

# Multilingual Knowledge Graph Completion via Efficient Multilingual Knowledge Sharing

Cunli Mao<sup>1,2</sup>, Xiaofei Gao<sup>1,2</sup>, Ran Song<sup>1,2\*</sup>, Shizhu He<sup>3,4</sup>

Shengxiang Gao<sup>1,2</sup>, Kang Liu<sup>3,4</sup>, Zhengtao Yu<sup>1,2</sup>

<sup>1</sup>Faculty of Information Engineering and Automation,

Kunming University of Science and Technology, Kunming, China

<sup>2</sup>Yunnan Key Laboratory of Artificial Intelligence, Kunming, China

<sup>3</sup>The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,

Institute of Automation, Chinese Academy of Sciences, Beijing, China

<sup>4</sup>School of Artificial Intelligence, University of Chinese Academy of Science, Beijing, China

{maocunli,xiaofeigao\_g,song\_ransr}@163.com, {shizhu.he,kliu}@nlpr.ia.ac.cn,

{gaoshengxiang.yn,ztyu}@hotmail.com

## Abstract

Large language models (LLMs) based Multilingual Knowledge Graph Completion (MKGC) aim to predict missing facts by leveraging LLMs' multilingual understanding capabilities, improving the completeness of multilingual knowledge graphs (KGs). However, existing MKGC research underutilizes the multilingual capabilities of LLMs and ignores the shareability of cross-lingual knowledge. In this paper, we propose a novel MKGC framework that leverages multilingual shared knowledge to significantly enhance performance through two components: Knowledge-level Grouped Mixture of Experts (KL-GMoE) and Iterative Entity Reranking (IER). KL-GMoE efficiently models shared knowledge, while IER significantly enhances its utilization. To evaluate our framework, we constructed a mKG dataset containing 5 languages and conducted comprehensive comparative experiments with existing state-of-the-art (SOTA) MKGC method. The experimental results demonstrate that our framework achieves improvements of 5.47%, 3.27%, and 1.01% in the Hits@1, Hits@3, and Hits@10 metrics, respectively, compared with SOTA MKGC method. Further experimental analysis revealed the properties of knowledge sharing in settings of unseen and unbalanced languages. We have released the dataset and code for our work on <https://github.com/gaoxiaofei07/KL-GMoE>.

## 1 Introduction

Knowledge Graphs (KGs) (Weikum, 2021) are structured semantic knowledge bases designed to represent and organize knowledge about the real world. Most KGs possess multilingual characteristics, including Wikidata (Vrandečić and Krötzsch,

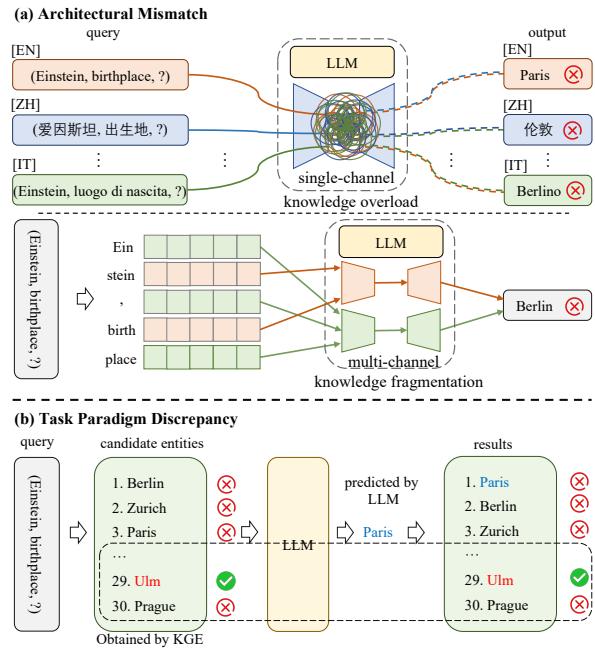


Figure 1: This figure depicts the problems encountered when applying LLMs directly to the MKGC task. (a) illustrates that existing PEFT is not suitable for the MKGC task. (b) indicates a discrepancy between the task paradigms LLMs excel at and MKGC tasks.

2014) and DBpedia (Lehmann et al., 2015). Existing multilingual KGs are often incomplete, which limits their effectiveness in practical applications (Ji et al., 2022). Multilingual knowledge graph completion (MKGC) aims to leverage known multilingual knowledge to complete missing triples and improve the completeness of the KGs.

Studies have focused on embedding-based methods (Ge et al., 2024) for MKGC, mapping entities and relations into a low-dimensional vector space to achieve completion. Recent advances in language models have shifted MKGC research toward generation-based approaches (Chen et al., 2022; Saxena et al., 2022) that reformulate KG comple-

\*Corresponding author

tion as text generation. Methods (Song et al., 2023; Zhou et al., 2022) employ a single pretrained language model (PLM) to consolidate multilingual knowledge within a unified semantic space, achieving superior performance in MKGC. Furthermore, recent work like DIFT (Liu et al., 2024) explores task adaptation through LLMs fine-tuning, achieving strong performance on monolingual KGC tasks. Modern LLMs, pretrained on diverse corpora, inherently possess multilingual capabilities (Huang et al., 2024), enabling the representation and knowledge sharing across languages (Hu et al., 2025) within a unified model. Crucially, this capacity for internal multilingual knowledge sharing is vital for MKGC, offering substantial potential to enhance completion performance. Motivated by this recognized potential, our research investigates effective methods for harnessing LLMs’ inherent capabilities to improve MKGC.

However, directly applying LLMs to MKGC presents several challenges, primarily stemming from two key aspects: model architecture and task paradigm. 1) **Architectural Mismatch**: Existing Parameter-Efficient Fine-Tuning (PEFT) methods (Han et al., 2024) for LLMs are mainly designed for text-centric tasks and exhibit significant gaps when applied to knowledge-level tasks. Specifically, single-channel methods struggle with the complex multilingual nature of KGs. As shown in Figure 1(a) top, processing numerous multilingual queries through a single channel often results in knowledge overload. This overload impacts the model’s ability to understand similar knowledge across languages, leading to incorrect predictions. For example, queries in English, Chinese, and Italian concerning *Einstein’s birthplace* all yield incorrect results. Conversely, multi-channel methods tend to disrupt the atomicity of knowledge, thereby causing knowledge fragmentation. As shown in Figure 1(a) bottom, query tokens are processed by disparate channels. Such fragmented processing consequently leads to incorrect entity predictions. 2) **Task Paradigm Discrepancy**: The MKGC task involves entity ranking, which presents a discrepancy with the text generation paradigm. As shown in Figure 1(b), for the query (*Einstein, birthplace, ?*), the LLM erroneously predicted *Paris* as the answer. This selection failed to improve the ranking of the correct entity *Ulm*.

To address the 1) **Architectural Mismatch**, specifically knowledge overload, we propose increasing the number of dedicated knowledge chan-

nels. This allows each channel to focus on processing semantically similar information, thereby enhancing the LLM’s capacity to understand and leverage cross-lingual shared knowledge. Concurrently, by enabling each channel to independently process complete knowledge, we can effectively mitigate knowledge fragmentation and facilitate the model’s comprehensive understanding of multilingual information. To resolve the 2) **Task Paradigm Discrepancy**, we propose adjusting the LLM’s training objective to enable it to iteratively refine the ranking of multiple entities. This approach aims to enhance the ranking of correct entities by increasing the frequency with which the LLM utilizes cross-lingual shared knowledge.

