Inductive Reasoning on Few-Shot Knowledge Graphs with Task-Aware Language Models

Cheng Yan[1,2], Feng Zhao[1*], Ruilin Zhao[1], Hong Zhang[1],
[1] Natural Language Processing and Knowledge Graph Lab,
School of Computer Science and Technology,
Huazhong University of Science and Technology, Wuhan, China
[2] National Engineering Research Center of Eco-Environment
in the Yangtze River Economic Belt,
China Three Gorges Corporation, Wuhan, China
{yancheng,zhaof,ruilinzhao,zhang_hong}@hust.edu.cn

Abstract

Knowledge graphs are dynamic structures that continuously evolve as new entities emerge, often accompanied by only a handful of associated triples. Current knowledge graph reasoning methods struggle in these few-shot scenarios due to their reliance on extensive structural information. To address this limitation, we introduce ENGRAM, a novel approach that enables inductive reasoning on few-shot KGs by innovatively enriching the semantics from both textual and structural perspectives. Our key innovation lies in designing a task-aware language model that activates the language model's in-context learning ability for structured KG tasks, effectively bridging the gap between unstructured natural language and structured tasks. Unlike prior methods that inefficiently employ classification over exhaustive candidate sets, we recast knowledge graph reasoning from a generative perspective, allowing for direct computation of inference results without iterative enumeration. Additionally, we propose a distant neighborhood awareness strategy to enrich the sparse structural features of few-shot entities. Our experimental findings indicate that our method not only achieves state-of-the-art performance in few-shot scenarios. The tunable parameters of our model are approximately 1% of those in previous language model-based methods, and the inference time has been reduced to $1/10$ of that required by previous methods.

1 Introduction

In the real world, knowledge is constantly evolving, leading to the iterative development of knowledge graphs (KGs) as structured representations of real-world knowledge. Over time, these KGs accumulate a vast amount of new knowledge, resulting in a long-tail phenomenon (Kandpal et al., 2023). This segment is characterized by the emergence of numerous new entities, each associated with only a limited number of triples (Yan et al., 2025). Given this context, enriching the associative information of these low-resource entities becomes crucial, especially considering the challenges posed by scarce data resources which intensify the complexity of reasoning tasks. Therefore, the efficient utilization of limited data to perform inductive reasoning on the associative information of such entities is an urgent issue (Qi et al., 2023a).

Current mainstream methods for knowledge graph reasoning (KGR), such as those based on knowledge graph embeddings (KGE), effectively capture the associative structures of triples (Cao et al., 2024). However, KGE is fundamentally a transductive approach and struggles to represent entities that were not present during training, limiting its capability for inductive reasoning. As a result, graph neural network (GNN)-based methods have been developed. These methods inductively represent emerging entities by aggregating features from their neighboring entities (Ding et al., 2025; Geng et al., 2023). Nevertheless, in low-resource settings, these GNNs often produce ambiguous structural features due to the limited associated data, as depicted in Figure 1(a). Furthermore, the recent rise in the use of language models (LMs) has led to the development of KGR approaches that leverage LMs (Chen et al., 2023; Zhao et al., 2025). These methods extract features from natural language descriptions of triples. However, the accuracy of language model-based methods has not significantly exceeded those based on GNN, and these models also suffer from low training efficiency (Wang et al., 2022; Ao et al., 2025).

We believe the suboptimal performance of language models in few-shot scenarios can primarily be attributed to: (1) **Heterogeneity between structured tasks and natural language**: Language models typically process unstructured natural language, which significantly differs from the struc-

[*] Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12656–12666
November 4-9, 2025 ©2025 Association for Computational Linguistics

