MusKGC: A Flexible Multi-source Knowledge Enhancement Framework for Open-World Knowledge Graph Completion

Xin Song[1*], **Haiyan Liu**[1*], **Haiyang Wang**[1], **Ye Wang**[1†], **Kai Chen**[1], **Bin Zhou**[1†],
[1]National University of Defense Technology, Changsha, China
{songxin, haiyan_liu, wanghaiyang19, ye.wang, chenkai_, binzhou}@nudt.edu.cn

Abstract

Open-world knowledge graph completion (KGC) aims to infer novel facts by enriching existing graphs with external knowledge sources while maintaining semantic consistency under the open-world assumption (OWA). Generation-based KGC methods leverage the inherent strengths of large language models (LLMs) in language understanding and creative problem-solving, making them promising approaches. However, they face limitations: (1) The unreliable external knowledge from LLMs can lead to hallucinations and undermine KGC reliability. (2) The lack of an automated and rational evaluation strategy for new facts under OWA results in the exclusion of some new but correct entities. In the paper, we propose MusKGC, a novel multi-source knowledge enhancement framework based on an LLM for KGC under OWA. We induce relation templates with entity type constraints to link structured knowledge with natural language, improving the comprehension of the LLM. Next, we combine intrinsic KG facts with reliable external knowledge to guide the LLM in accurately generating missing entities with supporting evidence. Lastly, we introduce a new evaluation strategy for factuality and consistency to validate accurate inferences of new facts, including unknown entities. Extensive experiments show that our proposed framework achieves SOTA performance across benchmarks, and our evaluation strategy effectively assesses new facts under OWA.

1 Introduction

Knowledge Graphs (KGs) commonly store structured information about specific domains or real-world, and are widely applied in areas of question-answering (Sun et al., 2019a) or recommendation systems (Huang et al., 2018). Knowledge graph completion (KGC) aims to improve KG completeness by automatically inferring missing facts. Traditional KGC methods(Bordes et al., 2013; Trouillon et al., 2016; Vashishth et al., 2020) are typically based on the closed-world assumption (CWA), which assumes that any knowledge unseen in KGs is incorrect. Such methods prove inadequate for real-world applications where knowledge evolves dynamically, which arises from real-world KG applications' inherent requirements to continuously integrate emerging knowledge. This demand motivates KGC under the open-world assumption (OWA) (Yang et al., 2022; Lv et al., 2022), which allows the incorporation of external knowledge beyond the KG to infer emerging entities and facts.

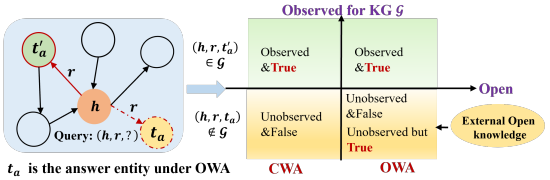


Figure 1: The difference between CWA and OWA. Under CWA, the fact $(h, r, t'_a) \in \mathcal{G}$ is observed and correct. The fact $(h, r, t_a) \notin \mathcal{G}$ is unobserved for KG, so false. However, under OWA, introducing external knowledge may result in $(h, r, t_a) \notin \mathcal{G}$ that is still correct, even though it is not observed in the KG.

With the enhanced capabilities of LLMs, generation-based KGC methods (Xie et al., 2022; Yao et al., 2023; Yang et al., 2025) have shown great potential under OWA by leveraging their vast parameterized knowledge as an external knowledge source. Moreover, the strong language understanding and reasoning capabilities of LLMs significantly aid in inferring new facts. These methods are mainly divided into two categories: (1) Directly generating missing entities based on LLMs (Xie et al., 2022; Yao et al., 2023), which formulates KGC as a text generation task to output target entities. Unfortunately, entities generated

[*] Equal contributions.
[†] Corresponding authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 10031–10049
November 4-9, 2025

by LLMs often require manual evaluation due to the uncontrollable and diverse nature of the generation process. (2) To address the shortcomings of the former, generative re-ranking methods based on LLMs (Wei et al., 2023; Wang et al., 2024) are proposed. They employ a lightweight KGE model to obtain an initial ranking of candidate entities, then leverage LLMs to rerank them. These approaches have gained substantial recognition, with their effectiveness rigorously demonstrated across multiple open-domain applications, including question answering (Khalifa et al., 2023) and KGC (Lv et al., 2022; Liu et al., 2024a; Wang et al., 2024).

Despite the significant progress achieved, the aforementioned methods still face two limitations: **(L1) The unreliable external knowledge from LLMs.** The existing works (Lv et al., 2022; Yang et al., 2025) excessively rely on the parametric knowledge within LLMs as external knowledge, but such reliance is problematic due to the opaque nature of LLMs' parametric knowledge, which leads to poor interpretability and frequent hallucination issues. **(L2) The lack of an automated and rational evaluation strategy for new facts under OWA.** In existing studies, they consider a predicted fact $(h, r, t'_a) \in \mathcal{G}$ is correct, otherwise false. This reflects a closed-world perspective based solely on the KG. However, a new fact $(h, r, t_a) \notin \mathcal{G}$ is not false, but rather unknown (Lv et al., 2022). As shown in Figure 1, we clearly demonstrate the difference between CWA and OWA.

To address the first limitation, inspired by Retrieval-Augmented Generation (RAG) (Huang and Huang, 2024), we attempt to incorporate non-parametric knowledge from reliable external sources. This helps reduce reliance on the parametric knowledge of LLMs, thereby alleviating hallucinations and enhancing interpretability. To tackle the second limitation, we argue that evaluating new facts under OWA should follow two principles: grounding in verifiable evidence for accuracy, and maintaining consistency with existing KG for coherence. Both principles jointly guide the evaluation of new facts, ensuring correctness and compatibility with the existing KG.

In the paper, we propose a flexible **Mu**lti-**S**ource knowledge enhancement framework based on an LLM for **KGC** under OWA, named **MusKGC**. To enhance the understanding and adaptability of the LLM to structured knowledge, we first conduct automated relation template induction with type constraints. It effectively transforms structured triplets into natural language sentences. Then, to mitigate hallucinations and unreliability in generation-based KGC under OWA, we design a multi-source knowledge enhancement module that jointly leverages the intrinsic knowledge from the KG and non-parametric information retrieved from a reliable external source. They jointly guide the LLM to generate more accurate and rational entities. Finally, we present a novel evaluation strategy to assess the factuality and consistency of new facts, enabling the discovery of correct but missing facts beyond the CWA.

Our contributions are threefold:

- We propose a flexible multi-source knowledge enhancement framework (MusKGC) for KGC under OWA. It mitigates hallucinations and enhances the reliability of generation-based KGC methods by retrieving key external knowledge from non-parametric sources.
- We develop a factuality and consistency evaluation strategy for generated facts, breaking limitations of CWA and facilitating the discovery of new facts beyond the KG.
- Extensive experiments and analysis are conducted to demonstrate the effectiveness and strengths of our MusKGC compared to SOTA baselines.

2 Related Works

2.1 Structure-based KGC methods

Structure-based KGC methods are built on the KG's structural information. The first group is translation-based models (TransE (Bordes et al., 2013) and its variants TransH (Wang et al., 2014), TransR (Lin et al., 2015)). The second group is tensor decomposition models, including ComplEx (Trouillon et al., 2016), RESCAL (Nickel et al., 2011). The third group is the GCN-based methods. For example, R-GCN (Schlichtkrull et al., 2018) and CompGCN (Vashishth et al., 2020). However, these methods overlook vast external knowledge, limiting their ability to address unseen knowledge and entities in open-world scenarios.

2.2 Text-based KGC methods

Recently, text-based KGC methods have gained increasing attention with the advancement of language models. Textual data serves as a complementary knowledge source, offering rich semantics to

enhance knowledge representation and inference. KG-BERT (Yao et al., 2019) is the first model that uses PLMs to perform KGC simply by concatenating entity and relation labels as input. Subsequent approaches (for example, MTL-KGC (Kim et al., 2020), PKGC (Lv et al., 2022), SimKGC (Wang et al., 2022a), CP-KGC (Yang et al., 2024), and GS-KGC (Yang et al., 2025)) have been successively proposed using entity and relation descriptions, along with PLMs, to accomplish the KGC task more effectively. These approaches, relying on encoder-only PLMs, often require task-specific fine-tuning and struggle to generalize across diverse scenarios.

