training. Compared with the open domain strategy, aligning knowledge in LLMs with the KG schema of FB15K-237 in the other two strategies introduces less noise. Therefore, in summary, LLMs consistently improve the KGR performance under all three proposed strategies, while showing different characteristics and influence in various scenarios.

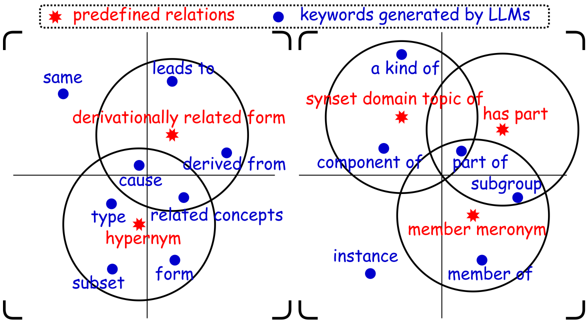


Figure 4: The positions of the predefined relations in WN18RR and keywords generated by ChatGPT in the open domain alignment strategy in the embedding space. We can see the predefined relations have overlapping and more delicate semantics, which LLMs realize.

		ChatGPT				LLAMA3-70B			
		Top-10		Top-20		Top-10		Top-20	
		Hits@3	*Imp.*	Hits@3	*Imp.*	Hits@3	*Imp.*	Hits@3	*Imp.*
WN18RR	10%	33.2	**+5.8**	35.5	**+8.2**	33.0	**+7.8**	37.0	**+9.7**
	40%	35.2	**+5.5**	37.1	**+8.4**	36.6	**+9.3**	36.7	**+7.3**
	70%	35.4	**+4.7**	37.6	**+7.0**	34.5	**+5.9**	38.9	**+7.3**
	100%	56.6	**+3.9**	59.2	**+9.7**	55.4	**+4.2**	60.7	**+8.5**
FB15K-237	10%	23.3	**+4.5**	26.3	**+9.4**	24.2	**+5.4**	26.7	**+7.9**
	40%	27.0	**+4.3**	29.6	**+7.4**	28.0	**+5.9**	28.4	**+6.4**
	70%	28.0	**+4.0**	31.1	**+8.8**	29.9	**+6.0**	30.1	**+6.2**
	100%	42.5	**+3.4**	43.4	**+10.9**	43.1	**+4.3**	43.6	**+5.8**

Table 6: The results of LLMs as reranker for top-10 and top-20 entities. *Imp.* is the improvement of the entity reranking stage after alignment and reasoning.

4.9 Effects of Reranking Entity Numbers

Table 6 shows conspicuous performance enhancement of LLMs as rerankers, which suggests the effectiveness of our proposed pipeline. The sparser the datasets, the more significant the enhancement of the entity reranking stage and the top-20 scenario gives better results than the top-10 scenario because LLMs have more chances to recall correct answers from candidates. These results prove that after the knowledge alignment stage, LLMs can further enhance the KGR performance based on the semantic differences between candidate entities. Moreover, LLAMA3-70B and ChatGPT have competitive overall results (Hits@3) and performance improvement (*Imp.*) in all the datasets, showing the generalizability of our pipeline.

5 Conclusion

This paper introduces a new pipeline for LLMs to assist and enhance KGR models without finetuning. We propose three knowledge alignment strategies to enrich KGs before reasoning and leverage LLMs as rerankers to recall correct answers. Experiments illustrate the effectiveness of our pipeline, both in incomplete and general situations, and the accuracy and stability of the proposed knowledge alignment. The case study reveals the various outputs of LLAMA3-70B and ChatGPT.

Limitations and Future Work

During the use of LLMs, we cannot anticipate whether the output is valuable before the call of LLMs, resulting in the quality of each answer of LLMs can not be controlled. Moreover, the error analysis in Table 9 also shows that there are some imperfections in the output of LLMs. Therefore, in the future, we can add a module to make further corrections using the ability of KGR models while KG reasoning.

Additionally, our proposed pipeline is scalable. The rapidly evolving RAG technology (Gao et al., 2024) may further improve the quality of knowledge alignment and reranking. We also hope the pipeline can inspire more thinking about how to utilize closed-source LLMs to enhance the performance of other KG-related tasks from the perspectives of knowledge alignment and reranking.