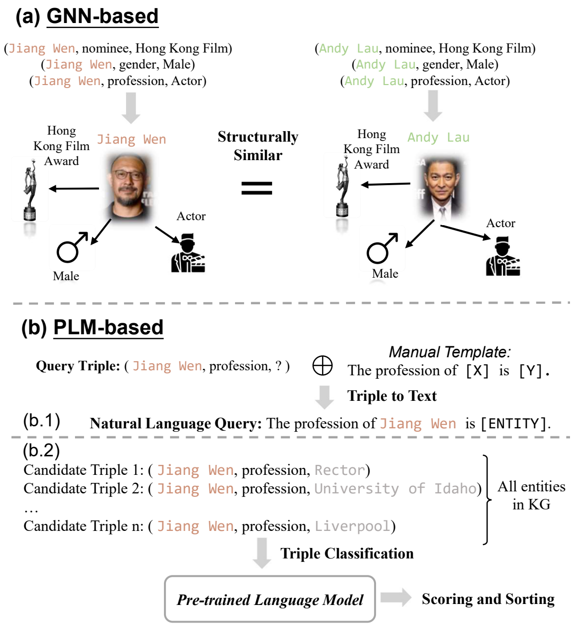
(a) GNN-based
(Jiang Wen, nominee, Hong Kong Film)
(Jiang Wen, gender, Male)
(Jiang Wen, profession, Actor)
(Andy Lau, nominee, Hong Kong Film)
(Andy Lau, gender, Male)
(Andy Lau, profession, Actor)
Hong Kong Film Award
Jiang Wen
Structurally Similar
Hong Kong Film Award
Andy Lau
Actor
Male
Actor
Male
(b) PLM-based
Query Triple: (Jiang Wen, profession, ?)
Manual Template:
The profession of [X] is [Y].
Triple to Text
(b.1)
Natural Language Query: The profession of Jiang Wen is [ENTITY].
(b.2)
Candidate Triple 1: (Jiang Wen, profession, Rector)
Candidate Triple 2: (Jiang Wen, profession, University of Idaho)
…
Candidate Triple n: (Jiang Wen, profession, Liverpool)
All entities in KG
Triple Classification
Pre-trained Language Model
Scoring and Sorting

Figure 1: (a) In few-shot settings, GNN-based methods often struggle with structural feature similarity due to the sparse nature of the graph structure; (b) when processing triple queries, language models typically rely on manually crafted templates and need to iterate through all candidate triples.

tured knowledge in KGs. Consequently, language models face difficulties in directly comprehending structured triples when applied to KGR tasks. Some methods employ manually designed prompt templates to transform triple queries into natural language queries (Lv et al., 2022), Figure 1(b.1) illustrates the activation of relevant knowledge in the language model by means of knowledge probing (Petroni et al., 2019). (2) **Inefficiency of classification-based inference pattern**: Current methods are primarily implemented through triple classification (Wang et al., 2021), such as the link prediction task in Figure 1(b.2). This process requires iterating through all candidate triples, each assessed by the language model to determine the probability of the triple's accuracy. Additionally, this approach incurs significant computational redundancy since the variation only occurs in the candidate entities, resulting in inefficient use of computational resources.

To bridge the gap between structured tasks and language models, and to enhance inference efficiency, we propose an inductiv**E** reaso**N**in**G** method for few-shot KGs using a task-awa**R**e langu**A**ge **M**odel (ENGRAM). First, we introduce a task-guided prompting method that provides the language model with task-specific examples drawn from the few associated triples of few-shot entities. By doing so, we activate the language model's in-context learning abilities (Akyürek et al., 2023), enabling it to understand and perform structured KG tasks without the need for manual template design. Second, moving away from the traditional classification paradigm, we reconceptualize link prediction as a generative task. This shift allows the language model to produce candidate entities directly, significantly reducing computational overhead by eliminating the need to iterate over all possible candidates. We leverage a pre-built contextual representation pool that stores the textual features of observed entities, facilitating simultaneous scoring of all candidate entities. Additionally, we propose a strategy that incorporates information from distant but semantically related entities within the KG. By doing so, we construct a more informative context for each entity, which improves feature differentiation and, consequently, reasoning performance. Finally, Our method integrates both textual and structural information by computing context scores and triple scores, which are then used to jointly train the language model and the GNN components.

We compared our method with current state-of-the-art approaches based on language models and graph structural features on benchmark datasets. Our model demonstrates superior computational efficiency, featuring significantly fewer tunable parameters and reducing the computational complexity during the inference phase. Further experiments focus on visual analyses of the in-context learning ability of language models and explore the effects of language models (BERT, RoBERTa, T5, Llama) with different parameter scales. In summary, our contributions are as follows:

- We propose ENGRAM to address few-shot entities in KGs from the combined perspectives of graph structure and textual information, enabling LMs to handle structured tasks and fully utilize the limited associated data.

- We propose a task-aware language model that leverages the in-context learning capabilities of LMs to complete structured link prediction tasks, transforming link prediction into a generative task, which allows for direct computation of candidate entities without iterative classification.

- Extensive experiments on benchmark datasets demonstrate that ENGRAM achieves SOTA performance. Specifically, ENGRAM uses approximately 1% of the tunable parameters required by previous language model-based methods and reduces inference time by an order of magnitude.