2.3 Generation-based KGC methods

Generation-based methods treat KGC as a text generation task. GenKGC (Xie et al., 2022) converts KGC to a sequence-to-sequence generation task. Besides, KGT5 (Saxena et al., 2022) designs a unified framework for KGC and question answering but discards the pre-trained weights and trains T5 from scratch. KG-S2S (Chen et al., 2022) unifies input formats across various types of KGs, enabling it to tackle static, temporal, and few-shot KGC tasks. Recently, KG-LLM (Yao et al., 2023) utilizes descriptions of entities and relations as prompts to generate missing entities. KICGPT (Wei et al., 2023) employs an in-context learning strategy to guide ChatGPT to rerank candidate entities through multi-turn interactions. Although existing generation-based methods have achieved promising performance, their heavy reliance on the opaque and parameterized knowledge of LLMs leads to hallucinations and errors, hindering the reliability of KGC tasks under the OWA.

3 Preliminary

Knowledge Graph (KG). A KG can be denoted as a set of triplets $\mathcal{G} = \{(h, r, t) \,|h, t \in \mathcal{E}, r \in \mathcal{R}\}$, where $\mathcal{E}$ and $\mathcal{R}$ denote the sets of entities and relations, respectively. Here, h and t are the head and tail entities, and r is the relation between them. In most KGs, the relation set $\mathcal{R}$ is relatively fixed (Yang et al., 2022; Jiang et al., 2024). Additionally, the $\mathcal{G}$ often includes textual attributes such as entity names, descriptions, and relation names. In this work, we focus on static KGs that integrate structural and textual information.

Knowledge Graph Completion (KGC). In the paper, we focus on the link prediction task. **Under the closed-world assumption (CWA)**, given a query $q = (h, r, ?)$, the goal is to predict the most likely entity as the answer entity t_a. If the fact $(h, r, t_a) \in \mathcal{G}$, it is considered a successful hit of the correct answer; otherwise, it is a failure.

Under the open-world assumption (OWA), a KG can be considered an observation or understanding of the world where there could be unknown but correct facts (Yang et al., 2022). Thus, the KGC under OWA aims to predict missing entities by incorporating external knowledge beyond the KG, which in turn expands the scope of observable knowledge. For a query $q = (h, r, ?)$, the above process is as follows:

$$t_a \sim \chi(\cdot \mid q, P, \mathcal{K}_{in}, \mathcal{K}_{ext}), \tag{1}$$

where χ is an LLM. P is the KGC task description under OWA. The $\mathcal{K}_{in}$ is the intrinsic knowledge about q from $\mathcal{G}$, and $\mathcal{K}_{ext}$ is the external knowledge beyond $\mathcal{G}$. Under OWA, even if the fact $(h, r, t_a) \notin \mathcal{G}$, it can still be considered a correct fact.

4 Methodology

4.1 Model Architecture

We propose MusKGC, a novel multi-source knowledge enhancement framework for KGC via LLMs under OWA. As illustrated in Figure 2, MusKGC consists of three key modules: **(1) Relation-Template Induction with Type Constraint module** automatically induces relation templates from the KG, converting structured triplets into natural language sentences to improve the understanding and adaptability of LLMs to structured knowledge. **(2) Multi-source Knowledge Enhancement module** contains intrinsic KG facts and open knowledge from external reliable data sources, reducing hallucinations and enhancing the reliability of generated entities. **(3) Factuality and Consistency Evaluation module** assesses the generated unseen entities under the OWA, ensuring candidate entities are considered correct even if not in the KG, thereby aligning with KGC tasks in open-world scenarios.

4.2 Relation-Template Induction with Type Constraint

To effectively convert structured triplets into more understandable natural language sentences by LLMs, we design an automatic relation-template induction method with type constraints. Existing works (Lv et al., 2022; Wei et al., 2023) face two main issues: (1) relying on manually constructed

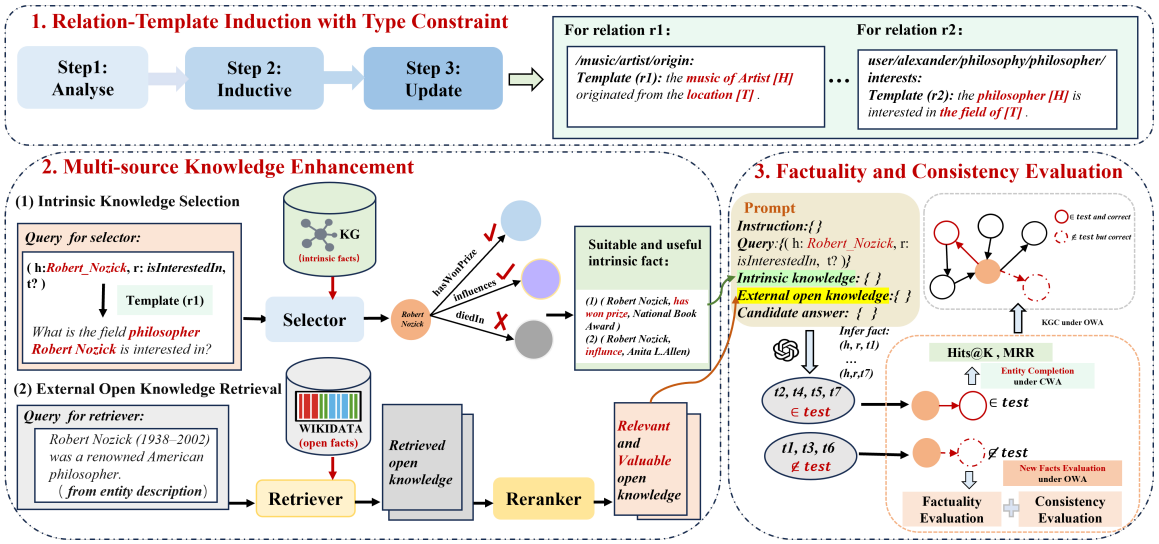


Figure 2: The architecture of our proposed MusKGC.

templates, which require costly expert effort; (2) neglecting the entity type information linked by relations, which leads to poor performance.

Therefore, to address these issues, we leverage the Chain-of-Thought (CoT) reasoning ability (Wei et al., 2022) of LLMs to automatically induce relation templates with entity type constraints. This process consists of the following steps:

- *Step 1*: Analyze the semantics of structured triplets and describe them in natural language.
- *Step 2*: Induce the relation template $\mathcal{T}_r'$ incorporating entity type information to better align with the triplet structure.
- *Step 3*: Update the $\mathcal{T}_r'$ with new triplets to enhance the adaptability of the generated relation template. We denote the updated relation template as $\mathcal{T}_r$.

Based on the above steps, we construct a set of relation templates $\mathcal{T}_r = \{(\mathcal{T}_{r_1}, \mathcal{T}_{r_2}..., \mathcal{T}_{r_m}) \mid r_i \in \mathcal{R}\}$. Given a typical example (as shown in Figure 2), we get the $\mathcal{T}_{r_i}$ is *the music of Artist [H] originated from the location [T]* when the r_i is */music/artist/origin*, where *[H]* and *[T]* mean the head and tail entities respectively. In our relation-template, we limit the head entity type linked to relation r_i to *Artist* and the tail entity type is *Location*. Detailed prompts are placed in Appendix A.

4.3 Multi-source Knowledge Enhancement

To expand the knowledge scope, we design a multi-source knowledge enhancement module that first selects relevant intrinsic neighboring knowledge from the KG and then retrieves key knowledge from external sources to assist in KGC under OWA.

4.3.1 Intrinsic Knowledge Selection

To select the most relevant triplets from the KG for the query q, we design a semantic similarity-driven intrinsic knowledge selection method. The motivation is that (1) directly remaining all neighboring triplets often introduces irrelevant or redundant information, and (2) in large-scale KGs, incorporating extensive neighboring triplets greatly increases prompt length, leading to diminished model performance and elevated computational costs.