Ethics Statement

In this paper, we use datasets WN18RR and FB15K-237, including eight versions of them. The data is all publicly available. Our task is knowledge graph reasoning, which is performed by finding missing entities given existing knowledge. This work is only relevant to NLP research and will not be put to improper use by ordinary people. We acknowledge the importance of the ACM Code of Ethics and totally agree with it. We ensure that this work is compatible with the provided code, in terms of publicly accessed datasets and models.

Risks and harms of LLMs include the generation of harmful, offensive, or biased content. These models are often prone to generating incorrect information, sometimes referred to as hallucinations. The ChatGPT used in this paper was licensed under the terms of OpenAI. We are not recommending the use of our proposed pipeline for alignment or ranking tasks with social implications, such as job candidates or products, because LLMs may exhibit racial bias, geographical bias, gender bias, etc., in the reasoning results. In addition, the use of LLMs in critical decision-making sessions may pose unspecified risks.

References

Ivana Balazevic, Carl Allen, and Timothy Hospedales. 2019. TuckER: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5185–5194, Hong Kong, China. Association for Computational Linguistics.

Chen Chen, Yufei Wang, Bing Li, and Kwok-Yan Lam. 2022. Knowledge is flat: A Seq2Seq generative framework for various knowledge graph completion. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4005–4017, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chen Chen, Yufei Wang, Aixin Sun, Bing Li, and Kwok-Yan Lam. 2023a. Dipping PLMs sauce: Bridging structure and text for effective knowledge graph completion via conditional soft prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11489–11503, Toronto, Canada. Association for Computational Linguistics.

Zhongwu Chen, Chengjin Xu, Fenglong Su, Zhen Huang, and Yong Dou. 2023b. Incorporating structured sentences with time-enhanced bert for fully-inductive temporal relation prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 889–899, New York, NY, USA. Association for Computing Machinery.

Zhongwu Chen, Chengjin Xu, Fenglong Su, Zhen Huang, and Yong Dou. 2023c. Meta-learning based knowledge extrapolation for temporal knowledge graph. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2433–2443, New York, NY, USA. Association for Computing Machinery.

Zhongwu Chen, Chengjin Xu, Fenglong Su, Zhen Huang, and Yong Dou. 2023d. Temporal extrapolation and knowledge transfer for lifelong temporal knowledge graph reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6736–6746, Singapore. Association for Computational Linguistics.

Alla Chepurova, Aydar Bulatov, Yuri Kuratov, and Mikhail Burtsev. 2023. Better together: Enhancing generative knowledge graph completion with language models and neighborhood information. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5306–5316, Singapore. Association for Computational Linguistics.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. 2018. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. In *ICLR*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2017. Convolutional 2d knowledge graph embeddings. In *AAAI Conference on Artificial Intelligence*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2023. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting.

Junheng Hao, Muhao Chen, Wenchao Yu, Yizhou Sun, and Wei Wang. 2019. Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 1709–1719, New York, NY, USA. Association for Computing Machinery.

Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. 2023. Temporal knowledge graph forecasting without knowledge using in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 544–557, Singapore. Association for Computational Linguistics.

Rui Li, Jianan Zhao, Chaozhuo Li, Di He, Yiqi Wang, Yuming Liu, Hao Sun, Senzhang Wang, Weiwei Deng, Yanming Shen, Xing Xie, and Qi Zhang. 2022. HousE: Knowledge graph embedding with householder parameterization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13209–13224. PMLR.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. 2022. Reasoning over different types of knowledge graphs: Static, temporal and multi-modal. *arXiv preprint arXiv:2212.05767*.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. 2018. Multi-hop knowledge graph reasoning with reward shaping. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3243–3253, Brussels, Belgium. Association for Computational Linguistics.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Guanglin Niu, Bo Li, Yongfei Zhang, and Shiliang Pu. 2022. CAKE: A scalable commonsense-aware framework for multi-view knowledge graph completion. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2867–2877, Dublin, Ireland. Association for Computational Linguistics.

Guanglin Niu, Yongfei Zhang, Bo Li, Peng Cui, Si Liu, Jingyang Li, and Xiaowei Zhang. 2019. Rule-guided compositional representation learning on knowledge graphs. In *AAAI Conference on Artificial Intelligence*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 2022. Sequence-to-sequence knowledge graph completion and question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2814–2828, Dublin, Ireland. Association for Computational Linguistics.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.

Yiming Tan, Dehai Min, Y. Li, Wenbo Li, Na Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Workshop on the Semantic Web*.