In this paper, we propose a novel framework for effectively leveraging multilingual shared knowledge to enhance the performance of MKGC. This proposed framework comprises two synergistic components: Knowledge-level Grouped Mixture of Experts (KL-GMoE) and Iterative Entity Reranking (IER). KL-GMoE introduces a knowledge-level expert routing mechanism and a group-based Mixture-of-Experts (MoE) architecture. This design aims to mitigate knowledge fragmentation while substantially enhancing LLMs’ capacity to capture cross-lingual shared knowledge. IER modifies both the training objective and the decoding strategy of LLMs. This enables the models to significantly improve their leveraging of multilingual shared knowledge through multiple iterative refinements. The experimental results demonstrate that our framework achieves improvements of 5.47%, 3.27%, and 1.01% in the Hits@1, Hits@3, and Hits@10 metrics, respectively, compared with SOTA MKGC method. Further experimental analysis revealed the properties of knowledge sharing in settings of unseen and unbalanced languages.

In summary, our contributions are as follows:

- We propose KL-GMoE to address the model architecture mismatch, efficiently modeling shared knowledge.
- We propose IER to address the discrepancy in the task paradigm, enhancing the utilization of shared knowledge.
- Experiments show that our framework significantly outperforms the SOTA MKGC method, with average improvements of 5.47%, 3.27%, and 1.01% in Hits@1, Hits@3, and Hits@10.

Language	Entity	Relation	Training	Validation	Testing
EN	86,539	512	708,267	49,782	49,777
FR	89,754	478	839,623	49,908	30,000
IT	65,434	445	613,014	49,883	20,000
JA	46,294	432	321,237	49,939	10,000
ZH	63,278	397	546,626	49,969	10,000
SUM	351,299	2,264	3,028,767	249,481	119,777

Table 1: Statistics of the multilingual knowledge graph completion dataset.

## 2 Datasets

### 2.1 Dataset Construction

We utilize *Wikidata5M* (Wang et al., 2021) as the foundational seed library, which is a million-scale English KG dataset integrating *Wikidata* and *Wikipedia*. Based on this, we further expanded the dataset to include French, Italian, Chinese, and Japanese, by collecting data from *Wikidata*. As shown in Table 1, we present statistics on the number of entities, relations, training, validation and testing triples. The KG contains 351,299 entities and 2,264 relations, with the total number of triples exceeding 3 million.

Based on the characteristics of multilingual knowledge distribution, the knowledge across different languages is not entirely aligned but exhibits certain linguistic specificity (Song et al., 2025). This asymmetry of knowledge across languages indicates that some knowledge is confined to specific languages. Therefore, the dataset we constructed follows the natural distribution patterns of knowledge. Some knowledge is shared across multiple languages, reflecting the similarities between languages. Other knowledge is unique to each language, reflecting the distinctive characteristics of each language.

### 2.2 Prompt Construction

We adopt the prompt construction method proposed by DIFT (Liu et al., 2024). Since the embedding-based model has learned the training data, it tends to rank the correct entity at the first in the candidate entities for most training facts. Constructing the prompt using these ranked candidates may cause LLMs to develop a bias toward selecting the first entity as the answer. Therefore, we partition a subset from the validation set to construct prompts, which are utilized as training data during the fine-tuning phase of the LLM.

For the query  $q = (h, r, ?)$ , the constructed Prompt P consists of four parts: Query Q, Descrip-

tion D, Neighbor facts N, and Candidate entities  $M_c$ . This can be represented as:

$$P(q) = [Q; D; N; M_c]. \quad (1)$$

Description provides specific descriptive information about entity  $h$ , enabling the model to comprehend the entity’s meaning more accurately. Neighbor facts are triples that include the entity  $h$ , and these triples are randomly sampled from the Knowledge Graph Embedding (KGE) model’s training data. These neighboring facts are intended to enhance the LLM’s comprehension of the entity  $h$ . Candidate entities  $M_c = [e_1, e_2, \dots, e_m]$  are composed of the top- $m$  entities selected from the ranking results generated by the KGE model (Bordes et al., 2013). To enable LLMs to adapt to the task paradigm of MKGC, we processed the number of entities  $m$  in the  $M_c$  during training. The specific processing method is detailed in Section 3.3. We provide specific prompt examples in Appendix A.1.

## 3 Methodology

### 3.1 Task Definition

In this paper, we integrate KGE model with LLM to perform the MKGC task. First, for a query  $q = (h, r, ?)$ , we use the KGE model to obtain the top- $m$  ranked entities, which form the candidate entities  $M_c$ . Next, we leverage LLMs to select the optimal entity from the candidate entities  $M_c$  to complete the query  $q$ . The completion process can be formulated as follows:

$$\hat{e} = \underset{e_i \in M_c}{\operatorname{argmax}} P(e_i | h, r, M_c), \quad (2)$$

where  $\hat{e}$  denotes the optimal entity for completing the query  $q$ , and  $P(e_i | h, r, M_c)$  represents the probability of selecting entity  $e_i$  given the head entity  $h$ , relation  $r$  and candidate entities  $M_c$ .

### 3.2 KL-GMoE Architecture

KL-GMoE is specifically tailored for MKGC task. This architecture is designed with multiple expert groups to alleviate the knowledge overload caused by single-channel and enhance the ability of LLMs to capture shared knowledge. Furthermore, KL-GMoE employs a knowledge-level expert routing mechanism to ensure that each sample is processed by a specific expert, rather than involving all experts collectively. As shown in Figure 2(a), for

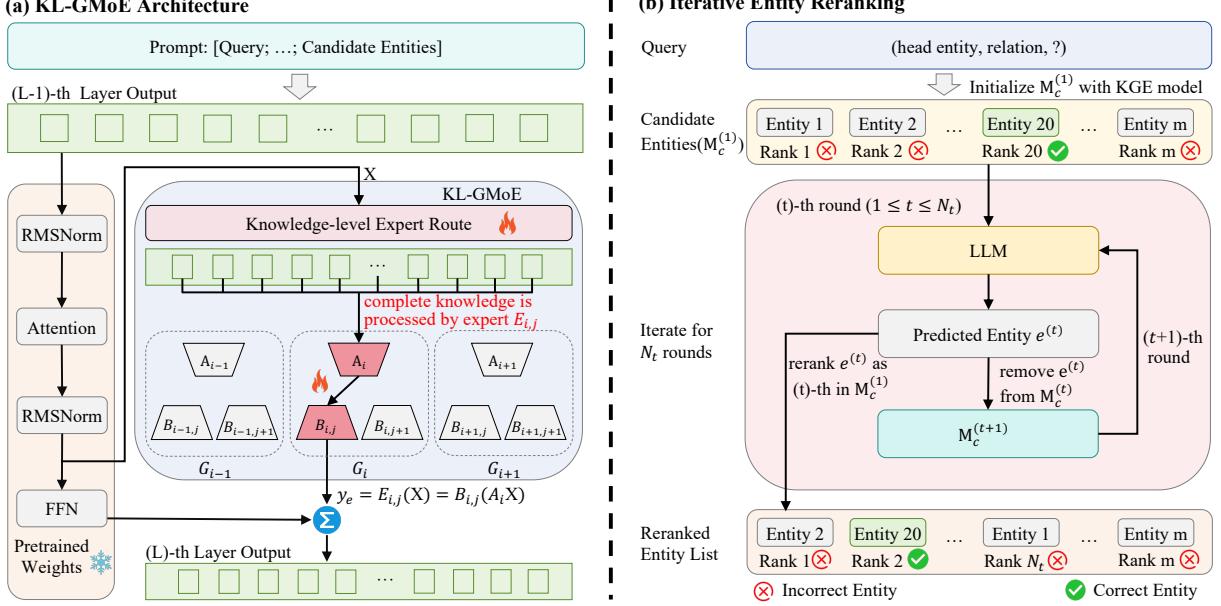


Figure 2: The figure illustrates our proposed framework. Figure (a) depicts the architecture and workflow of the KL-GMoE, where the matrices  $A_i$  and  $B_{i,j}$  highlighted in red represent the currently activated expert. Figure (b) illustrates the workflow of the IER method. After  $N_t$  iterations, we can obtain a reranked list of entities.