2 Related Work

Inductive reasoning Inductive reasoning in KGs primarily addresses the emergence of new entities or relations within the graph, utilizing associative information to represent these entities or relations and establishing connections with the existing KG. MEAN (Hamaguchi et al., 2017) initially focused on the out-of-knowledge-base entity problem, highlighting that KGE methods are transductive and cannot cope with more realistic inductive settings. Building on MEAN, LAN (Wang et al., 2019) introduced an attention-based feature aggregation method. GraIL (Teru et al., 2020) and CoMPILE (Mai et al., 2021) extract subgraphs formed around target nodes and then encode these subgraphs to perform reasoning on local subgraphs. RMPI (Geng et al., 2023) and INGRAM (Lee et al., 2023) take into account the new entities and relations appearing in the KG, constructing a relation-view graph and obtaining representations of new relations and entities from both relation-level and entity-level perspectives.

Few-shot reasoning To make more efficient use of few-shot data, a series of meta-learning-based methods have been proposed for modeling few-shot relations and entities. GMatching (Xiong et al., 2018) was the first to focus on one-shot relations in KGs. MetaR (Chen et al., 2019), FSRL (Zhang et al., 2020), GANA (Niu et al., 2021), and MetaP (Jiang et al., 2021) utilize intricately designed GNNs to learn meta-knowledge with generalization capabilities. GEN (Baek et al., 2020) was the first to consider few-shot entities within KGs. BayesKGR (Zhao et al., 2023a) estimated the uncertainty in few-shot KGs and used Bayesian neural networks to model the uncertainty in the inference process. RawNP (Zhao et al., 2023b) utilized neural processes to model the distribution of limited data.

Language models for reasoning In recent years, with the rise of language models (Hu et al., 2022; Liu et al., 2021; Zhao et al., 2024), research has begun to focus on the textual information in KGs (Qi et al., 2023b; Jiang et al., 2023). For example, KG-BERT (Yao et al., 2019) was the first to use the BERT model to achieve link prediction tasks through triple classification. StAR (Wang et al., 2021) introduced a Siamese-style textual encoder that uses a shared-parameter model to encode both query and candidate entities separately, enhancing inference efficiency. SimKGC (Wang et al., 2022) proposed a contrastive learning framework, which improved the quality and efficiency of negative sampling. CSProm-KG (Chen et al., 2023) introduced a method based on soft prompts, embedding learnable soft prompts into the triplet inputs to reduce the training overhead of language models. TAGREAL (Jiang et al., 2023) implemented an automated template construction method, achieving better knowledge probing results and enhancing the method's scalability.

3 Preliminary

Definition 1 (Knowledge Graph with Text). *A knowledge graph can be defined as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where: $\mathcal{E}$ is the set of entities, $\mathcal{R}$ is the set of relations, $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is the set of triples. At the textual level, each entity and relation is not only identified by its graph structure but also enriched with textual descriptions that provide additional semantic information. Formally, this can be defined as: X^E is an extended set of entities, where each entity $e \in \mathcal{E}$ is associated with a descriptive text $\mathbf{x}_e^E$, $X^E = \{(e, \mathbf{x}_e^E) | e \in \mathcal{E}\}$; X^R is an extended set of relations, where each relation $r \in \mathcal{R}$ is associated with a descriptive text $\mathbf{x}_r^R$, $X^R = \{(r, \mathbf{x}_r^R) | r \in \mathcal{R}\}$; X^T is an extended set of triples, where each triple (e_1, r, e_2) not only represents the relation between entities but also includes the descriptions of these entities and relations, $X^T = \{((e_1, \mathbf{x}_{e_1}^E), (r, \mathbf{x}_r^R), (e_2, \mathbf{x}_{e_2}^E)) | (e_1, r, e_2)) \in \mathcal{T}\}$.*

Definition 2 (Few-shot link prediction). *Emerging entities refer to entities that do not appear in the original set of entities, denoted as $e' \in \mathcal{E}'$, where $\mathcal{E}' \cap \mathcal{E} = \varnothing$. These emerging entities typically have only a few associated triples and are also known as few-shot entities. Each few-shot entity is associated with K triples: $|\{(e_i, r_i, e')$ or $(e', r_i, e_i)\}_{i=1}^{K}| = K$, where $e_i \in \mathcal{E}$ and K is a small number, such as 1 or 3. Few-shot KGR involves completing link prediction tasks related to few-shot entities, such as $(e', r, ?)$ and $(?, r, e')$.*

4 Method

4.1 Task-Guided Prompt

Previous approaches that employed language models for link prediction tasks often reformulated