To alleviate these obstacles, we first aggregate all neighboring triplets of the query $q = (h_q, r_q, ?)$ as $\mathcal{N}_q = \{(h, r, t) \mid h = h_q \text{ or } t = h_q\}$. Each triplet is then converted into a natural language sentence using the induced relation templates, forming a corresponding sentence set $\mathcal{N}_{q_s} = \{s_1, s_2, ..., s_n\}$, where $|\mathcal{N}_q| = |\mathcal{N}_{q_s}|$. Next, we leverage the pre-trained Sentence-Bert[1] (Reimers and Gurevych, 2019) to encode query q and each neighboring sentence s_i into a common representation space, obtaining embedding vectors $\mathbf{q}$ and $\mathbf{s}_i$, respectively. By computing the semantic similarity between $\mathbf{q}$ and each $\mathbf{s}_i$, we rank the neighboring facts and select the top-k as effective intrinsic knowledge from the KG. This ensures that the selected knowledge is semantically closely aligned with the query q.

[1]https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

The above process can be formulated as follows:

$$\mathbf{q}, \mathbf{s}_i = \text{Sentence-Bert}(q, s_i), \quad (2)$$

$$\mathcal{K}_{in} = \{s_{i1}, ..., s_{ik}\} = \text{Top-}k\ \{\ \frac{\mathbf{q} \cdot \mathbf{s}_i}{\|\mathbf{q}\|\|\mathbf{s}_i\|}\ \}, \quad (3)$$

where $\frac{\mathbf{q} \cdot \mathbf{s}_i}{\|\mathbf{q}\|\|\mathbf{s}_i\|}$ is the semantic similarity score between the $\mathbf{q}$ and the $\mathbf{s}_i$ via the cosine similarity function. We view the $\mathcal{K}_{in}$ as the top-k relevant intrinsic knowledge from the KG.

4.3.2 External Open Knowledge Retrieval

To mitigate the challenges of hallucinations and knowledge opacity inherent in generation-based LLMs for KGC under the OWA , inspired by RAG (Huang and Huang, 2024), we design an external knowledge retrieval module.

Specifically, given the initial query $q = (h, r, ?)$, we utilize the relation template to convert it into a natural language question s_q. We then construct a retrieval query q^{re} by concatenating s_q with the description of the head entity h, enabling accurate and comprehensive external knowledge retrieval. To retrieve relevant paragraphs from the external Wikipedia database, we employ the off-the-shelf Contriever-MS MARCO (Asai et al., 2024), denoted as the retriever $\mathcal{R}_{retriever}$. Subsequently, we get an initial set of retrieved open knowledge, denoted as $\mathcal{K}'_{ext}$. These processes can be formalized as $\mathcal{K}'_{ext} = \mathcal{R}_{retriever}(q^{re}, Wiki)$.

Next, inspired by (Huang and Huang, 2024), we introduce a reranker $\mathcal{R}_{rank}$ to perform deep reranking, ensuring the high relevance of external knowledge to the query. Specifically, we leverage the latest BGE M3 model (Chen et al., 2024a) to calculate relevance scores and select the top-n paragraphs. This process is formalized as $\mathcal{K}_{ext} = \text{Top-}n\ \{\mathcal{R}_{rank}(\mathcal{K}'_{ext}, q)\}$. This ensures that the retrieved external knowledge is relevant and valuable for the query q. Notably, our framework is not limited to the Wikipedia corpus and flexibly supports offline and online external sources, making it adaptable to domain-specific KGs.

4.3.3 Target Entity Generation

We aim to generate the answer entity t_a using an LLM-based generator by designing multi-source knowledge-enhanced prompts. Additionally, following the existing work (Wei et al., 2023; Wang et al., 2024), we conduct a generative re-ranking approach and ask the LLM to re-rank the top-m candidate entities, denoted as $\mathcal{A}$:

$$\mathcal{A} = LLM(P, q, \mathcal{K}_{in}, \mathcal{K}_{ext}, C), \quad (4)$$

where P is the KGC task description under OWA. C denotes initial candidates from the initial KGE Model $\mathcal{M}_e$. We place the details in Appendix D.

4.4 FC Evaluation for New Entity under OWA

The main difference between KGC evaluation under the CWA and OWA is that a predicted entity t_a can be correct even if it is not in the test set labels. Therefore, we design the FC evaluation strategy tailored to the OWA setting. For a generated fact $\mathcal{T}_{new} = (h, r, t_a)$ of t_a, we evaluate it from two aspects: **(1) Factuality:** assesses whether the facts described by $\mathcal{T}_{new}$ are correct; **(2) Consistency:** examines whether the facts described by $\mathcal{T}_{new}$ are consistent and compatible with the existing KG.

4.4.1 Factuality Evaluation

To ensure the factual correctness of the new fact $\mathcal{T}_{new}$, we evaluate it from the perspective of evidence support. Inspired by the previous work (Hu et al., 2024), we utilize an LM-based factuality checker to evaluate the correctness of new facts.

Specifically, based on the retrieved external knowledge, we define the factuality evaluation as the probability that the $\mathcal{T}_{new}$ is entailed by the retrieved knowledge. Thus, we construct knowledge-triplet pairs (k_i, $\mathcal{T}_{new}$), where $k_i \in \mathcal{K}_{ext}$. We apply an LM-based factuality checker [2], denoted as $Checker(\cdot)$, to calculate the likelihood of entailment for each knowledge-triplet pair. Finally, we use the average likelihood of entailment across all evidence-triplet pairs as the factuality score. It can be formalized as follows:

$$score_{fact}(\mathcal{T}_{new}) = \frac{1}{n}\Sigma_{k_i \in \mathcal{K}_{ext}} Checker(k_i, \mathcal{T}_{new}), \quad (5)$$

where n denotes the number of external knowledge items retrieved by the External Open Knowledge Retrieval module (Section 4.3.2).

4.4.2 Consistency Evaluation

To verify the compatibility and consistency between the newly generated entity and existing KG data, we design a consistency evaluation.

In detail, we employ the pre-trained text-based KGE model $\mathcal{M}_e$ to calculate the score of all newly generated triplets. We view these scores as the degree of consistency between the new facts $\mathcal{T}_{new} = (h, r, t_a)$ and KG intrinsic knowledge. This process can be formalized as follows:

$$score_{cons}(\mathcal{T}_{new}) = \mathcal{M}_e(\mathcal{T}_{new}). \quad (6)$$

[2] https://github.com/amazon-science/RefChecker

The larger the $score_{cons}(\mathcal{T}_{new})$, the higher the consistency with the existing KG.

5 Experiments

To comprehensively validate the performance of our MusKGC, we design a two-stage experiment: (1) KGC under the CWA: We selected SOTA baselines for comparison to verify the performance of MusKGC on known entities. (2) KGC under the OWA: Since MusKGC leverages generation-based LLMs and incorporates external knowledge to assist KGC, it may generate new facts. To address this, we further design an additional experiment to demonstrate the effectiveness of our proposed FC evaluation method in automatically assessing new facts, thereby bridging the gap between KGC in closed-world and open-world scenarios.

5.1 Datasets and Compared Methods

Datasets. We evaluate two benchmark datasets: FB15K-237 (Toutanova et al., 2015) and WN18RR (Dettmers et al., 2018). We place comprehensive dataset statistics in Appendix B.

Compared Methods. We compare our proposed MusKGC with a number of structure-based, text-based, and generation-based baselines. Among them, KG-FIT (Jiang et al., 2024) is the latest method to introduce open-world knowledge from LLMs. PKGC(Lv et al., 2022) utilizes the parameterized knowledge as open-world knowledge and uses manual evaluation to verify the correctness of new facts under OWA. The descriptions of baselines are presented in Appendix C.