each sample processed by KL-GMoE, only the matrix  $A_i$  and one matrix  $B_{i,j}$  from the expert group  $G_i$  are activated. The KL-GMoE is applied exclusively to the Feed-Forward Network (FFN) layer of the LLM. Specifically, the matrix operations in the FFN layer during forward propagation can be represented as follows:

$$\mathbf{y} = \mathbf{W}_0 \mathbf{X} + \mathbf{y}_e, \quad (3)$$

where  $\mathbf{W}_0 \in \mathbb{R}^{\text{dout} \times \text{din}}$  represents the original FFN layer parameter matrix, which is frozen during training.  $\mathbf{X} = [\mathbf{x}_h : \mathbf{x}_r : \mathbf{x}_t]$  represents the input of the FFN layer.  $\mathbf{y}_e$  represents the output calculated by KL-GMoE based on the input  $\mathbf{X}$ .

The design of KL-GMoE is inspired by the asymmetric fine-tuning architecture proposed in HydraLoRA (Tian et al., 2024). We adopted a grouped MoE design architecture, where each group can be represented as follows:

$$G_i = (A_i, \{B_{i,j} \mid j \in \{1, 2, \dots, N_b\}\}), \quad (4)$$

where  $i \in \{1, 2, \dots, N_g\}$ ,  $N_g$  denotes the total number of expert groups.  $N_b$  is the number of B matrices in group  $G_i$ . Within each group  $G_i$ , the pairing of the  $A_i$  matrix with a  $B_{i,j}$  matrix is considered an expert  $E_{i,j} = (A_i, B_{i,j})$ . The  $A_i$  matrix is designed to capture a category of similar knowledge. The different  $B_{i,j}$  matrices within the group  $G_i$  are regarded as modules that capture subtle differences in this category of knowledge. This design

aims to enhance LLMs' ability to capture shared knowledge across multiple languages.

Simultaneously, our proposed knowledge-level expert routing mechanism includes three different routes:  $\mathbf{R}_g$ ,  $\mathbf{R}_k$  and  $\mathbf{R}_l$ . First, an expert group is selected based on  $\mathbf{R}_g$ . Within this group, a specific expert is then determined by combining  $\mathbf{R}_k$  and  $\mathbf{R}_l$ . The following describes the process of selecting a specific expert based on these three routes.

$\mathbf{R}_g$  is the group routing selection module that determines which expert group processes the  $\mathbf{X}$ . The group selection is formulated as follows:

$$\begin{aligned} G_i &= \operatorname{argmax}_{i \in \{1, 2, \dots, N_g\}} (\mathbf{R}_g(\mathbf{X})) \\ &= \operatorname{argmax}_{i \in \{1, 2, \dots, N_g\}} \left( \sum_{m \in \{h, r, t\}} \operatorname{Softmax}(\mathbf{W}_g \mathbf{x}_m) \right), \end{aligned} \quad (5)$$

where  $\mathbf{W}_g \in \mathbb{R}^{N_g \times \text{din}}$  is the routing matrix for group selection.  $G_i$  represents the expert group selected to process  $\mathbf{X}$ .

$\mathbf{R}_k$  and  $\mathbf{R}_l$  represent expert routing selection modules that operate within the group  $G_i$ . These modules comprehensively consider the input  $\mathbf{X}$  and the output from the  $A_i$  matrix to perform expert selection. Specifically,  $\mathbf{R}_k$  generates expert selection scores  $\mathbf{S}_k$  based on  $\mathbf{X}$ . The formula for

calculating  $\mathbf{S}_k$  is as follows:

$$\mathbf{S}_k = \mathbf{R}_k(\mathbf{X}) = \sum_{m \in \{h, r, t\}} \text{Softmax}(\mathbf{W}_k \mathbf{x}_m), \quad (6)$$

where  $\mathbf{S}_k \in \mathbb{R}^{N_b}$ , and  $\mathbf{W}_k \in \mathbb{R}^{N_b \times \text{din}}$  is the routing matrix that receives  $\mathbf{X}$  as input.  $\mathbf{R}_l$  generates expert selection scores  $\mathbf{S}_l$  based on the output  $A_i \mathbf{X}$  of matrix  $A_i$ . The calculation of  $\mathbf{S}_l$  can be expressed as follows:

$$\mathbf{S}_l = \mathbf{R}_l(\mathbf{X}) = \sum_{m \in \{h, r, t\}} \text{Softmax}(\mathbf{W}_l(A_i \mathbf{x}_m)), \quad (7)$$

where  $\mathbf{S}_l \in \mathbb{R}^{N_b}$ , and  $\mathbf{W}_l \in \mathbb{R}^{N_b \times r}$  is the routing matrix that receives  $A_i \mathbf{X}$  as input.  $r$  represents the size of the rank in LoRA (Hu et al., 2022). Then, select the matrix  $B_{i,j}$  from group  $G_i$  based on the scores of  $\mathbf{S}_k$  and  $\mathbf{S}_l$  to process  $\mathbf{X}$ :

$$B_{i,j} = \underset{j \in \{1, 2, \dots, N_b\}}{\text{argmax}} (\mathbf{S}_k + \mathbf{S}_l). \quad (8)$$

Finally, we determine that the expert  $E_{i,j} = (A_i, B_{i,j})$  processes  $\mathbf{X}$  based on the knowledge-level expert routing mechanism.

After determining the expert  $E_{i,j}$ , the output of KL-GMoE is expressed as follows:

$$\mathbf{y}_e = E_{i,j}(\mathbf{X}) = B_{i,j}(A_i \mathbf{X}). \quad (9)$$

Then, the expert output  $\mathbf{y}_e$  is added to the original FFN output, as shown in Equation 3.

### 3.3 Iterative Entity Reranking

We propose a method called Iterative Entity Reranking (IER), aimed at enhancing LLMs’ utilization of cross-lingual shared knowledge. As shown in Figure 2(b), the IER method fully leverages shared knowledge through multiple iterations, significantly improving the accuracy of correct entity ranking. IER adjusts the training task and decoding strategy of LLMs. In the training phase, we randomly set the number of candidate entities  $m$  to a variable value, to train the LLM to be capable of iteratively adjusting the ranking of multiple entities. In the decoding stage, IER allows the LLMs to perform multiple rounds of entity prediction to adjust the ranking of multiple entities.