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66, Beijing, China. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Fanqi Wan, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Mitigating hallucinations of large language models via knowledge consistent alignment.

Haochun Wang, Chi Liu, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. Chatglm-med. In `https://github.com/SCIR-HI/Med-ChatGLM`. GitHub.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023b. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Yanbin Wei, Qiushi Huang, Yu Zhang, and James Kwok. 2023. KICGPT: Large language model with knowledge in context for knowledge graph completion. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8667–8683, Singapore. Association for Computational Linguistics.

Xin Xie, Ningyu Zhang, Zhoubo Li, Shumin Deng, Hui Chen, Feiyu Xiong, Mosha Chen, and Huajun Chen. 2022. From discrimination to generation: Knowledge graph completion with generative transformer. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 162–165, New York, NY, USA. Association for Computing Machinery.

Derong Xu, Ziheng Zhang, Zhenxi Lin, Xian Wu, Zhihong Zhu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024. Multi-perspective improvement of knowledge graph completion with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11956–11968, Torino, Italia. ELRA and ICCL.

Liang Yao, Jiazhen Peng, Chengsheng Mao, and Yuan Luo. 2023. Exploring large language models for knowledge graph completion.

Denghui Zhang, Zixuan Yuan, Hao Liu, Xiaodong lin, and Hui Xiong. 2022. Learning to walk with dual agents for knowledge graph reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5932–5941.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, Zhoufan Zhu, Anbo Wu, Xin Guo, and Yun Chen. 2023a. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models.

Yichi Zhang, Zhuo Chen, Yin Fang, Lei Cheng, Yanxi Lu, Fangming Li, Wen Zhang, and Huajun Chen. 2023b. Knowledgeable preference alignment for llms in domain-specific question answering.

Yichi Zhang, Zhuo Chen, Wen Zhang, and Huajun Chen. 2023c. Making large language models perform better in knowledge graph completion.

Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. 2020. Learning hierarchy-aware knowledge graph embeddings for link prediction. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 3065–3072. AAAI Press.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. 2024. Large language models for information retrieval: A survey.

Appendix

A Results based on more KGR models

In this section, in order to demonstrate the generalizability of the proposed pipeline, we list the results of our pipeline using RPJE (Niu et al., 2019) in Table 7. From Table 7, we find RPJE is a powerful baseline. The combination of RPJE and our proposed pipeline further confirms our contributions.

	FB15K-237-100%
	MRR
RPJE	0.470
RPJE+our pipeline (ChatGPT)	0.519
RPJE+our pipeline (LLAMA3-70B)	0.526

Table 7: Results of our pipeline with RPJE.

B Prompts for knowledge alignment

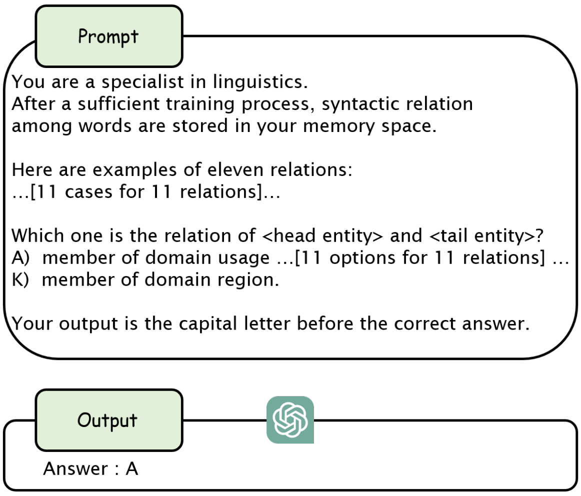
Currently, many works are trying to explore how to incorporate structural information stored in KGs into the knowledge in LLMs (Chepurova et al., 2023). They either explicitly linearize the neighbourhood edges and use LLMs as answer generators, or fine-tune LLMs by incorporating the structured KG embedding into the input of LLMs. As mentioned in the introduction, our motivation is the opposite of the recent papers. We want to figure out whether the knowledge stored in the LLMs itself can be aligned with the predefined schema of KGs. Therefore, for our designed prompts input into LLMs, we should not introduce any structural information, such as neighbourhoods, paths or subgraphs. Following the conclusions of (Min et al., 2022), we design several prompts and select the best in our experiments.