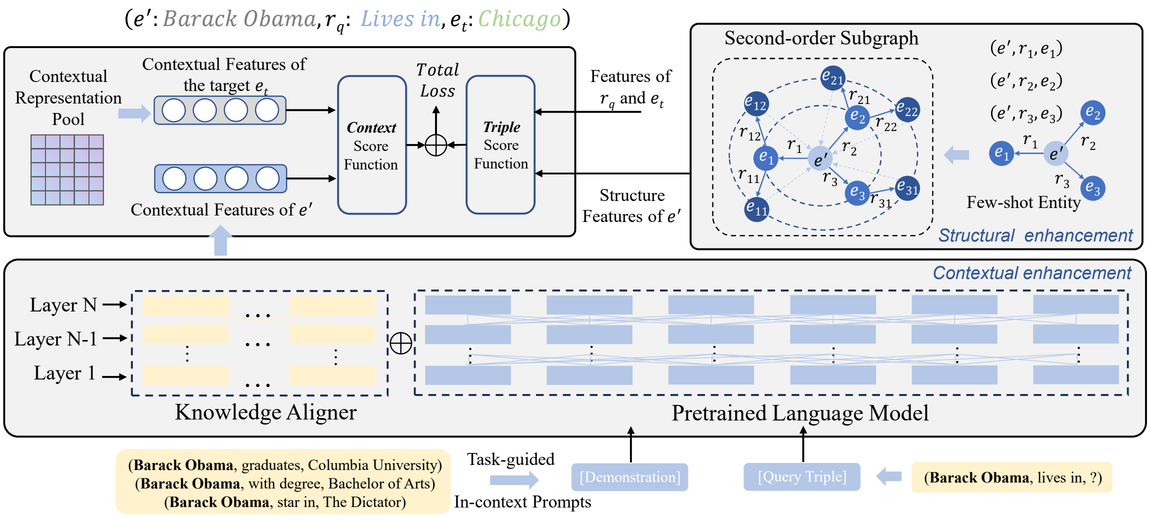

Figure 2: Overview of the proposed ENGRAM framework.

these tasks into cloze tasks. (Lv et al., 2022; Jiang et al., 2023). For instance, a query like *(Barack Obama, with degree, ?)* would require the construction of a prompt template such as *"with degree"* → *"[X] can grant a [Y]"*, transforming the original query into *"Barack Obama can grant a [MASK]"*. This process activates related knowledge in the language model through knowledge probing (Petroni et al., 2019). However, the conversion of triples to natural language can be resource-intensive, and evaluating the efficacy of template-based knowledge probing presents significant challenges.

To address these issues, we propose a task-guided prompting method that utilizes triples directly as in-context prompts. This innovative approach provides a more intuitive and seamless connection between structured knowledge and natural language. Specifically, for each triple (e_1, r, e_2), we construct formalized query-answer pairs, such as $(e_1, r, ?) \rightarrow e_2$. This reformulation transforms the original triple into a link prediction query-answer pair, which enables the model to more effectively understand and process structured knowledge graph link prediction tasks. For an emerging entity e', the observed associated triples are $\{(e_i, r_i, e')$ or $(e', r_i, e_i)\}_{i=1}^{K}$. Thus, we can construct query text $x_i = (\mathbf{x}_{e'}^E, \mathbf{x}_{r_i}^R, ?)$ and answer text $y_i = \mathbf{x}_{e_i}^E$, with the in-context prompts for emerging entity e' formalized as the string $C = [x_1; y_1; x_2; y_2; ...; x_K; y_K]$.

4.2 Generative Strategy for Inductive Reasoning

Current language model-based KGR methods were mainly implemented through triple classification (Yao et al., 2019; Wang et al., 2021). Consequently, link prediction tasks are often transformed into triple classification tasks, where all entities are considered as candidate entities in the query, and the classification results for all triples are computed. The logits produced by the language model act as scores, indicating the likelihood of each triple's correctness. As a result, such methods are computationally intensive and exhibit considerable redundancy since only the candidate entities vary while the query remains unchanged.

To address this, we propose a generative strategy for inductive reasoning. Our first step involves constructing a Contextual Representation Pool (CRP) to store the textual representations of observed entities. This paper primarily addresses establishing connections between emerging entities and existing entities in the KG, where entities in the existing KG are pre-observed. Therefore, we can pre-input all the observed entities into the language model LM to obtain their textual features to form the feature matrix: $\mathbf{P} = [\mathbf{emb}_{e_0}; \mathbf{emb}_{e_1}; ...; \mathbf{emb}_{e_N}]$, where $\mathbf{emb}_{e_i} = \mathrm{LM}(\mathbf{x}_{e_i}^E)$ represents the embedding vector of entity e_i.