For the query  $q = (h, r, ?)$ , the initial set of candidate entities is generated by the KGE model and denoted as  $M_c^{(1)} = [e_1, e_2, \dots, e_m]$ . The list of entities to be sorted is initialized as  $L^{(1)} = M_c^{(1)}$ . The LLM performs  $N_t$  rounds of entity prediction. In

the  $t$ -th round, where  $t \in \{1, 2, \dots, N_t\}$ , the entity prediction operation can be expressed as follows:

$$e^{(t)} = \underset{e_i \in M_c^{(t)}}{\text{argmax}} P(e_i | h, r, M_c^{(t)}), \quad (10)$$

where  $M_c^{(t)}$  represents the candidate entity set in round  $t$ .  $e^{(t)}$  is the entity predicted by the LLM from  $M_c^{(t)}$ . Then, we update  $M^{(t)}$  to obtain  $M^{(t+1)}$  for the next iteration:

$$M_c^{(t+1)} = M_c^{(t)} \setminus \{e^{(t)}\}, \quad (11)$$

where  $M_c^{(t)} \setminus \{e^{(t)}\}$  denotes removing the entity  $e^{(t)}$  from  $M_c^{(t)}$ . Finally, we update the ranking of entity  $e^{(t)}$  in  $L^{(t)}$ :

$$L^{(t+1)} = \text{Insert}(L^{(t)} \setminus \{e^{(t)}\}, t, e^{(t)}), \quad (12)$$

where  $\text{Insert}(L^{(t)} \setminus \{e^{(t)}\}, t, e^{(t)})$  denotes first removing  $e^{(t)}$  from  $L^{(t)}$ , and then inserting  $e^{(t)}$  into the  $t$ -th position of  $L^{(t)}$ . After iterating for  $N_t$  rounds, we obtain the final ranked list of entities  $L^{(N_t+1)}$ . The implementation of IER is detailed in Appendix A.2.

## 4 Experiment

### 4.1 Implementation Details

In the experiment, we selected TransE (Bordes et al., 2013) to obtain candidate entities. Additionally, we utilized Llama-2-7b-chat-hf<sup>1</sup> as the base model for fine-tuning. The model training hyperparameters are set as follows: the learning rate is 2e-5, the LoRA rank is 4, and the length of the candidate entities  $M_c$  is between 25 and 30. The number of iterations  $N_t$  for the IER is 10.

### 4.2 Multilingual Knowledge Graph Completion

We compared the performance of the proposed framework with embedding-based and generation-based methods on our constructed dataset. The experimental results demonstrate that our method achieves optimal performance on the average metrics across all languages. Specifically, as shown in Table 2, the performance of our proposed framework surpasses all the aforementioned methods in the three languages: EN, FR, and IT. For JA and ZH, our framework performed excellently on all metrics except Hits@10. Our framework failed

<sup>1</sup><https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

	MODEL	EN	FR	IT	JA	ZH	AVG
H@1	TransE	8.52	9.07	9.36	8.00	11.77	9.34
	Analogy	13.40	15.81	14.58	15.11	6.43	13.07
	ComplEx	10.92	11.75	11.49	13.95	17.79	13.18
	Distmult	6.89	7.73	7.93	8.16	5.66	7.27
	RotatE	24.08	24.61	25.57	29.49	31.36	27.02
	HAKE	31.64	32.92	30.99	35.53	52.24	36.66
	ICL	1.79	1.07	1.26	1.93	2.27	1.66
	GC-PLM	33.37	32.51	30.38	36.65	49.13	36.41
	DIFT ( <i>Single</i> )	36.05	35.75	34.22	38.31	56.65	40.19
H@3	Ours	<b>36.50</b>	<b>36.72</b>	<b>35.93</b>	<b>41.60</b>	<b>58.63</b>	<b>41.88</b>
	TransE	37.02	39.17	37.57	44.59	60.51	43.78
	Analogy	28.39	30.45	29.56	35.84	19.61	28.77
	ComplEx	23.05	23.92	23.05	29.68	41.12	28.17
	Distmult	14.59	14.82	15.79	19.16	17.94	16.46
	RotatE	40.73	42.13	41.68	49.78	62.36	47.34
	HAKE	43.30	43.27	41.88	47.52	63.22	47.84
	ICL	34.99	37.09	35.33	42.84	59.15	41.88
	GC-PLM	40.99	42.21	40.47	50.95	63.67	47.66
H@10	DIFT ( <i>Single</i> )	42.21	42.28	40.50	47.83	64.50	47.46
	Ours	<b>46.25</b>	<b>45.30</b>	<b>44.22</b>	<b>51.97</b>	<b>66.93</b>	<b>50.93</b>
	TransE	50.25	51.23	49.60	58.10	71.80	56.19
	Analogy	39.17	41.73	40.30	48.78	64.45	46.89
	ComplEx	34.84	37.51	35.70	45.02	59.97	42.61
	Distmult	26.74	26.78	27.41	36.04	50.25	33.44
	RotatE	52.66	53.17	51.50	<b>61.68</b>	<b>74.58</b>	58.72
	HAKE	53.37	52.06	50.49	57.85	70.04	56.76
	ICL	49.99	51.05	49.30	57.99	71.71	56.01
M@R	GC-PLM	52.76	52.81	51.76	59.53	71.98	57.77
	DIFT ( <i>Single</i> )	52.48	52.35	50.30	58.74	72.08	57.19
	Ours	<b>54.71</b>	<b>53.45</b>	<b>52.31</b>	60.85	72.56	<b>58.78</b>
	TransE	24.85	25.97	25.46	28.38	37.51	28.43
	Analogy	22.85	25.01	23.82	27.44	21.04	24.03
	ComplEx	19.18	20.29	19.56	24.30	32.12	23.09
	Distmult	13.33	13.80	14.28	16.86	16.69	14.99
	RotatE	34.63	35.39	35.49	41.75	48.74	39.20
	HAKE	39.41	39.80	38.10	43.36	59.04	43.94
R@M	ICL	20.11	20.70	19.80	23.98	31.06	23.13
	GC-PLM	36.66	37.21	37.18	42.21	55.39	41.73
	DIFT ( <i>Single</i> )	40.99	40.64	39.09	44.54	61.38	45.33
	Ours	<b>42.96</b>	<b>42.58</b>	<b>41.69</b>	<b>48.33</b>	<b>63.74</b>	<b>47.86</b>

Table 2: This table presents the MKGC results across five languages. The embedding-based methods TransE (Bordes et al., 2013), Analogy (Liu et al., 2017), ComplEx (Trouillon et al., 2016), DistMult (Yang et al., 2014), and RotatE (Sun et al., 2019) are all implemented using the OpenKE framework (Han et al., 2018). The results of HAKE (Zhang et al., 2020) were reproduced using its open-source code. ICL refers to evaluation using the LLaMA-2-7b-chat model without fine-tuning. GC-PLM (Song et al., 2023) represents the current SOTA method for MKGC. DIFT (Liu et al., 2024) is a SOTA LLM-based monolingual KGC method. The *Single* refers to training a separate model for each language independently. The numbers in bold represent the best results among the methods and languages considered.

to surpass RotatE’s performance on Hits@10, primarily attributed to our use of a relatively weaker-

	MODEL	EN	FR	IT	JA	ZH	AVG
H@1	LoRAMoE	36.28	36.36	35.81	40.22	56.87	41.11
	@ HydraLoRA	35.68	35.60	35.05	40.49	57.92	40.95
	Ours	<b>36.50</b>	<b>36.72</b>	<b>35.93</b>	<b>41.60</b>	<b>58.63</b>	<b>41.88</b>
H@3	LoRAMoE	42.54	42.74	41.40	48.75	64.96	48.08
	@ HydraLoRA	42.49	42.58	41.22	48.68	64.99	47.99
	Ours	<b>42.87</b>	<b>43.08</b>	<b>41.70</b>	<b>48.94</b>	<b>65.06</b>	<b>48.33</b>
H@10	LoRAMoE	52.27	52.50	50.91	59.07	72.19	57.39
	@ HydraLoRA	52.51	52.41	50.81	<b>59.10</b>	72.24	57.41
	Ours	<b>52.63</b>	<b>52.62</b>	<b>51.17</b>	59.09	<b>72.32</b>	<b>57.57</b>
M@R	LoRAMoE	41.15	41.13	40.30	45.90	61.52	46.00
	@ HydraLoRA	40.80	40.62	39.81	46.06	62.18	45.89
	Ours	<b>41.42</b>	<b>41.44</b>	<b>40.51</b>	<b>46.80</b>	<b>62.67</b>	<b>46.57</b>

Table 3: This table compares the KL-GMoE with the existing SOTA fine-tuning methods LoRAMoE (Dou et al., 2024) and HydraLoRA (Tian et al., 2024).