To make LLMs better understand the semantics of relations, we randomly choose some triple examples of relations and expect LLMs to capture their meanings. We also include a description of KG domains, since the relations are highly correlated with it.

Figure 5, Figure 6, Figure 7 and Figure 8 represent four prompts in our proposed knowledge alignment settings for two datasets.

C Prompts for LLMs as reranker

Inspired by LLMs as rerankers in Information Retrieval (IR) (Zhu et al., 2024), we design two prompts for LLMs as rerankers in Figure 9 and Figure 10.



Prompt

You are a specialist in linguistics.
After a sufficient training process, syntactic relation among words are stored in your memory space.

Here are examples of eleven relations:
...[11 cases for 11 relations]...

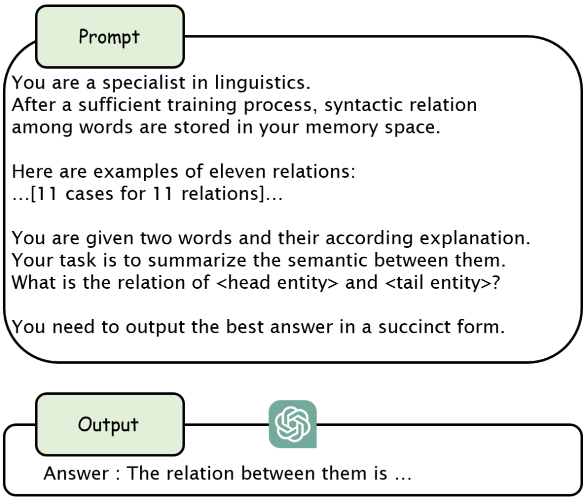
Which one is the relation of <head entity> and <tail entity>?
A) member of domain usage ...[11 options for 11 relations] ...
K) member of domain region.

Your output is the capital letter before the correct answer.

Output

Answer : A

Figure 5: Prompt in the closed-domain knowledge alignment setting for WN18RR.



Prompt

You are a specialist in linguistics.
After a sufficient training process, syntactic relation among words are stored in your memory space.

Here are examples of eleven relations:
...[11 cases for 11 relations]...

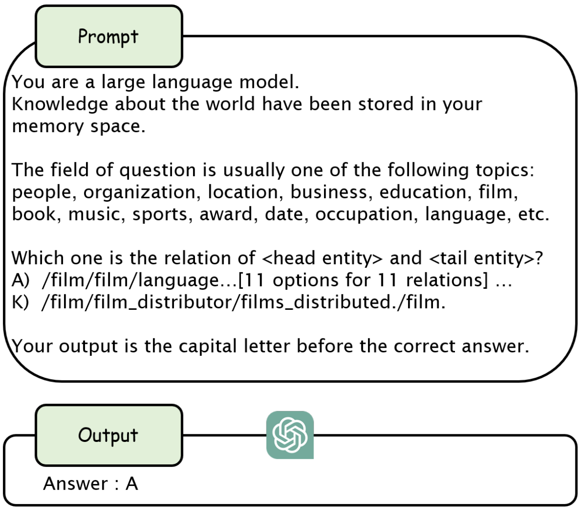
You are given two words and their according explanation.
Your task is to summarize the semantic between them.
What is the relation of <head entity> and <tail entity>?

You need to output the best answer in a succinct form.

Output

Answer : The relation between them is ...

Figure 6: Prompt in the open-domain knowledge alignment setting for WN18RR.



Prompt

You are a large language model.
Knowledge about the world have been stored in your memory space.

The field of question is usually one of the following topics:
people, organization, location, business, education, film, book, music, sports, award, date, occupation, language, etc.

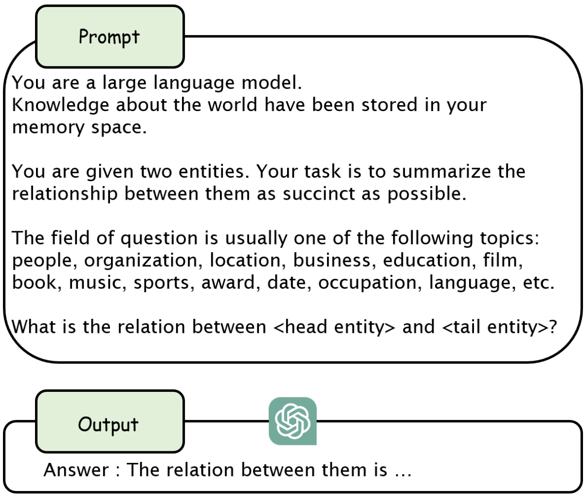
Which one is the relation of <head entity> and <tail entity>?
A) /film/film/language...[11 options for 11 relations] ...
K) /film/film_distributor/films_distributed./film.