5.2 Experimental Setup

Implementation Details. Our MusKGC can support various retrievers and LLMs without fine-tuning. For the initial KGE model $\mathcal{M}_e$, we use SimKGC with the same hyperparameter settings as in (Wang et al., 2022a), though $\mathcal{M}_e$ can be replaced by any suitable KGC model. For the intrinsic knowledge selection, we set $k = 8$. For external open knowledge retrieval, we use the off-the-shelf Contriever-MS MARCO (Asai et al., 2024) as the retriever and BGE M3 (Chen et al., 2024a) as the reranker. Then we set $n = 2$ to remain the top-2 relevant paragraphs. In target entity generation, we generate $m = 20$ candidate entities. For the ChatGPT API, we adopt GPT-4o-mini because of its flexibility and shorter API call time. We set temperature, presence_penalty, and frequency_penalty to 0, and top_p to 1 to avoid randomness.

Metrics. Given a positive test triple (h, r, t) , we convert it into a query $(h, r, ?)$ or $(?, r, t)$. Our evaluation includes two parts: (1) We use two widely used metrics (MRR and Hits@$k(k = 1, 3, 10)$) to assess the prediction performance on known test set entities. (2) Under the OWA, since the correctness of new facts is unknown, we construct a gold standard via human annotations. Based on this, we compute F1 and Recall scores by comparing the MusKGC-FC evaluation with the human annotations. Details on the annotation criteria and evaluation process are provided in Appendix F.

5.3 Main Experimental Results

We show the performance comparison in Table 1, where our MusKGC outperforms existing methods in most cases, particularly on Hits@1. Specifically, (1) structure-based methods generally exhibit lower performance, indicating that structural information alone is insufficient. While KG-FIT integrates open-world knowledge via LLMs, the reliability and interpretability of implicit Parametric knowledge within LLMs remain concerns. (2) Text-based methods rely solely on semantic descriptions of entities and relations and require fine-tuning on specific datasets, which is not only time-consuming but also yields suboptimal results; (3) Generation-based methods benefit from LLMs language understanding capability but suffer from hallucinations when relying solely on LLMs. In contrast, our MusKGC alleviates this issue by incorporating reliable external knowledge.

5.4 Ablation Study

To verify the effectiveness of each module, we conduct an ablation study in Table 2. We can observe: (1) Removing the relation template can lead to performance degradation. This suggests that relation templates provide a "language bridge", enabling LLMs to more effectively understand and process structured knowledge; (2) Removing the intrinsic knowledge selection also reduces performance. We conclude that the module can minimize irrelevant facts and reduce the input of noisy data, thereby ensuring the correctness of the results. (3) Removing the external knowledge retrieval module also showed varying degrees of decline. On the one hand, this confirms that external knowledge enhances the understanding of KG semantics. On the other hand, it shows that external non-parametric knowledge helps mitigate the hallucination in LLMs to improve performance.

Model	FB15K-237				WN18RR			
	MRR↑	Hits@1↑	Hits@3↑	Hits@10↑	MRR↑	Hits@1↑	Hits@3↑	Hits@10↑
Structure-based methods								
RESCAL (Nickel et al., 2011) ‡	0.356	0.266	0.390	0.535	0.467	0.439	0.478	0.516
TransE (Bordes et al., 2013) †	0.279	0.198	0.376	0.441	0.243	0.043	0.441	0.532
DistMult (Yang et al., 2015) †	0.241	0.155	0.263	0.419	0.430	0.390	0.440	0.490
ComplEx (Trouillon et al., 2016) †	0.247	0.158	0.275	0.428	0.440	0.410	0.460	0.510
RotatE (Sun et al., 2019b)	0.338	0.241	0.375	0.533	0.476	0.428	0.492	0.571
TuckER (Balazevic et al., 2019)	0.358	0.266	0.394	0.544	0.470	0.443	0.482	0.526
HAKE (Zhang et al., 2020a)	0.346	0.250	0.381	0.542	0.497	0.452	0.516	0.582
CompGCN (Vashishth et al., 2020)	0.355	0.264	0.390	0.535	0.479	0.443	0.494	0.546
pro_CBR (Das et al., 2020)	-	-	-	-	0.480	0.430	0.490	0.550
HittER (Chen et al., 2021)	0.344	0.246	0.380	0.535	0.496	0.449	0.514	0.586
KG-FIT (Jiang et al., 2024)	0.362	0.275	-	0.572	0.553	0.488	-	0.695
Text-based methods								
Pretrain-KGE (Zhang et al., 2020b)	0.332	-	-	0.529	0.235	-	-	0.557
KG-BERT (Yao et al., 2019) †	-	-	-	0.420	0.216	0.041	0.302	0.524
StAR (Wang et al., 2021) †	0.263	0.171	0.287	0.452	0.364	0.222	0.436	0.647
SimKGC (Choi et al., 2021)	0.338	0.252	0.364	0.511	0.671	0.595	0.719	0.802
PKGC(Lv et al., 2022)	0.381	0.192	0.308	0.476	0.422	0.351	0.484	0.537
MPIKGC (Xu et al., 2024)	0.327	0.241	0.354	0.497	0.656	0.571	0.712	0.803
Generation-based methods								
GenKGC (Xie et al., 2022)§	-	0.192	0.355	0.439	-	0.287	0.403	0.535
KGT5 (Saxena et al., 2022)§	0.276	0.210	-	0.414	0.508	0.487	-	0.544
KG-S2S (Chen et al., 2022)§	0.336	0.257	0.373	0.498	0.574	0.531	0.595	0.661
ChatGPT$_{zero-shot}$ (Zhu et al., 2024)	-	0.237	-	-	-	0.190	-	-
ChatGPT$_{one-shot}$ (Zhu et al., 2024)	-	0.267	-	-	-	0.212	-	-
KICGPT (Wei et al., 2023)	0.412	0.327	0.448	0.554	0.549	0.474	0.585	0.641
GS-KGC (Yang et al., 2025)	-	0.280	0.426	-	-	0.346	0.516	-
MusKGC	**0.443**	**0.393**	**0.462**	**0.577**	**0.673**	**0.620**	**0.721**	**0.816**
$\Delta\%$	↑7.5	↑20.2	↑3.1	↑0.9	↑0.3	↑4.2	↑2.8	↑1.6

Table 1: Comparison between our MusKGC and baseline methods. We reproduce the results of PKGC using their source code in two datasets. †: results are from (Wang et al., 2021). ‡: results are from (Chen et al., 2021). §: results are from (Liu et al., 2024b). Other results are taken from their original papers. $\Delta\%$ is the percentage difference between the best and suboptimal result.

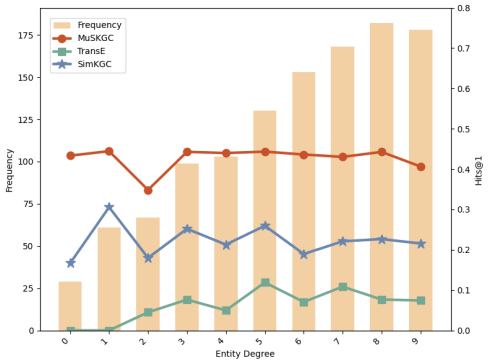
Models	FB15K-237				WN18RR			
	MRR↑	Hits@1↑	Hits@3↑	Hits@10↑	MRR↑	Hits@1↑	Hits@3↑	Hits@10↑
MusKGC	**0.443**	**0.393**	**0.462**	**0.577**	**0.673**	**0.620**	**0.721**	**0.816**
w/o Relation Template	0.425	0.378	0.441	0.531	0.652	0.602	0.666	0.770
w/o Intrinsic Knowledge Selection	0.435	0.385	0.456	0.544	0.664	0.615	0.682	0.776
w/o External Open Knowledge	0.431	0.385	0.447	0.536	0.667	0.617	0.687	0.782

Table 2: Results of ablation study.