Model	Trainable Params	Activated Params	Lora Rank
TransE	106.1m	106.1m	-
DIFT(LoRA)	159.9*5 m	159.9*5 m	64
LoRAMoE	19.2m	19.2m	4
HydraLoRA	12.5m	12.5m	4
Ours	32.9 m	9.4 m	4

Table 4: This table shows the comparison of our method with other methods in terms of parameter count.

performing TransE model for generating candidate entities. We replaced TransE with RotatE in the candidate entities retrieval and conducted experiments. The corresponding results and analysis are presented in Appendix A.3. Compared to the existing SOTA MKGC method GC-PLM, our framework achieved significant performance advantages in Hits@1, Hits@3, Hits@10, and MRR metrics, with improvements of 5.47%, 3.27%, 1.01%, and 6.13%, respectively. Furthermore, our framework achieves a substantial improvement in performance compared with DIFT, the SOTA LLM-based monolingual KGC method. Overall, the experimental results clearly demonstrate the effectiveness and superiority of our proposed framework.

### 4.3 Model Architecture Comparison and Parameter Analysis

We compared the MKGC performance between the KL-GMoE architecture and existing SOTA fine-tuning methods, including LoRAMoE and HydraLoRA. These two SOTA methods both utilize multiple channels to process a single query, which can lead to the problem of knowledge fragmentation. As shown in Table 3, KL-GMoE outperforms

Model	Avg Tokens Num	TFLOPs
LoRAMoE	353.89	2.37814
HydraLoRA	353.89	2.37580
DIFT(LoRA)	353.89	2.42721
Ours	353.89	2.37472

Table 5: This table compares the computational efficiency of our method with that of other methods. The data presented are average values calculated from 1,000 samples.

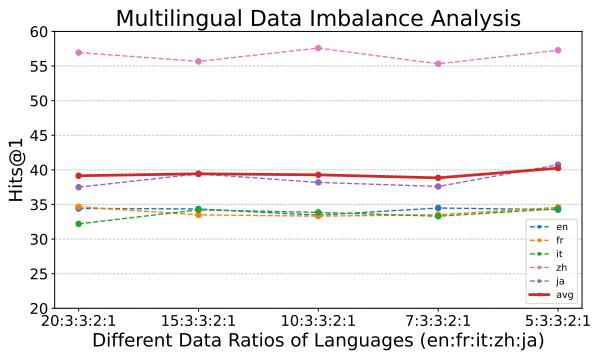


Figure 3: This figure shows the variation in Hits@1 scores of our framework under training data settings with five different language proportions.

these methods on average metrics. This experimental result demonstrates that our method effectively addresses knowledge fragmentation, thereby enhancing performance on the MKGC task.

We further analyzed the advantages of KL-GMoE in terms of model parameters. As shown in Table 4, compared to the embedding-based method TransE, KL-GMoE has 3.2 times fewer trainable parameters and 11.3 times fewer activated parameters. Among LLM-based methods, KL-GMoE has significantly fewer activated parameters than all other methods. In particular, compared to DIFT, KL-GMoE has approximately 24.3 times fewer trainable parameters and about 85.1 times fewer activated parameters, which demonstrates its significant advantages in terms of parameter count. At the same time, we compared the proposed method with other methods in terms of FLOPs during inference. As shown in Table 5, our method reduces the FLOPs by approximately 0.053 TFLOPs compared to the current state-of-the-art LLM-based KGC method, DIFT, demonstrating superior computational efficiency.

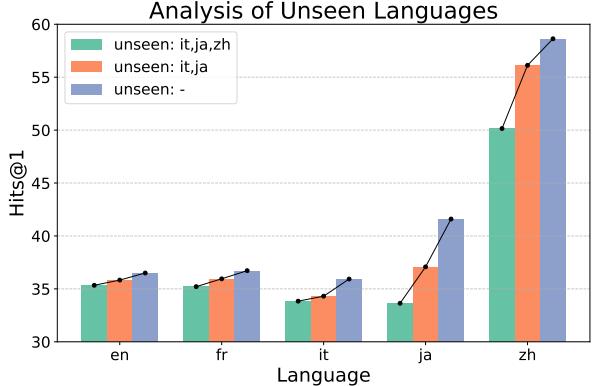


Figure 4: The figure illustrates the Hits@1 performance of our method on five languages under three different training language settings.

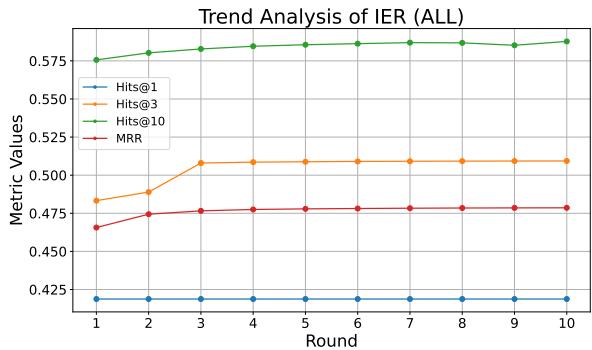


Figure 5: The figure illustrates the impact of the number of iterations of the IER method on performance.

#### 4.4 Analysis of Language Imbalance

To evaluate the robustness of our framework in scenarios with imbalanced language distribution in the training data, we conducted experiments. Specifically, we conducted experiments with imbalanced training data ratios across five languages, while keeping the total amount of training data constant. As shown in Figure 3, despite the changes in language proportions, the Hits@1 scores for each language (dashed lines) and the average score across the five languages (solid red line) remained relatively stable. It is evident that our framework is insensitive to variations in the language distribution. Based on this analysis, our framework can effectively leverage cross-lingual shared knowledge, thereby demonstrating strong robustness.

#### 4.5 Analysis of Unseen Languages

To evaluate the generalization capabilities of the proposed framework on languages not included in the training data, we conducted analysis experiments. These experiments were conducted with three distinct training configurations: (1) trained

Language		Query		Share	Answer	Prediction By DIFT	Prediction By Ours
EN	(Towelhead, composer, ?)	FR	Thomas Newman		David Kitay   Towelhead   <b>Thomas Newman</b>   James Newton Howard   Chris Messina   Rolfe Kent   Carrie Preston	<b>Thomas Newman</b>   Rolfe Kent   James Newton Howard   Mychael Danna   Theodore Shapiro   Mark Isham   Carter Burwell	
FR	(Shtji Terayama, occupation, ?)	IT	parolier		metteur ou metteuse en scène   journaliste   <b>parolier</b>   producteur de cinéma   musicien ou musicienne   compositeur ou compositeuse   autrice-compositrice-interprète	<b>parolier</b>   producteur de cinéma   guitariste   metteur en scène   compositeur ou compositeuse   journaliste   artiste	
IT	(The Pixar Story, interpetri, ?)	EN	John Lasseter		Joe Ranft   Eric Larson   <b>John Lasseter</b>   Milt Kahl   Glen Keane   James Algar   Don Hahn	<b>John Lasseter</b>   Eric Larson   Milt Kahl   Glen Keane   Chris Buck   Colin Hanks   James Algar	
JA	(パウロス, 同一とされる事物, ?)	ZH	バール		パウルス   パウロス   <b>バール</b>   ポーラ   クリストイーン   コンスタンティン   クリストイナ	バール   パウルス   ポーラ   パウロス   クリストイーン   コンスタンティン   クリストイナ	
ZH	(拉科西·马加什, 口头、书写语言, ?)	JA	俄語		匈牙利   拉科西·马加什   <b>俄語</b>   德语   中文   西班牙语   拉丁语	<b>俄語</b>   德语   日语   英语   拉丁语   匈牙利   波兰语	