Then, we concatenate the query triple $x' = (\mathbf{x}_{e'}^E, \mathbf{x}_{r_i}^R, ?)$ and in-context prompts $C = [x_1; y_1; x_2; y_2; ...; x_K; y_K]$ as input to the language model. The output features of the language model LM represent the predicted results:

$$\mathbf{emb}_{e_{pred}} = \mathrm{LM}([C; x']). \qquad (1)$$

These generated feature embeddings are then multiplied by the CRP entity feature matrix and normalized to compute the score for each candidate entity:

$$\text{Context Score} = \text{softmax}(\mathbf{P} \cdot \mathbf{emb}_{e_{pred}}^T). \quad (2)$$

By establishing the contextual representation pool in advance, we circumvent the high computational expense associated with iterating through all candidate triples, reducing the number of forward passes through the language model from linear $\mathcal{O}(|\mathcal{E}|)$ to constant $\mathcal{O}(1)$.

4.3 Knowledge Distribution Alignment

Language models are generally trained on large-scale corpora, often referred to as open-world knowledge (Song et al., 2023). In contrast, KGs are typically constructed within specific domains, known as closed-world knowledge. This fundamental difference leads to issues of inconsistent knowledge distribution when applying language models to KGR tasks. To address the knowledge distribution bias, we propose an efficient knowledge alignment strategy. This strategy involves concatenating each layer of the language model with k trainable vectors. During the training process, we fine-tune only the prefix vectors for each layer, while keeping the original model parameters frozen (Liu et al., 2021).

Assuming the language model comprises L layers, each layer θ_l is characterized by parameters θ_l. For each layer, we introduce an additional prefix parameter $P_l \in \mathbb{R}^{k \times d}$, where d is the dimension of the word vector associated with the language model. At each layer l, the original input x_l is modified by concatenating it with the prefix parameters P_l:

$$x'_l = \text{Concatenate}(P_l, x_l), \quad (3)$$

where x'_l is the modified input. With the modified input x'_l, the output y_l of each layer is calculated as follows:

$$y_l = F_l(x'_l; \theta_l), \quad (4)$$

where F_l is the transformation function of layer l, and θ_l are the layer's fixed parameters. The final output of Eq. (1) is obtained by sequentially processing the modified inputs through all layers:

$$\text{LM}([C; x']) = F_L(...F_2(F_1(x'_1; \theta_1); \theta_2)...; \theta_L). \quad (5)$$

During training, only the prefix parameters P_l are tunable, while the core parameters of the layer θ_l remain fixed.

4.4 Distant Neighborhood Awareness

To address the structural sparsity issue, we propose a distant neighborhood awareness method to enhance the structural features of few-shot entities. We classify the neighbors of few-shot entities into two categories: direct neighbors (those directly connected) and distant neighbors (second-order neighbors). For a few-shot entity e', the features of their direct and distant neighbors are processed by two distinct GNNs (Direct Neighbors Network and Distant Neighbors Network), respectively. The feature extraction processes are formulated as follows:

$$\mathbf{h}_{direct} = \text{ReLU}(\sum_{e_i \in \mathcal{N}(e')} \frac{1}{|\mathcal{N}(e')|} \mathbf{W}_{direct} \cdot \mathbf{e}_i), \quad (6)$$

$$\mathbf{h}_{distant} = \text{ReLU}(\sum_{e_i \in \mathcal{N}_2(e')} \frac{1}{|\mathcal{N}(e')|} \mathbf{W}_{distant} \cdot \mathbf{e}_i), \quad (7)$$

where $\mathcal{N}(e')$ and $\mathcal{N}_2(e')$ represent the sets of direct and second-order neighbor entities of e', respectively, and $\mathbf{W}_g$ and $\mathbf{b}_g$ are the network weights. To integrate these two types of features effectively, we have designed a gating mechanism that adaptively adjusts the fusion weights based on the significance of the distant neighbor features:

$$\mathbf{e}' = g(\mathbf{h}_{distant}) \cdot \mathbf{h}_{direct} + (1 - g(\mathbf{h}_{distant})) \cdot \mathbf{h}_{distant}, \quad (8)$$

$$g(\mathbf{h}_{distant}) = \sigma(\mathbf{W}_g \mathbf{h}_{distant} + \mathbf{b}_g), \quad (9)$$

where σ is the activation function, $\mathbf{W}_g$ and $\mathbf{b}_g$ are the network weights and bias parameters, $\mathbf{e}_i$ is the final structural feature of entity e'.