Your output is the capital letter before the correct answer.

Output

Answer : A

Figure 7: Prompt in the closed-domain knowledge alignment setting for FB15K-237.



You are a large language model.
Knowledge about the world have been stored in your memory space.

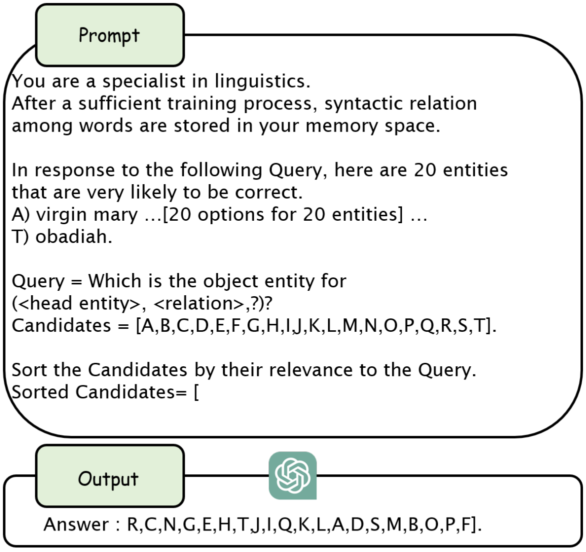
You are given two entities. Your task is to summarize the relationship between them as succinct as possible.

The field of question is usually one of the following topics: people, organization, location, business, education, film, book, music, sports, award, date, occupation, language, etc.

What is the relation between <head entity> and <tail entity>?

Answer : The relation between them is ...

Figure 8: Prompt in the open-domain knowledge alignment setting for FB15K-237.



You are a specialist in linguistics.
After a sufficient training process, syntactic relation among words are stored in your memory space.

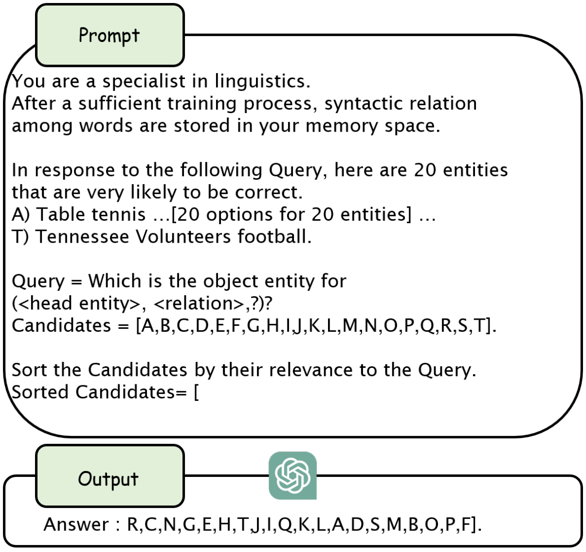
In response to the following Query, here are 20 entities that are very likely to be correct.
A) virgin mary ...[20 options for 20 entities] ...
T) obadiah.

Query = Which is the object entity for (<head entity>, <relation>,?)?
Candidates = [A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T].

Sort the Candidates by their relevance to the Query.
Sorted Candidates= [

Answer : R,C,N,G,E,H,T,J,I,Q,K,L,A,D,S,M,B,O,P,F].

Figure 9: Prompt of LLMs as Reranker for WN18RR.



You are a specialist in linguistics.
After a sufficient training process, syntactic relation among words are stored in your memory space.

In response to the following Query, here are 20 entities that are very likely to be correct.
A) Table tennis ...[20 options for 20 entities] ...
T) Tennessee Volunteers football.

Query = Which is the object entity for (<head entity>, <relation>,?)?
Candidates = [A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T].

Sort the Candidates by their relevance to the Query.
Sorted Candidates= [

Answer : R,C,N,G,E,H,T,J,I,Q,K,L,A,D,S,M,B,O,P,F].

Figure 10: Prompt of LLMs as Reranker for FB15K-237.

D Some case studies of the aligned knowledge

Table 8 shows some cases of the output of ChatGPT and LLAMA3-70B.