Figure 6: The figure presents a comparison of the prediction results between our method and DIFT in the knowledge shared case. The **Share** column indicates that the knowledge of these queries exists in the LLM’s training data but is presented in other languages.

on EN and FR; (2) trained on EN, FR, and ZH; and (3) trained on five languages. As shown in Figure 4, the green bar indicates that LLMs trained solely on EN and FR data demonstrated significant KGC performance on unseen languages IT, JA, and ZH. This clearly demonstrates that knowledge sharing is effective not only among languages seen during LLM training, but also shows significant cross-lingual generalization capability among unseen languages. Furthermore, we observed a consistent improvement in performance across all languages as the number of training languages increased. This finding suggests that training data in more languages provides richer knowledge signals to LLMs, which facilitates the sharing of multilingual knowledge.

#### 4.6 Analysis of IER Trends

To evaluate the impact of the number of iterations in the IER method on MKGC performance, we conducted analytical experiments. Figure 5 illustrates the changes in all metrics as the number of iterations increases. From the results, it can be observed that Hits@3, Hits@10, and MRR significantly improved in the first three iterations and reached their optimal values by the tenth iteration. This trend indicates that with an increasing number of iterations, IER allows LLMs to leverage multilingual shared knowledge more effectively, thereby significantly improving the performance of MKGC.

#### 4.7 Ablation Experiment

To verify the effectiveness of each component in our proposed framework, we conducted ablation experiment. We evaluated the contribution of each component by removing it sequentially. As shown in Table 6, removing the KL-GMoE component resulted in a drop in Hits@1 from 41.88 to 40.28, Hits@3 from 50.93 to 49.71, Hits@10 from 58.78

Model	H@1	H@3	H@10	MRR
Ours	<b>41.88</b>	<b>50.93</b>	<b>58.78</b>	<b>47.86</b>
Ours w/o <i>kg</i>	40.28	49.71	58.07	46.55
Ours w/o <i>kg+ier</i>	40.28	47.66	57.29	45.42

Table 6: This table shows the results of ablation experiments on the KL-GMoE (*kg*) and IER (*ier*) components. All results are the average of the five language metrics.

to 58.07, and MRR from 47.86 to 46.55. This indicates that the KL-GMoE component is crucial for improving the performance of MKGC. Furthermore, when we removed both KL-GMoE and IER simultaneously, the values of Hits@3, Hits@10, and MRR further decrease compared to removing only KL-GMoE. This demonstrates that the IER component also makes a positive contribution to the performance of MKGC. These ablation experiment results strongly prove the effectiveness of our proposed KL-GMoE and IER components.

#### 4.8 Case Study

We conducted a case study to evaluate the framework’s performance in cross-lingual knowledge sharing. These case’ queries are knowledge that the LLM learned during its training, but expressed in another language. As shown in Figure 6, for the English query (*Towelhead, composer, ?*), the LLM has already learned this knowledge in the French training data. Our framework successfully leverages this French knowledge to accurately predict the entity as *Thomas Newman*. In contrast, SOTA LLM-based methods incorrectly predict *David Kitay*. This demonstrates that our framework can effectively utilize cross-lingual shared knowledge to improve completion accuracy.

## 5 Related Work

Embedding-based methods map entities and relations in KGs to low-dimensional vector spaces. For example, TransE (Bordes et al., 2013) based on the translation principle of entities and relations. RotatE (Sun et al., 2019) treats each relation as rotation in complex vector space. DMoG (Song et al., 2022a) represents the unseen relations of the factual graph by fusing ontology and textual graphs. TransH (Wang et al., 2014) models relation as hyperplane. HOLEX (Xue et al., 2018) interpolate between a high model complexity method and HolE (Nickel et al., 2016). TR-GCN (Song et al., 2022b) proposes an ontology-guided zero-shot relation learning method to represent unseen relations.

Generation-based Methods transform KGC task into text generation task. For example, KGT5 (Saxena et al., 2022) posing KG link prediction as a sequence-to-sequence task. GC-PLM (Song et al., 2023) enhances the performance of MKGC by introducing global and local knowledge constraints. GenKGC (Xie et al., 2022) introduces a hierarchical decoding strategy of relation-guided demonstration and entity awareness. KICGPT (Wei et al., 2023) integrates LLMs and KGE model, adopting a knowledge-prompted contextual learning strategy to rerank multiple entities. DIFT (Liu et al., 2024) fine-tunes LLMs using LoRA (Hu et al., 2022) to select the most optimal entity from candidate entities obtained by the KGE model.

## 6 Conclusion

In this paper, we propose a novel MKGC framework. This framework integrates two components: KL-GMoE and IER. KL-GMoE significantly improves completion performance by efficiently capturing shared knowledge across languages. IER fully utilized cross-lingual shared knowledge through a multi-round iterative approach, further improving completion performance. The experimental results demonstrate that our framework exhibits superior performance in the MKGC task.

## Limitations

Our framework is limited by the token length of the LLM, therefore it is unable to perform entity selection based on all entities in the KG. Moreover, the framework processes text information exclusively. This limitation impedes its application to

multimodal KG datasets, as it cannot integrate information from other modalities.

## Ethics Statement

The paper proposes a method for MKGC and conducts experiments on a multilingual dataset extended from public available datasets. Therefore, data privacy implications are non-existent in this scenario.

## Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant Nos. U21B2027, U23A2038, 62166023, 62376270), the Yunnan Provincial Major Science and Technology Special Plan Projects (Grant Nos. 202402AG050007, 202502AD080012, 202502AD080016), the General Projects of Basic Research in Yunnan Province (Grant Nos. 202301AS070047, 202201BE070001-021).

## References

- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. *Translating embeddings for modeling multi-relational data*. *Advances in neural information processing systems*, 26.
- 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.
- Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhi-heng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. *LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945, Bangkok, Thailand. Association for Computational Linguistics.
- Xiou Ge, Yun Cheng Wang, Bin Wang, C-C Jay Kuo, and 1 others. 2024. *Knowledge graph embedding: An overview*. *APSIPA Transactions on Signal and Information Processing*, 13(1).
- Xu Han, Shulin Cao, Lv Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. *Openke: An open toolkit for knowledge embedding*. In *Proceedings of EMNLP*.

- Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. Large language models are cross-lingual knowledge-free reasoners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1525–1542, Albuquerque, New Mexico. Association for Computational Linguistics.
- Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, and 1 others. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.
- Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and 1 others. 2015. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.
- Hanxiao Liu, Yuexin Wu, and Yiming Yang. 2017. Analogical inference for multi-relational embeddings. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 2168–2178. JMLR.org.
- Yang Liu, Xiaobin Tian, Zequn Sun, and Wei Hu. 2024. Finetuning generative large language models with discrimination instructions for knowledge graph completion. In *International Semantic Web Conference*, pages 199–217. Springer.
- Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.
- Apoorv Saxena, Adrian Kochsieck, 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.
- Ran Song, Shengxiang Gao, Xiaofei Gao, Cunli Mao, and Zhengtao Yu. 2025. Mke-pllm: A benchmark for multilingual knowledge editing on pretrained large language model. *Neurocomputing*, 651:130979.
- Ran Song, Shizhu He, Shengxiang Gao, Li Cai, Kang Liu, Zhengtao Yu, and Jun Zhao. 2023. Multilingual knowledge graph completion from pretrained language models with knowledge constraints. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7709–7721, Toronto, Canada. Association for Computational Linguistics.
- Ran Song, Shizhu He, Suncong Zheng, Shengxiang Gao, Kang Liu, Zhengtao Yu, and Jun Zhao. 2022a. Decoupling mixture-of-graphs: Unseen relational learning for knowledge graph completion by fusing ontology and textual experts. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2237–2246, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Ran Song, Shizhu He, Suncong Zheng, Shengxiang Gao, Kang Liu, Jun Zhao, and Zhengtao Yu. 2022b. Ontology-guided and text-enhanced representation for knowledge graph zero-shot relational learning. In *ICLR 2022 Workshop on Deep Learning on Graphs for Natural Language Processing*.
- 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*.
- Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. 2024. Hydralora: An asymmetric lora architecture for efficient fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML’16, page 2071–2080. JMLR.org.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.
- Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.
- Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI’14, page 1112–1119. AAAI Press.

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.

Gerhard Weikum. 2021. **Knowledge graphs 2021: A data odyssey**. *Proceedings of the VLDB Endowment*, 14(12):3233–3238.

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.

Yexiang Xue, Yang Yuan, Zhitian Xu, and Ashish Sabharwal. 2018. Expanding holographic embeddings for knowledge completion. *Advances in neural information processing systems*, 31.

Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations*.

Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. 2020. Learning hierarchy-aware knowledge graph embeddings for link prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3065–3072.

Wenxuan Zhou, Fangyu Liu, Ivan Vulić, Nigel Collier, and Muhao Chen. 2022. **Prix-LM**: Pretraining for multilingual knowledge base construction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5412–5424, Dublin, Ireland. Association for Computational Linguistics.

## A Appendix

### A.1 Prompt Example

We present prompt examples for candidate entity lists of varying lengths during the training phase.

Prompt	Given a triplet with a missing tail entity t: (Saint George and the Dragon, material used, t).  The following provides descriptive information about entity Saint George and the Dragon: Saint George and the Dragon, Saint George and the Dragon or Saint George Killing the Dragon is a 1555 or 1558 painting by the Venetian artist Tintoretto. It was later acquired by the English collector  Here are some triplets containing entity Saint George and the Dragon: [(Saint George and the Dragon, depicts, hill); (Saint George and the Dragon, depicts, spear); (Saint George and the Dragon, creator, Jacopo Tintoretto); (Saint George and the Dragon, depicts, combat); (Saint George and the Dragon, depicts, woman); (Saint George and the Dragon, depicts, sky)]  What is the entity name of t? Select one from the list of entities below: [oil paint; Saint George and the Dragon; wood; tempera; textile; brick; pearl; metamorphic rock; schist; sandstone; paint; igneous rock; tissue; gemstone; brass; copper; woven fabric; volcanic rock; marble; dragon; basalt; sedimentary rock; The Three Graces; limestone; steel]  [Answer]:
Number of entities	25
Prompt	Given a triplet with a missing tail entity t: (Jason Lee, instance of, t).  The following provides descriptive information about entity Jason Lee: Jason Lee, Jason Michael Lee (born April 25, 1970) is an American actor, photographer, producer, skateboarder, comedian, and writer. He is best known for his roles as Earl Hickey in the television  Here are some triplets containing entity Jason Lee: [(Mallrats, cast member, Jason Lee); (Jason Lee, ethnic group, Scottish American); (Jason Lee, occupation, screenwriter); (Jason Lee, occupation, actor); (Jason Lee, occupation, film producer); (Jason Lee, occupation, businessperson)]  What is the entity name of t? Select one from the list of entities below: [Jason Lee; human; twin; Jason Alexander; Sofía Vergara; Kevin Smith; Screen Actors Guild Award; David Cross; 3D film; college; Primetime Emmy Award; sports season; MTV Movie Awards; Kaley Cuoco; municipality of Spain; Jason Mewes; decade; military rank; suburb; animation studio; Jane Lynch; Hank Azaria; Satellite Award; Breckin Meyer; My Name Is Earl; Patrick Warburton; business]  [Answer]:
Number of entities	27

Table 7: Prompt examples for candidate entity lists of varying lengths.

## A.2 Details of the Iterative Entity Reranking Algorithm

---

**Algorithm 1** Iterative Entity Reranking (IER)

---

- 1: **Input:** Query  $q = (h, r, ?)$ ,  $M_c^{(1)} = [e_1, e_2, \dots, e_m]$ : the top- $m$  entities generated by the KGE model,  $N_t, L^{(1)} = M_c^{(1)}$
- 2: **for**  $t = 1$  to  $N_t$  **do**
- 3:    $e^{(t)} = \operatorname{argmax}_{e_i \in M_c^{(t)}} P(e_i | h, r, M_c^{(t)})$ ;
- 4:    $M_c^{(t+1)} = M_c^{(t)} \setminus \{e^{(t)}\}$ ;
- 5:    $L^{(t+1)} = \text{Insert}(L^{(t)} \setminus \{e^{(t)}\}, t, e^{(t)})$ ;
- 6: **end for**
- 7: **Output:**  $L^{(N_t+1)}$

---

## A.3 Analysis of Knowledge Graph Embedding Models

The experimental results in Table 8 clearly demonstrate that when using RotatE to retrieve candidate entities, our proposed method achieves a significant performance improvement compared to the original RotatE model, with a 14.78% increase in Hits@1. Notably, Ours+RotatE exhibits slightly lower performance than Ours+TransE on several language-specific metrics. This phenomenon can be attributed to the differing top-1 ranking rates of correct entities within the candidate sets generated by each KGE model. Specifically, the proportion of correct entities ranked as top-1 was 14.38% when using TransE, whereas this proportion significantly increased to 30.51% with RotatE. Therefore, this feature has had some impact: during the fine-tuning stage, LLM is more inclined to choose the entity that ranks first in the candidate set as the final answer. We hypothesize that this "top-1 bias" may, to some extent, suppress the model's exploration of other potentially correct answers, leading to Ours+RotatE performing slightly worse than Ours+TransE on some languages. In future work, we plan to further investigate how to construct a more stable fine-tuning instruction set that does not rely on traditional KGE models.

	MODEL	EN	FR	IT	JA	ZH	AVG
H@1	RotatE	24.08	24.61	25.57	29.49	31.36	27.02
	Ours+TransE	36.50	<b>36.72</b>	<b>35.93</b>	41.60	58.63	<b>41.88</b>
	Ours+RotatE	<b>36.55</b>	35.70	35.46	<b>41.78</b>	<b>59.49</b>	41.80
@3	RotatE	40.73	42.13	41.68	49.78	62.36	47.34
	Ours+TransE	46.25	<b>45.30</b>	<b>44.22</b>	51.97	66.93	50.93
	Ours+RotatE	<b>46.38</b>	45.05	44.20	<b>52.27</b>	<b>67.13</b>	<b>51.01</b>
H@10	RotatE	52.66	53.17	51.50	61.68	74.58	58.72
	Ours+TransE	54.71	53.45	52.31	60.85	72.56	58.78
	Ours+RotatE	<b>54.76</b>	<b>54.13</b>	<b>52.53</b>	<b>62.31</b>	<b>74.59</b>	<b>59.66</b>
M@R	RotatE	34.63	35.39	35.49	41.75	48.74	39.20
	Ours+TransE	42.96	<b>42.58</b>	<b>41.69</b>	48.33	63.74	47.86
	Ours+RotatE	<b>43.04</b>	42.18	41.58	<b>49.01</b>	<b>64.46</b>	<b>48.05</b>

Table 8: The impact of different KGE models on the performance of our proposed framework during the candidate entities retrieval process.