4.5 Joint Training

During the training phase, given a query $(e', r_q, ?)$ where the target entity is e_t, the feature of the answer entity is initially predicted as $\mathbf{emb}_{e_{pred}}$ using the language model, as defined by Eq. (1). Subsequently, the structural feature of e' is calculated as $\mathbf{e}'$ based on the distant neighborhood awareness. The final scoring function for the candidate triple is therefore composed of two components:

$$\begin{aligned} Score(e', r_q, e_t) = \sigma \cos(\mathbf{emb}_{e_{pred}}, \mathbf{emb}_{e_t}) \\ + \alpha f(e', r_q, e_t), \end{aligned} \quad (10)$$

where α is a hyperparameter that modulates the balance between the textual and structural contributions to the score, and $f(e', r_q, e_t) = -\|\mathbf{e}' + \mathbf{r}_q - \mathbf{e}_t\|$ is the score function of KGE methods, commonly used to assess the correctness of triples.

In calculating the loss function, each few-shot entity e' possesses a query set $\mathcal{Q}_i$ encompassing related queries. The entire model is optimized using a hinge loss function:

$$\mathcal{L}(\mathcal{Q}_i) = \sum_{(e',r_q,e_t)\in\mathcal{Q}_i} \max(\gamma - Score(e', r_q, e_t) + Score(e', r_q, e_t^-)), \tag{11}$$

where e_t^- is the corrupted entity, and γ is the margin hyperparameter used to distinguish between positive and negative samples.

5 Experiment

5.1 Experiment Setup

Datasets. Aligned with prior research (Baek et al., 2020), we evaluated the performance of few-shot link prediction on the FB15K-237 and NELL-995 datasets. NELL-995 includes 3,000 emerging entities with 31,873 associated triples, while FB15K-237 contains 5,000 emerging entities with 88,178 associated triples. In terms of textual descriptions, we utilized the text descriptions of entities and relations collected by KG-BERT (Yao et al., 2019).

Baselines. We have classified the baselines into two categories: (1) GNN-based: MEAN (Hamaguchi et al., 2017) and LAN (Wang et al., 2019) can aggregate associated triplet features of emerging entities; FSRL (Zhang et al., 2020), and GANA (Niu et al., 2021) focus on few-shot relationships within KGs and utilize a meta-learning framework to model these relations, GEN (Baek et al., 2020) and RawNP (Zhao et al., 2023b) also employ a meta-learning framework to model few-shot entities within KGs. (2) LM-based: KG-BERT (Yao et al., 2019), StAR (Wang et al., 2021), PKGC (Lv et al., 2022), and SimKGC (Wang et al., 2022) are based on language models. KICGPT (Wei et al., 2023), DIFT(Liu et al., 2024), and KoPA (Zhang et al., 2024) are based on LLMs.

Implementation Details. We used *RoBERTa-base* (Liu et al., 2019) as the initial language model to encode textual knowledge and employed TransE (Bordes et al., 2013) to initialize the embeddings for entities and relations, setting the embedding dimension at 200. AdamW was chosen as the optimizer, configured with a learning rate of 5×10^{-5} and a weight decay of 0.05. The balance between textual and structural features, denoted by α, was maintained at 1. For each triple in the training process, the size of negative sampling is 8. The number of prefix parameters introduced per layer k is set to 4. The datasets are sourced from GEN (Baek et al., 2020) and KG-BERT (Yao et al., 2019). The experimental environment is on Ubuntu 20 with RTX 3090 * 4.

Evaluation Protocol and Metrics. During the inference phase, for a correct triplet involving an emerging entity, we replace the non-emerging entity in the triple with other remaining entities to form multiple corrupted triples. Both the correct triplet and the corrupted triples are scored and ranked to determine the rank of the correct triple. We use Hits@n and MRR as evaluation metrics, where Hits@n is the proportion of correct triples within the top-n during testing, and MRR is the average reciprocal rank of all test triples.

5.2 Main Results

Table **??** illustrates the performance of all methods under 1-shot and 3-shot scenarios, demonstrating that our method significantly outperforms current GNN-based and LM-based methods. Among the baselines, GNN-based methods like MEAN and LAN perform poorly in few-shot scenarios, whereas in the meta-learning category, methods such as MetaR, FSRL, and GANA, although focused on modeling few-shot relations, do not adequately address few-shot entities. We noted that methods employing graph structural features, such as GEN and RawNP, deliver performance comparable to those based on language models. These approaches represent two distinct strategies for tackling few-shot challenges—via graph structure and textual analysis, respectively. GEN and RawNP efficiently utilize limited training data to model few-shot entities through meta-learning. In contrast, LM-based methods enhance few-shot information leveraging knowledge accumulated during the pre-training process. Our method outstrips both categories, rectifying the shortcomings prevalent in current language model-based approaches.

In Table **??**, we evaluate the effectiveness of our proposed method across a range of backbone LMs, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020), and Llama3-8B (Dubey et al., 2024). Given that our method has demonstrated superior performance compared to certain LLM-based approaches (as indicated in Table **??**), the objective of this experiment is to further investigate its performance consistency across varying scales of language models.