E Error analysis

We analyse the incorrect output of LLMs in the open domain. Errors fall into the following three categories: error type 1) generating fabricated or misplaced facts (hallucination of LLMs); error type 2) outputting “not related” for those entity pairs that should have relations; and error type 3) outputting incorrectly formatted or meaningless sentences. We show some cases in the Table 9. These inconsistencies can be solved through further inconsistency detection and knowledge consistent alignment (Wan et al., 2024; Guan et al., 2023; Zhang et al., 2023b).

F Experimental detail

Our experiments use one 24G Tesla V100 GPU with Pytorch 1.8. The KG reasoning process needs 3h to 12h, depending on the sparsity level of the datasets. The implementation code of KGR models is obtained from their original papers. We use the optimal parameter reported in the original papers and code. All the results were mean values from multiple runs.

The keys of ChatGPT API were bought from the official channel. Each call time was about 0.5s to 2s. All the input and output of ChatGPT is in English. The collection of the output of ChatGPT was done by the authors. Since the used datasets are well constructed, there are no offensive content and identifiers. While collecting the output of ChatGPT, we still manually checked to anonymise the offensive content and identifiers in the output by removing them.

G Differences between our work with works of KG construction and works introducing the external knowledge

Currently, KG construction is mainly based on the ability of LLMs to extract the given text. For instance, Univeral IE (Lu et al., 2022) was fine-tuned for different information extraction domains respectively; InstructionUIE (Wang et al., 2023b) further utilized instruction fine-tuning to unify multiple information extraction tasks at the output side and achieved better performance. On the other hand, in order to understand if LLMs contain real-world knowledge, given head entities and relations, LAMA (Petroni et al., 2019) answered queries structured as "fill-in-the-blank" cloze statements to answer tail entities and found that LLMs can recover missing entities to some extent. In contrast, our proposed pipeline leveraged LLMs to answer missing relations and enriched KGs with predicted relations by LLMs, leading to the improvement of KGR models. However, constructing KG with LLMs is not our ultimate goal, our goal is to make the enriched new edges serve KGR models better.

In order to conduct KGC tasks, the researchers explored several ways of enriching KG information, including, KG text descriptions (Chen et al., 2023b), lifelong reasoning (Chen et al., 2023d), and the use of Meta-Learning (Chen et al., 2023c). Since LLMs are currently believed to contain a wealth of real-world knowledge, we want to know whether the knowledge of LLMs is effective for KGR models, and thus this exploratory work proposed two ways of using LLMs to help KGR, and formed a pipeline: firstly, introducing the knowledge of LLMs into KGs, and then using the knowledge of LLMs to rerank the results of KGR. Note that currently only LLMs can provide both two capabilities together for the KGR task.

H The choice of three knowledge alignment strategies

The results of our proposed strategies are related to the abstraction degree of the relations in KGs. We try to give a qualitative description: For KGs with abstract relations, Open Domain Strategy is the best; for KGs with concrete relations, Closed Domain Strategy is the best. Specifically, from Section 4.8, we find 1) the relations in WN18RR are high-dimensional and abstract (for instance: hypernym, derivationally related from). 2) the relations in FB15K-237 are concrete and non-subdivisible (for instance: award, has nationality). Because of the powerful generative and understanding capabilities of LLMs, for WN18RR, the LLM output usually goes beyond the predefined KG relations and provides fine-grained information, leading to the best performance in the Open Domain Strategy. In contrast, FB15K-237 contains detailed relations, so it is the best choice to correspond the knowledge of LLMs to the original relations in FB15K-237.

Datasets	Head Entity	Tail Entity	Golden Label	ChatGPT Outputs
WN18RR	Sesame Family	Genus Martynia	member meronym	Sesame family is a higher-level taxonomic grouping that includes the genus martynia.
	Tuberculosis	Infectious Disease	hypernym	Tuberculosis is a type of infectious disease.
	Prettify	Beauty	derivationally related form	Prettify is a verb that means to make something more attractive or beautiful.
FB15K-237	Star Trek Nemesis	Jerry Goldsmith	/film/film/music	Star Trek Nemesis is a 2002 science fiction film directed by Stuart Baird, with music composed by Jerry Goldsmith.
	White American	Sylvia Plath	/people/ethnicity/people	White American is the ethnicity of Sylvia Plath, an American poet and novelist.
	Robert Ridgely	USA	/people/person/nationality	Robert Ridgely was an American actor. He was active in the American film industry until his death in 1997.