## A.4 Analysis of Expert Routing

To verify the existence of knowledge sharing, we analyzed the expert selection in the test samples. We obtained the output of each expert routing and visualized it from both the linguistic and knowledge

dimensions. The left side of Figure 7 shows the expert routing analysis based on the linguistic dimension. Illustrates the selection of experts in each Transformer layer of the LLM for samples in different languages. The expert selections across different language samples are mostly consistent, suggesting that our method does not distinguish between languages in MKGC task. The right side of the figure illustrates the expert selection for each relation. The analysis from figure shows that samples with same relations across different languages are mostly handled by same experts. Overall, some relations exhibit consistent expert selection, while others show differences. Based on this analysis, it is validated that our method can effectively leverage knowledge sharing across different languages.

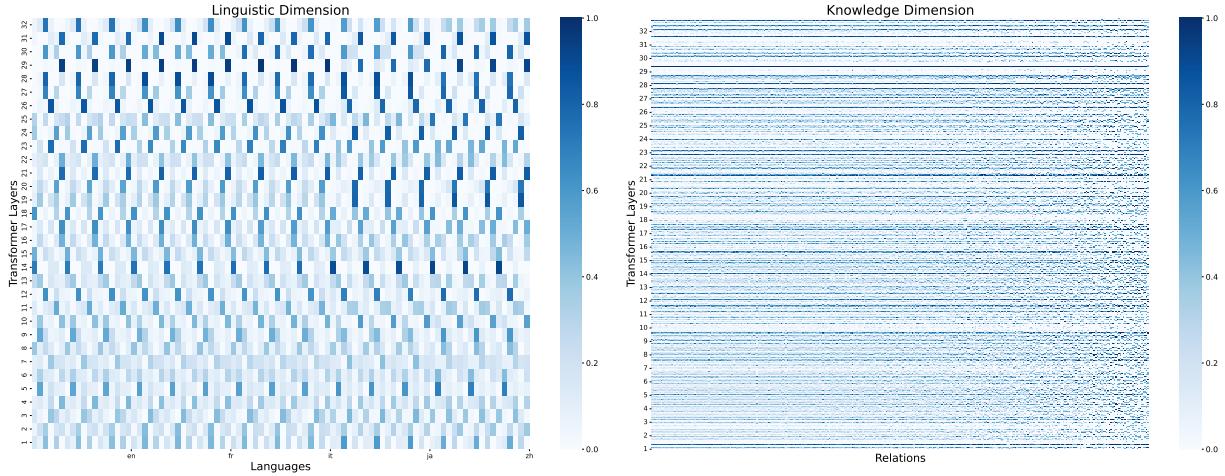


Figure 7: The left shows expert selection across five languages. The horizontal axis represents all languages, with each small bar within a language corresponding to an expert. The vertical axis indicates the layer numbers in the Transformer of the LLM. The color intensity of each blocks represents the frequency of samples selecting particular expert. The right depicts expert selection across all relations. The horizontal axis represents all relations. Vertical axis shows the layer numbers in Transformer, each row within a layer corresponding to an expert.

### A.5 IER Trend Analysis

We analyzed the impact of the number of iterations in the IER method on five language evaluation metrics. The experimental results reveal a clear performance improvement across all languages with increased iterations. These findings demonstrate the IER method’s ability to exploit cross-lingual shared knowledge.

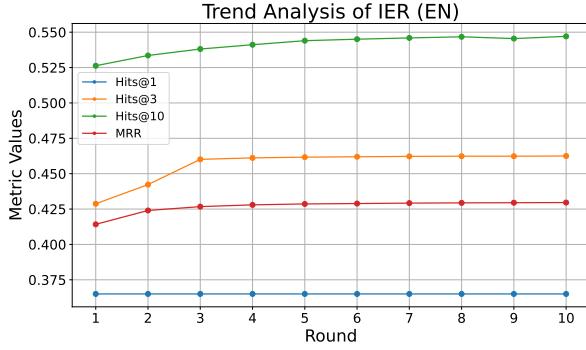


Figure 8: This figure shows how the metric performance of the IER method changes with the number of iterations on the English test set.

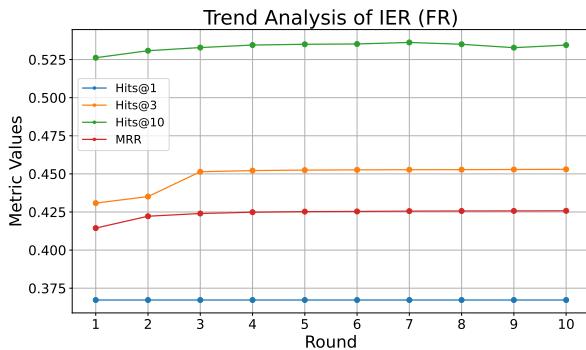


Figure 9: This figure shows how the metric performance of the IER method changes with the number of iterations on the French test set.

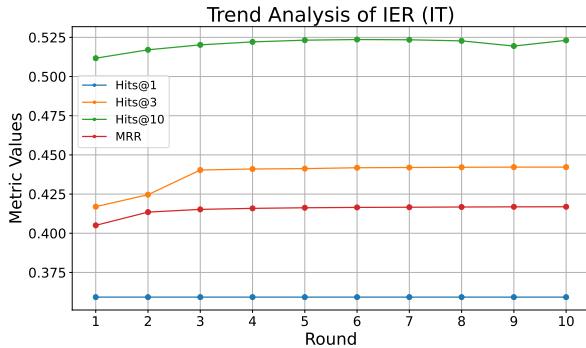


Figure 10: This figure shows how the metric performance of the IER method changes with the number of iterations on the Italian test set.

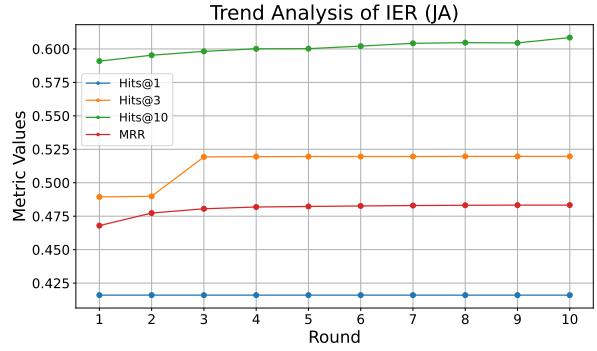


Figure 11: This figure shows how the metric performance of the IER method changes with the number of iterations on the Japanese test set.

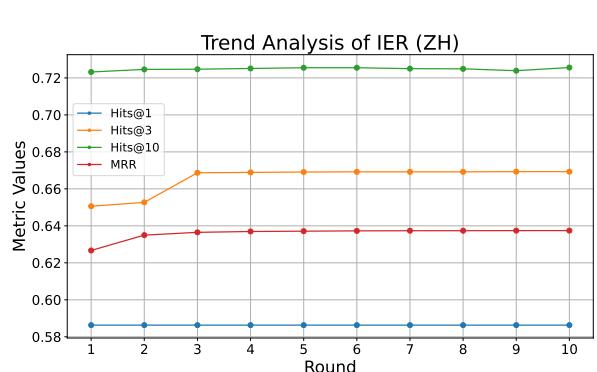


Figure 12: This figure shows how the metric performance of the IER method changes with the number of iterations on the Chinese test set.