The results presented clearly show a general trend where LMs with greater parameter counts

Method	NELL-995								FB15K-237							
	1-shot				3-shot				1-shot				3-shot			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
MEAN	0.158	0.107	0.173	0.263	0.180	0.124	0.189	0.296	0.105	0.052	0.109	0.207	0.114	0.058	0.119	0.217
LAN	0.163	0.109	0.179	0.274	0.185	0.127	0.195	0.307	0.113	0.061	0.121	0.223	0.119	0.064	0.125	0.231
FSRL	0.067	0.054	0.068	0.091	0.085	0.064	0.095	0.126	0.097	0.065	0.104	0.156	0.090	0.054	0.096	0.150
GANA	0.090	0.057	0.101	0.147	0.093	0.060	0.104	0.158	0.103	0.048	0.109	0.184	0.110	0.061	0.112	0.201
GEN	0.282	0.206	0.320	0.421	0.291	0.217	0.333	0.433	0.367	0.282	0.410	0.530	0.382	0.289	0.430	0.565
RawNP	0.283	0.210	0.316	0.419	0.314	0.243	0.352	0.452	0.371	0.289	0.411	0.532	0.409	0.323	0.453	0.575
KG-BERT	0.154	0.101	0.168	0.267	0.157	0.103	0.171	0.261	0.115	0.061	0.124	0.237	0.121	0.065	0.129	0.238
StAR	0.217	0.150	0.233	0.330	0.236	0.151	0.255	0.399	0.328	0.243	0.332	0.478	0.344	0.250	0.351	0.507
PKGC	0.210	0.147	0.244	0.345	0.231	0.148	0.249	0.387	0.321	0.239	0.322	0.462	0.339	0.248	0.343	0.493
SimKGC	0.231	0.163	0.257	0.381	0.245	0.159	0.268	0.419	0.336	0.245	0.343	0.501	0.352	0.253	0.361	0.534
KICGPT	0.287	0.192	0.323	0.434	0.290	0.194	0.336	0.441	0.374	0.277	0.421	0.542	0.378	0.278	0.439	0.578
KoPA	0.225	0.158	0.249	0.355	0.239	0.158	0.261	0.388	0.330	0.241	0.339	0.503	0.348	0.249	0.355	0.523
DIFT	0.242	0.169	0.261	0.394	0.258	0.183	0.284	0.421	0.356	0.260	0.387	0.539	0.367	0.272	0.381	0.552
ENGRAM	**0.323**	**0.238**	**0.355**	**0.453**	**0.357**	**0.275**	**0.391**	**0.519**	**0.401**	**0.357**	**0.443**	**0.576**	**0.409**	**0.324**	**0.456**	**0.589**
ENGRAM(std)	± 0.007	± 0.013	± 0.015	± 0.018	± 0.014	± 0.010	± 0.012	± 0.015	± 0.013	± 0.007	± 0.009	± 0.014	± 0.012	± 0.011	± 0.012	± 0.016

Table 1: 1-shot (1-S) and 3-shot (3-S) results on benchmark datasets.

Backbone	NELL-995		FB15K-237	
	MRR	Hits@10	MRR	Hits@10
BERT-Base	0.341	0.492	0.397	0.572
BERT-Large	0.350	0.501	0.408	0.586
RoBERTa-Base	0.357	0.519	0.409	0.589
RoBERTa-Large	0.361	0.526	0.413	0.596
T5-Base	0.353	0.515	0.407	0.584
T5-Large	0.358	0.523	0.413	0.596
Llama3-8B	0.351	0.509	0.402	0.581

Table 2: Results on different LMs as backbone.

tend to exhibit enhanced performance. Nonetheless, an intriguing deviation is observed with the Llama3-8B model, which underperforms relative to smaller models. Figure 3 provides additional insights into the effects of employing Llama3. We propose two main explanations for this phenomenon: (1) **Knowledge Distribution Gap**: a notable discrepancy exists between the knowledge distribution of Llama3 and the target KG; (2) **Model Complexity and Fine-Tuning Constraints**: larger-scale models such as Llama3 may face greater difficulty fitting limited few-shot data under standard fine-tuning constraints.

Method	NELL-995		
	MRR	Hits@1	Hits@10
-Prompt	0.294	0.219	0.451
-Tuning	0.110	0.073	0.161
-Distant	0.341	0.267	0.499
ENGRAM	0.357	0.275	0.519

Table 3: Results of ablation models.