Datasets	Head Entity	Tail Entity	Golden Label	LLAMA3-70B Outputs
WN18RR	Sesame Family	Genus Martynia	member meronym	Genus Martynia is a part of the Sesame Family (Pedaliaceae), meaning Martynia is a genus within the Pedaliaceae family.
	Tuberculosis	Infectious Disease	hypernym	Tuberculosis is a type of infectious disease, meaning that it is a disease caused by the bacterium Mycobacterium tuberculosis and is contagious, spreading from person to person through the air.
	Prettify	Beauty	derivationally related form	Prettify is a verb that means to make something or someone more beautiful or attractive, so it is a means to achieve Beauty.
FB15K-237	Star Trek Nemesis	Jerry Goldsmith	/film/film/music	Jerry Goldsmith composed the original score for the 2002 film Star Trek: Nemesis.
	White American	Sylvia Plath	/people/ethnicity/people	Sylvia Plath was a White American, meaning she was an American of European descent and therefore a member of the White American ethnic group.
	Robert Ridgely	USA	/people/person/nationality	Robert Ridgely was an American ornithologist, which means he was a citizen of the United States of America (USA) and made significant contributions to the field of bird study in the country.

Table 8: Some cases of ChatGPT and LLAMA3-70B output in the two datasets. Golden Label is the predefined KG relations. It is very interesting to note that there are indeed two Americans, both named Robert Ridgely. One is an actor and the other is an ornithologist. So ChatGPT and LLAMA3-70B both correctly complete the relationship between the two entities. Meanwhile, they also have the same problem: providing "redundant correct information".

Datasets	WN18RR
	Error Type 1: Generating fabricated or misplaced facts (hallucination of LLMs) .
Gloden Label	(genus dieffenbachia, _member_meronym, motherinlaw)
ChatPT Outputs	Mother-in-law plant is a common name for genus dieffenbachia.
LLAMA3-70B Output	The genus Dieffenbachia is commonly known as "Mother-in-Law".
	Error Type 2: Outputting “not related” for those entity pairs that should have relations.
Gloden Label	(bolt, _synset_domain_topic_of, politics)
ChatPT Outputs	Bolt and politics are not related.
LLAMA3-70B Output	Usain Bolt is a Jamaican sprinter who has no direct relation to politics.
	Error Type 3: Outputting incorrectly formatted or meaningless sentences.
Gloden Label	(close, _derivationally_related_form, shutting)
ChatPT Outputs	close: verb group -> shutting: verb.
Gloden Label	(righteous, _also_see, innocent)
LLAMA3-70B Outputs	Righteous and innocent are related, the former often encompassing the latter, but not always vice versa.
Datasets	FB15K-237
	Error Type 1: Generating fabricated or misplaced facts (hallucination of LLMs).
Gloden Label	(Canada, /olympic /participating_country/athletes./olympic_athlete/affiliation, 1994 Winter Olympics)
ChatPT Outputs	The 1994 Winter Olympics were held in Canada.
Gloden Label	(Philip Glass, /education/educational_institution/students_graduates./education/education/student, University of Chicago)
LLAMA3-70B Output	Philip Glass, the renowned American composer, attended the University of Chicago.
	Error Type 2: Outputting “not related” for those entity pairs that should have relations.
Gloden Label	(Cabbage, /food/nutrition_fact/nutrient, Iron)
ChatPT Outputs	Cabbage and iron are not related.
Gloden Label	(Japan, /film/film/release_date_s./film/film_regional_release_date/film_release_region, Iron)
LLAMA3-70B Outputs	There is no direct relation between Japan and Insidious, as Insidious is a 2011 American supernatural horror film and Japan is a country located in East Asia.
	Error Type 3: Outputting incorrectly formatted or meaningless sentences.
Gloden Label	(Drums, /music/instrument/instrumentalists, Jack White)
ChatPT Outputs	Drums and Jack White are associated as musicians and instruments.
Gloden Label	(Canada, /olympic /participating_country/athletes./olympic_athlete/affiliation, 1994 Winter Olympics)
LLAMA3-70B Outputs	Canada did not host the 1994 Winter Olympics, which were held in Lillehammer, Norway, from February 12 to February 27, 1994.

Table 9: Some error cases in the two datasets. Golden Label is the predefined KG schema. Error Types are described in Appendix E.