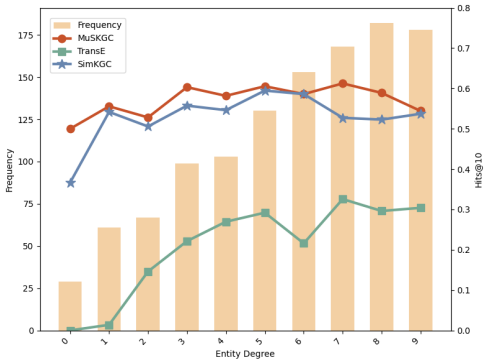
5.5 Further Analysis on Long-Tail Entities

To further validate the link prediction performance of the model in open-world scenarios, we treat long-tailed scenarios as approximations of the open world and conduct fine-grained analysis. Specifically, we follow existing research (Wang et al., 2022b) to group entities based on their degree in the KG, with lower-degree entities being more likely to be long-tail entities. In our paper, we focus on long-tailed entities, defined as those with a degree of less than 10, as shown in Figure 3.

By comparing the performance differences of triplet-based (TransE), text-based (SimKGC), and our MusKGC in predicting long-tailed entities, we have the following insights: (1) Our MusKGC performs much better in predicting long-tailed entities. By incorporating external knowledge, we enrich the semantic information of these entities, which are underrepresented in the training data. This leads to a significant improvement in prediction performance, especially for the hits@1 metric (as in Figure 3(a)). (2) For degree-0 entities, they com-



(a) Comparison of Hits@1



(b) Comparison of Hits@10

Figure 3: Results for the long-tail entities on FB15K-237 datasets.

pletely absent from the training set, the advantage of our MusKGC is even more pronounced. Unlike baseline models such as the TransE model, which rely solely on relations learned from the training data, our framework demonstrates that external knowledge is crucial to predict these new entities that are not present in the training datasets.

5.6 FC Evaluation Analysis under OWA

To verify the effectiveness of our FC evaluation strategy, we manually construct annotated labels as the gold standard. We then choose GPT-4o as a comparative evaluation strategy, relying solely on the LLM's intrinsic knowledge to assess the correctness of new facts (which is also used in GS-KGC (Yang et al., 2025) for evaluating new facts under OWA). Please refer to the Appendix E for specific implementation details.

Metric analysis on different evaluation strategies. From Table 3, we can find that our MusKGC-FC evaluation strategy achieves a higher recall and slightly better F1 score. This indicates that, compared to GPT-4o's direct evaluation, our MusKGC-FC evaluation strategy more closely aligns with

Models	Recall	F1 score
GPT-4o evaluation	0.654	0.768
MusKGC-FC evaluation	0.724	0.783

Table 3: Results of GPT-4o direct evaluation and our MusKGC-FC evaluation strategy (human annotated labels as the gold standard).

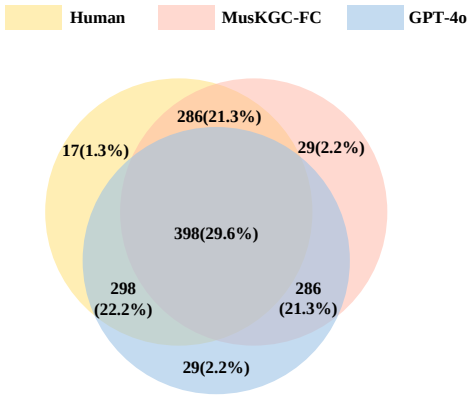


Figure 4: The Venn diagram of different evaluations.

human evaluation results. This improvement further highlights the effectiveness and reliability of our MusKGC-FC evaluation, while also avoiding the cost issues associated with using GPT-4o.

The Venn diagram analysis of different evaluation strategies. We calculate the prediction overlap among Human, MusKGC-FC, and GPT-4o evaluations. As shown in Figure 4. We can observe the following insights: (1) MusKGC-FC evaluation demonstrates strong alignment with human evaluation standards. Under OWA, the discrepancy between MusKGC-FC and GPT-4o evaluations is minimal, showcasing its reliability in emulating human judgment. (2) MusKGC-FC exhibits a degree of "innovative capability." Specifically, it identifies 29 unique correct entities (2.2%) that are not captured by the other methods. This highlights its potential to discover diverse and correct knowledge, contributing to the expansion of KG.

5.7 Case Study

We present representative cases in Table 4 to demonstrate the advantages of our MusKGC in evaluating new facts. **In case 1,** for the query= (?, */film/film/executive_produced_by*, *Laurence Mark*), MusKGC predicts "Dreamgirls" (different from the label), which is annotated as "Correct" by humans. However, LLM-based evaluation incorrectly labels it as "Wrong" due to hallucinations. In contrast, our FC evaluation strat-

	Case 1	Case 2
Facts	**Query:** (?, /film/film/executive_produced_by, Laurence Mark) **Label:** Working Girl	**Query:** (1980 Summer Olympics, /time/event/locations, ?) **Label:** Moscow
Top-1 Candidate	Dreamgirls	Moscow
Human Annotation	Correct	Correct
GPT Response	To evaluate the correctness of the sentence, let's break it down:... 1. Role of Laurence Mark: **Laurence Mark is a well-known producer who has worked on several notable films, including Dreamgirls**,...However, **Laurence Mark is credited not specifically as an executive producer, for Dreamgirls.**...Based on this analysis, ... The sentence was judged as **Wrong**.	To evaluate the correctness of the sentence, let's break it down: 1. Event and Year: ...The 1980 Summer Olympics did indeed occur in that year. 2.Location: **the Olympics took place in Moscow.**...Based on these points, **the sentence contains a grammatical error with the use of the before Moscow and an unnecessary period in the end**...The sentence was judged as **Wrong**.
GPT Output	Wrong	Wrong
Retrieval knowledge	[**Laurence Mark is a Golden Globe-winning executive producer of such films as ..., Dreamgirls (2006)**, I, Robot (2004) ...; Jerry Maguire is a 1996 American romantic ...]	[**The 1980 Summer Olympics, ..., were held in Moscow**, ...; **The 1980 Summer Olympics, ..., were an international multi-sport event held in Moscow**, Soviet Union from 19 July to 3 August...]
MusKGC Output	Correct	Correct

Table 4: Case Study. Two cases are demonstrated to explain the advantages of MusKGC.

egy successfully verifies the rationality of candidates from the factuality and consistency perspectives. **In case 2,** for query=(*1980 Summer Olympics*, */time/event/locations*, ?), MusKGC predicts "Moscow" (same as label), and the human annotation is "Correct". When we directly use GPT-4o for evaluation, the fact is incorrectly judged as "Wrong". However, through MusKGC-FC evaluation, the fact is correctly judged as "correct". This demonstrates that our proposed MusKGC-FC evaluation method ensures that new facts are grounded in evidence and consistent with existing KG. Additional cases are provided in Appendix J.

6 Conclusion

In the paper, we propose a novel and flexible **Mu**lti-source knowledge enhancement framework for **KGC** under the open-world assumption with LLM, named **MusKGC**. Our method enhances LLMs' understanding of structured knowledge and enables accurate, evidence-based missing entities generation via an automatic relation template induction module and a multi-source knowledge enhancement module. To address the challenge of assessing unseen entities under the OWA, we introduce a novel evaluation strategy that assesses the factuality and consistency of generated facts, ensuring uniform evaluation regardless of whether entities appear in the original dataset. This effectively expands the KG and aligns with OWA requirements. Extensive experiments show that our model achieves SOTA performance across benchmarks, and the proposed evaluation strategy effectively assesses new facts under OWA. Further experiments on domain-specific datasets demonstrate the flexibility and generalizability of MusKGC.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments and helpful suggestions. This work was supported by the National Natural Science Foundation of China (No.62302507).

Limitations

We identify that there may be some possible limitations in this study. First, to ensure a fair comparison with existing methods, we adopted a generation-based re-ranking approach for KGC under OWA. This method relies on the number of candidate entities selected in the re-ranking stage to some extent. Second, our work focuses on a single modality (text-based KGC) while neglecting other modalities (e.g., images, audio). These additional modalities, often studied in the context of multi-modal KGs, represent an important research direction. To this end, we further explore multi-modal integration strategies to enhance MusKGC's capabilities.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ivana Balazevic, Carl Allen, and Timothy M. 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 2019, Hong Kong, China, November 3-7, 2019, pages 5184–5193. Association for Computational Linguistics.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

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, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 4005–4017. International Committee on Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *CoRR*, abs/2402.03216.

Sanxing Chen, Xiaodong Liu, Jianfeng Gao, Jian Jiao, Ruofei Zhang, and Yangfeng Ji. 2021. Hitter: Hierarchical transformers for knowledge graph embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10395–10407. Association for Computational Linguistics.

Zhongwu Chen, Long Bai, Zixuan Li, Zhen Huang, Xiaolong Jin, and Yong Dou. 2024b. A new pipeline for knowledge graph reasoning enhanced by large language models without fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1366–1381. Association for Computational Linguistics.

Bonggeun Choi, Daesik Jang, and Youngjoong Ko. 2021. MEM-KGC: masked entity model for knowledge graph completion with pre-trained language model. *IEEE Access*, 9:132025–132032.

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 *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Rajarshi Das, Ameya Godbole, Nicholas Monath, Manzil Zaheer, and Andrew McCallum. 2020. Probabilistic case-based reasoning for open-world knowledge graph completion. In *Findings of EMNLP*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1811–1818. AAAI Press.

Cong Fu, Tong Chen, Meng Qu, Woojeong Jin, and Xiang Ren. 2019a. Collaborative policy learning for open knowledge graph reasoning. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2672–2681. Association for Computational Linguistics.

Cong Fu, Tong Chen, Meng Qu, Woojeong Jin, and Xiang Ren. 2019b. Collaborative policy learning for open knowledge graph reasoning. *arXiv preprint arXiv:1909.00230*.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *CoRR*, abs/2405.14486.

Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y. Chang. 2018. Improving sequential recommendation with knowledge-enhanced memory networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 505–514. ACM.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *CoRR*, abs/2404.10981.

Pengcheng Jiang, Shivam Agarwal, Bowen Jin, Xuan Wang, Jimeng Sun, and Jiawei Han. 2023. Text-augmented open knowledge graph completion via pre-trained language models. *arXiv preprint arXiv:2305.15597*.

Pengcheng Jiang, Lang Cao, Cao Xiao, Parminder Bhatia, Jimeng Sun, and Jiawei Han. 2024. KG-FIT: knowledge graph fine-tuning upon open-world knowledge. *CoRR*, abs/2405.16412.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Few-shot reranking for multi-hop QA via language model

prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15882–15897. Association for Computational Linguistics.

Bosung Kim, Taesuk Hong, Youngjoong Ko, and Jungyun Seo. 2020. Multi-task learning for knowledge graph completion with pre-trained language models. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1737–1743. International Committee on Computational Linguistics.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2181–2187. AAAI Press.

Yang Liu, Xiaobin Tian, Zequn Sun, and Wei Hu. 2024a. Finetuning generative large language models with discrimination instructions for knowledge graph completion. In *The Semantic Web - ISWC 2024 - 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11-15, 2024, Proceedings, Part I*, volume 15231 of *Lecture Notes in Computer Science*, pages 199–217. Springer.

Yang Liu, Xiaobin Tian, Zequn Sun, and Wei Hu. 2024b. Finetuning generative large language models with discrimination instructions for knowledge graph completion. *CoRR*, abs/2407.16127.

Xin Lv, Yankai Lin, Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. Do pre-trained models benefit knowledge graph completion? A reliable evaluation and a reasonable approach. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3570–3581. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 809–816. Omnipress.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2814–2828. Association for Computational Linguistics.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer.

Haitian Sun, Tania Bedrax-Weiss, and William W. Cohen. 2019a. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2380–2390. Association for Computational Linguistics.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019b. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1499–1509. The Association for Computational Linguistics.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. 2020. Composition-based multi-relational graph convolutional networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. 2021. Structure-augmented text representation learning for efficient knowledge graph completion. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1737–1748. ACM / IW3C2.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022a. Simkgc: Simple contrastive knowledge graph completion with pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4281–4294. Association for Computational Linguistics.

Xintao Wang, Qianyu He, Jiaqing Liang, and Yanghua Xiao. 2022b. Language models as knowledge embeddings. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 2291–2297. ijcai.org.

Yilin Wang, Minghao Hu, Zhen Huang, Dongsheng Li, Dong Yang, and Xicheng Lu. 2024. Kc-genre: A knowledge-constrained generative re-ranking method based on large language models for knowledge graph completion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 9668–9680. ELRA and ICCL.

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, July 27 -31, 2014, Québec City, Québec, Canada*, pages 1112–1119. AAAI Press.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yanbin Wei, Qiushi Huang, Yu Zhang, and James T. Kwok. 2023. KICGPT: large language model with knowledge in context for knowledge graph completion. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8667–8683. 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 of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, pages 162–165. ACM.

Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. RC-NET: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 1219–1228. ACM.

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, 20-25 May, 2024, Torino, Italy*, pages 11956–11968. ELRA and ICCL.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Haotong Yang, Zhouchen Lin, and Muhan Zhang. 2022. Rethinking knowledge graph evaluation under the open-world assumption. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Rui Yang, Jiahao Zhu, Jianping Man, Li Fang, and Yi Zhou. 2024. Enhancing text-based knowledge graph completion with zero-shot large language models: A focus on semantic enhancement. *Knowl. Based Syst.*, 300:112155.

Rui Yang, Jiahao Zhu, Jianping Man, Hongze Liu, Li Fang, and Yi Zhou. 2025. Gs-kgc: A generative subgraph-based framework for knowledge graph completion with large language models. *Information Fusion*, 117:102868.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *CoRR*, abs/1909.03193.

Liang Yao, Jiazhen Peng, Chengsheng Mao, and Yuan Luo. 2023. Exploring large language models for knowledge graph completion. *CoRR*, abs/2308.13916.

Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. 2020a. Learning hierarchy-aware knowledge graph embeddings for link prediction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3065–3072. AAAI Press.

Zhiyuan Zhang, Xiaoqian Liu, Yi Zhang, Qi Su, Xu Sun, and Bin He. 2020b. Pretrain-kge: Learning knowledge representation from pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 259–266. Association for Computational Linguistics.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: recent capabilities and future opportunities. *World Wide Web (WWW)*, 27(5):58.

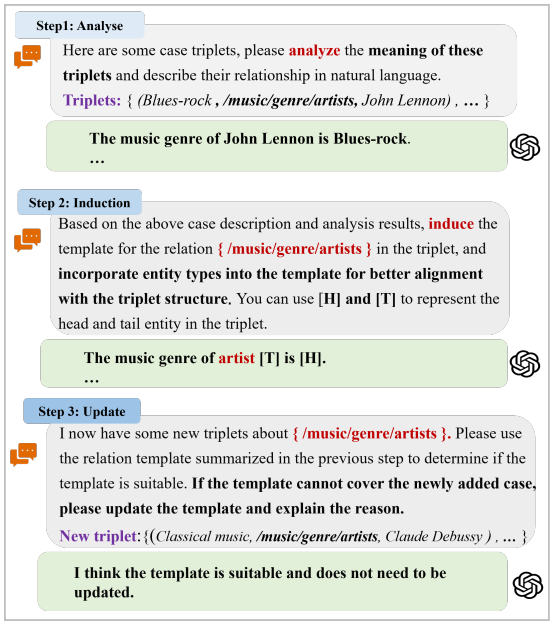


Figure 5: Overview of Relation-Template Induction with Type Constraint. It conducts analysis, induction, and update step by step to obtain relation templates automatically.

A Prompts for Relation-Template induction with Type Constraint

The detailed conduct process and prompts are shown in Figure 5. For example, for the relation */music/genre/artists*, we get the relation template *The music genre of artist [T] is [H].*

B Datasets Details

We choose two benchmark datasets for evaluation: FB15K-237 (Toutanova et al., 2015) and WN18RR (Dettmers et al., 2018). Statistics of the datasets are summarized in Table 5 (Bordes et al., 2013) proposed the WN18 and FB15K datasets. Due to the test set leakage problem, (Toutanova et al., 2015; Dettmers et al., 2018) improved these datasets by removing reverse relations, releasing FB15k-237 and WN18RR, respectively. The WN18RR dataset comprises $\sim 41k$ synsets and 11 relations derived from WordNet(Miller, 1995), focusing on the relationships between words. The FB15K-237 dataset consists of $\sim 15k$ entities and 237 relations from Freebase (Bollacker et al., 2008), covering topics such as movies, sports, awards, and traveling. The statistics of the two datasets are shown in Table 5.

Table 5: Statistics of datasets we use.

dataset	#entity	#relation	#train	#valid	#test
FB15K-237	14,541	237	272,115	17,535	20,466
WN18RR	40,943	11	86,835	3,034	3,134

C Baselines description

We compared our MusKGC with three categories of SOTA methods:

(1) Structure-based KGC methods.. We select eleven classical structural based models, including RESCAL (Nickel et al., 2011), TransE (Bordes et al., 2013), DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), RotatE (Sun et al., 2019b), TuckER (Balazevic et al., 2019), HAKE (Zhang et al., 2020a), CompGCN (Vashishth et al., 2020), pro_CBR (Das et al., 2020), HittER (Chen et al., 2021), and KG-FIT (Jiang et al., 2024). Among them, pro_CBR (Das et al., 2020) models uncertainty by integrating probabilistic graphical models with case-based reasoning for open-world KGC. KG-FIT (Jiang et al., 2024) introduces a fine-tuning framework for KGs that dynamically adapts to open-world knowledge by integrating external information sources.

(2) Text-based KGC methods. We select six text-based models as the competitors, namely KG-BERT (Yao et al., 2019), Pretrain-KGE (Zhang et al., 2020b), StAR (Wang et al., 2021), SimKGC (Wang et al., 2022a), PKGC (Lv et al., 2022), and MPIKGC (Xu et al., 2024). KG-BERT (Yao et al., 2019) introduces entity and relation descriptions and utilizes Bert for KGC for the first time. Pretrain-KGE (Zhang et al., 2020b) uses the external parameter knowledge of the PLM to enrich the knowledge representation. StAR(Wang et al., 2021) achieves structure-enhanced KGC by combining text and structure knowledge. SimKGC(Wang et al., 2022a) introduces contrastive learning into text-based methods for the first time, using three negative sampling strategies. MPIKGC (Xu et al., 2024) represents a cutting-edge text-based approach. It compensates for the deficiency of contextualized knowledge and improves KGC by querying LLMs from various perspectives. PKGC(Lv et al., 2022) utilizes the parameterized knowledge as open-world knowledge and uses manual evaluation to verify the correctness of new facts under OWA.

(3) Generation-based methods. We select the following six generation-based KGC models, where GenKGC(Xie et al., 2022), KGT5(Saxena et al., 2022), and KG-S2S(Chen et al., 2022) are built on either BART or T5 model. Then, we choose three latest models based on ChatGPT models, including ChatGPT$_{one-shot}$, ChatGPT$_{few-shot}$, and KICGPT(Wei et al., 2023). KICGPT is the most similar work to our proposed method, but it only considers the parametric knowledge within LLM and lacks consideration for external non-parametric knowledge.

D Target Entity Generation

This sub-module aims to generate the answer entity using an LLM-based generator by designing a multi-source knowledge-enhanced prompt. A prompt of candidate entity re-ranking based on multi-source knowledge enhancement is shown in Figure 6. For the neighbor triplets from $\mathcal{K}_{in}$ and q, we apply the relation template $\mathcal{T}_r$ to convert them into natural language sentences, which are then fed into the LLM prompt. Additionally, Figure 7 illustrates an example of candidate entities re-ranking from the FB15K-237 dataset.

E FC Evaluation Details and Principles

Sample Principles. Specifically, we randomly select 1,000 queries from the test dataset for our FC evaluation experiment. To ensure fairness, our sampling criteria are as follows: (1) an equal number of queries are chosen for both the head entity predictions and the tail entity predictions; (2) a proportionate selection based on entity degree, ensuring that a diverse range of entity degrees, including long-tail and non-long-tail entities, are represented.

Factuality and Consistency Score. For each query $(h, r, ?)$, we concatenate the top-1 answer entity output by the LLM, denoted as t_a with the query to form a new fact (h, r, t_a). Please note that for (h, r, t_a), we first calculate the consistency score. We leverage the pre-trained text-based KGC model (SimKGC in our work) to calculate the consistency score of the newly generated facts. Entities with scores above the top-20 threshold are considered relevant to the existing KG. The knowledge expressed by these entities is deemed consistent with the existing KG knowledge. Subsequently, leveraging the external knowledge obtained during the retrieval phase, we compute the factuality score. Specifically, we employed the *roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli* model to perform factuality verification. This model effectively evaluates the logical relationship between two text segments, displaying confidence scores for three possible labels, including neutral and contradiction. In our work, we consider a new fact to be correct if and only if the entailment score is the highest.

	Model	Hits@1	Hits@5	Hits@10
KGE-based	TransE (Bordes et al., 2013)	-	40.83	53.62
	ComplEx (Trouillon et al., 2016)	-	34.54	49.30
	TuckER (Balazevic et al., 2019)	-	30.22	45.33
	RotatE (Sun et al., 2019b)	-	40.15	53.82
Text&KGE-based	RC-Net (Xu et al., 2014)	-	9.26	12.00
	TransE+Line (Fu et al., 2019a)	-	22.31	33.65
RL-based	MINERVA (Das et al., 2018)	-	57.63	63.83
	CPL (Fu et al., 2019b)	-	58.10	65.16
PLM-based	PKGC (Lv et al., 2022)	-	55.05	59.43
	TagReal (Jiang et al., 2023)	-	60.68	62.88
	Our MusKGC	**54.9**	**86.05**	**93.10**

Table 6: **Performance Results on UMLS dataset**. The highest score is highlighted in **bold.**

F Manual Annotation

In order to establish a referenceable evaluation standard, we manually annotated 1000 queries as the gold standard for evaluation under OWA. Specifically, we selected 5 volunteers between the ages of 20 and 30 with a background in KG. Among them, there are 3 males and 2 females. There are three PhD students, and two are master's students. We have created evaluation questionnaires, each consisting of 200 pairs of "text triplet lists" (input text from the model and corresponding output t_a). Volunteers evaluate the correctness of triplets.

Then we used human evaluation results as the fundamental facts for recall and F1 score.

G Domain-specific Experiments

We use the UMLS-PubMed dataset provided by (Fu et al., 2019b), where UMLS serves as a biomedical knowledge graph and PubMed as the accompanying textual corpus. This dataset focuses on the medical domain, encompassing a wide range of biomedical concepts and relations. We conduct our evaluation on the UMLS-PubMed dataset, following the CPL(Fu et al., 2019b) and TagReal(Jiang et al., 2023), which are two early benchmarks for open-world KGC, as shown in Table 6. Experimental results show that our MusKGC has good adaptability to domain knowledge.

Prompt of Candidate Entities Re-ranking based on Multi-source Knowledge Enhancement

Instruction: You are a good assistant to perform the entity re-rank task. Given a goal question and a list of candidate answers, your task is to **evaluate the likelihood of correctness for each answer and re-rank the list so that the most likely correct answer appears at the top**. Please consider the context of the question and the plausibility of each answer while making your decisions.

Knowledge from the KG:
{$\mathcal{K}_{in}$ from Intrinsic Knowledge Selection.}

Background knowledge from WIKI:
{$\mathcal{K}_{ext}$ from External Open Knowledge Retrieval.}

the goal question is: { q }

the candidate answer list is:
{ [$candidate_1$, ... (20 options for candidate entities) ... , $candidate_{20}$] }

Please re-rank the candidate answers based on the knowledge mentioned above and your own knowledge. The output format is strictly in accordance with < The list of sorted candidate answers is [$answer_1$ | $answer_2$ | ... | $answer_{20}$] >

Output results:

Figure 6: Prompt of Candidate Entities Re-ranking.

H Computational Overhead

We conduct our experiments on an NVIDIA A800 80G GPU. To analyze the computational overhead of our MusKGC, we record the time consumption of each module for a single query, as presented in Table 8.

Taking WN18RR as an example, our MusKGC completes processing of the entire dataset in approximately $4h$, which is comparable to the strong baseline KICGPT (Wei et al., 2023), requiring around $4.5h$. This indicates only a marginal difference in computational time. Although MusKGC introduces additional components, including an external knowledge retrieval module and an FC evaluation mechanism, these do not significantly increase the overall computational load.

While the retrieval module requires some time for initial loading, subsequent queries are processed efficiently. The FC evaluation, which is essential for accurately validating newly inferred entities under open-world settings, also maintains a manageable computational overhead. Moreover, the modular design of MusKGC supports parallel processing on large-scale datasets, further improving efficiency. Therefore, the computational overhead introduced by our framework remains within practical limits and does not hinder its applicability in real-world scenarios.

datasets	Hits@10	Hits@15	Hits@20	Hits@25	Hits@30
WN18RR	80.22	82.66	84.21	85.55	86.86
FB15K-237	51.14	56.33	60.00	62.83	65.05

Table 7: Recall at rank m (Hits@m) when using simKGC for initial candidate ranking.

I Analysis of Candidate Entity Set Size in the KGE Model

Generative re-ranking methods have been widely adopted across various domains (Khalifa et al., 2023; Lv et al., 2022; Liu et al., 2024a; Wang et al., 2024). These approaches typically employ a lightweight KGE model to produce an initial ranking of candidate entities, which are subsequently re-ranked by LLMs.

In Table 7, we report the probability that the correct answer appears within the top-m candidates generated by the KGE model for different values of m, using Hits@k($k = 10, 15, 20, 25, 30$) as the evaluation metric. While increasing the candidate set size leads to higher Hits@k scores, the observed gains in recall diminish as m grows larger. More importantly, a larger candidate set substantially increases computational cost and LLM inference time. Therefore, following prior work (Chen et al., 2024b), we select the top-20 ($m = 20$) candidate entities for subsequent prediction.

Example of Candidate Entities Re-ranking based on Multi-source Knowledge Enhancement

Instruction: You are a good assistant to perform the entity re-rank task. Given a goal question and a list of candidate answers, your task is to **evaluate the likelihood of correctness for each answer and re-rank the list so that the most likely correct answer appears at the top**. Please consider the context of the question and the plausibility of each answer while making your decisions.

Knowledge from the KG:
{ ['Fran Walsh was nominated for Writers Guild of America Award for Best Adapted Screenplay.',
'Fran Walsh has the profession of Musician-GB.',
'Fran Walsh was nominated for Academy Award for Best Picture.',
'Fran Walsh is nominated alongside Annie Lennox for an award.',
'Fran Walsh is nominated alongside Peter Jackson for an award.',
'Fran Walsh was nominated for Academy Award for Best Original Screenplay.',
'Fran Walsh was nominated for Writers Guild of America Award for Best Original Screenplay.',
'Fran Walsh was nominated for Grammy Award for Best Song Written for a Motion Picture, Television or Other Visual Media.'] }

Background knowledge from WIKI:
{ ['Dame Frances Rosemary Walsh (born 10 January 1959) is a New Zealand screenwriter, film producer, and lyricist. The partner of filmmaker Peter Jackson, Walsh has contributed to all of his films since 1989: as co-writer since Meet the Feebles, and as producer since The Lord of the Rings Trilogy. She has won three Academy Awards for the final film of the trilogy, The Lord of the Rings: The Return of the King.',
'Sir Peter Robert Jackson (born 31 October 1961) is a New Zealand film director, screenwriter, and film producer. He is best known as the director, writer, and producer of the Lord of the Rings trilogy (2001–2013) and the Hobbit trilogy (2012–2014), both of which are adapted from the novels of the same name by J. R. R. Tolkien. Other notable films include the critically lauded drama Heavenly Creatures (1994), the horror comedy The Frighteners (1996), the epic monster remake film King Kong (2005), and the World War I documentary film They Shall Not Grow Old (2018). He is the third-highest-grossing film director of all-time, his films having made over $6.5 billion worldwide. Jackson began his career with the splatstick horror comedy Bad Taste (1987) and the black comedy Meet the Feebles (1989) before filming the zombie comedy Braindead (1992). He shared a nomination '] }

the goal question is: { Where is the nationality of Fran Walsh? }

the candidate answer list is:
{ ["Australia", "New Zealand", "South Africa", "United Kingdom", "India", "Canada", "Jamaica", "United States of America", "England", "Malaysia", "Saint Vincent and the Grenadines", "Republic of Ireland", "Tuvalu", "Bermuda", "Zimbabwe", "Cook Islands", "Papua New Guinea", "Cayman Islands", "Union of South Africa", "Grenada"] }

Please re-rank the candidate answers based on the knowledge mentioned above and your own knowledge. The output format is strictly in accordance with < The list of sorted candidate answers is [$answer_1$ | $answer_2$ | ... | $answer_{20}$] >

Output results:
The list of sorted candidate answers is ["New Zealand", "Australia", "United Kingdom", "England", "South Africa", "Canada", "United States of America", "India", "Jamaica", "Malaysia", "Republic of Ireland", "Zimbabwe", "Cook Islands", "Cayman Islands", "Grenada", "Tuvalu", "Papua New Guinea", "Union of South Africa", "Saint Vincent and the Grenadines", "Bermuda"]

Figure 7: An example of candidate Entities Re-ranking from the FB15K-237 dataset.

J Cases Study

We list additional cases to demonstrate the advantages of MusKGC as follows (Table 10 and Table 9).

Modules	Main Overhead	Time per Query	Others
Intrinsic Knowledge Selection	Computing the semantic similarity score	0.0465s	-
External Open Knowledge Retrieval	Initial loading: 185.60s	1.8s	After loading, query speed is fast with cache reuse
Target Entity Generation with prompt	GPT reasoning	2s	-
FC evaluation	-	1.5s	-

Table 8: The main computational overhead for a query.

	Case 2
Facts	**Query:** (?, /medicine/symptom/symptom_of, Typhoid fever) **Label:** Abdominal pain
Top-1 Candidate	Abdominal pain
Human Annotation	Correct
GPT Response	To evaluate the correctness of the sentence, let's break it down: 1. Subject Matter: The sentence discusses... 2. Symptom Identification: The sentence claims that ... Based on medical knowledge, **abdominal pain is indeed one of the symptoms associated with typhoid fever**. 3. Definite Article Usage: The sentence uses "the" before "Typhoid fever.". 4. Overall Evaluation: While the symptom mentioned is correct, **the sentence contains a grammatical error with the use of the and a punctuation issue**. The sentence was judged as **Wrong**.
GPT Output	Wrong
Retrieval knowledge	[**Typhoid fever, also known as typhoid, is a disease caused by Salmonella serotype Typhi bacteria...This is commonly accompanied by weakness, abdominal pain**, constipation, headaches, and mild vomiting...; Salmonellosis is a symptomatic infection caused by bacteria of the Salmonella type...]
MusKGC Output	Correct

Table 9: Case 2. The case is demonstrated to explain the advantages of MusKGC.

	Case 1
Facts	**Query:** (Golden State Warriors, /sports/professional_sports_team/draft-_picks./sports/sports_league_draft_pick/school, ?) **Label:** Xavier University
Top-1 Candidate	University of North Carolina
Human Annotation	Correct
GPT Response	To evaluate the correctness of the sentence, let's break it down: 1. Team Mentioned: The sentence refers to the G̈olden State Warriors,ẅhich is a professional basketball team in the NBA... 2. Historical Context: Based on historical NBA draft records, **the Golden State Warriors have not been known to draft a player from the University of North Carolina at Charlotte**...The sentence was judged as **Wrong**.
GPT Output	Wrong
Retrieval knowledge	[The 1992ˇ01393 NBA season was the Warriors' 47th season in the National Basketball Association, and 30th in the San Francisco Bay Area...; The 1998ˇ01399 NBA season was the Warriors' 53rd season... In the 1998 NBA draft, **the Golden State Warriors selected Vince Carter from the University of North Carolina** with the fifth overall pick, but soon traded him to the Toronto Raptors for his college teammate Antawn Jamison...]
MusKGC Output	Correct

Table 10: Case 1. The case is demonstrated to explain the advantages of MusKGC.