5.3 Ablation Study

To validate the effectiveness of each component in our model, we conducted a series of ablation experiments: (1) We simplified the Task-Guided Prompt to merely inputting the query directly into the language model (-Prompt); (2) To assess the impact of Knowledge Distribution Alignment, we did not fine-tune the model parameters (-Tuning); (3) Finally, to evaluate the impact of Distant Neighborhood Awareness, we represented few-shot entities using only features from direct neighbors (-Distant). The results, as shown in Table ??, indicate that "-Tuning" has the most substantial impact on the model, highlighting the critical role of knowledge distribution alignment. Moreover, task-guided prompting significantly contributes to the model's performance, demonstrating that in-context prompts can effectively guide the LM in completing link prediction tasks. Structurally, distant neighborhood awareness offers distinct advantages over solely using direct neighbor features.

Method	Tunable Parameters	Complexity	Inference Time
KG-BERT	125M	$\mathcal{O}(\|\mathcal{E}\|)$	7h
StAR	125M	$\mathcal{O}(\|\mathcal{E}\|)$	12min
SimKGC	218.9M	$\mathcal{O}(\|\mathcal{E}\|)$	15min
ENGRAM	1.9M	$\mathcal{O}(1)$	55s

Table 4: Comparison of computational efficiency.

5.4 Efficiency Analysis

Table ?? displays the size of tunable parameters, the computational complexity, and the inference time required to process the same test set across different language model-based approaches. Notably, the number of tunable parameters in our model is approximately 1% of that in SimKGC and our inference time is 1/10 of its. KG-BERT and StAR are both based on the BERT-base model and involve full parameter fine-tuning, resulting in substantial training overhead. During the inference phase, inference is achieved through triple classifi-

Query Triple: (Barack Obama, lives in, ?) **Target: Chicago**	**Prompt: (Barack Obama, graduates,?) → Columbia University** **(Barack Obama, with degree,?) → Bachelor of Arts** **(Barack Obama, star in,?) → The Dictator**
Llama2 (7B):	1. New York(0.17), 2. Washington (0.12), 3. **Chicago (0.09)**, 4. Hawaii (0.09), 5. Indonesia (0.04) …
RoBERTa-Large (355M):	1. **Chicago (0.16)**, 2. Homeland (0.03), 3. Oslo (0.02), 4. Brooklyn (0.02), 5. Chelsea (0.02) …
RoBERTa-Base (125M):	1. Harvard (0.06), 2. University (0.02), 3. UCLA (0.02), 4. Yale (0.01), 5. History (0.01) …
DeBERTaV3 (86M):	1. College (0.04), 2. Health (0.03), 3. History (0.01), 4. Opportunity (0.01), 5. Unicef (0.01) …

Figure 3: Results of language models with different parameter scales.

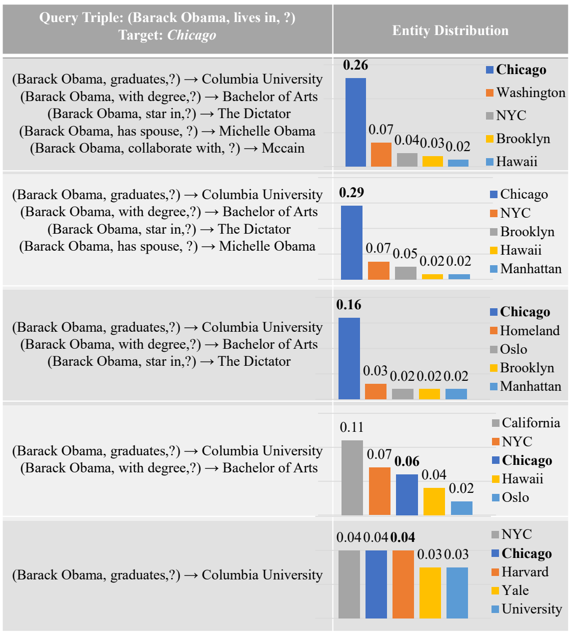

Figure 4: The distribution of entities output by the language model.

cation, which has a high computational complexity proportional to the number of entities, i.e., $\mathcal{O}(|\mathcal{E}|)$. SimKGC employs two BERT-base models and relies heavily on a large number of negative samples, which extends the training duration. In contrast, our model introduces only a minimal number of additional tunable parameters. Unlike SimKGC, our contextual representation pool stores the textual features of entities, enabling the language model to perform a single forward computation during the inference stage. This strategic modification substantially reduces the number of forward passes from $\mathcal{O}(|\mathcal{E}|)$ to a single pass, after which candidate scoring can be performed efficiently through lightweight similarity calculations. Although the overall complexity is not strictly constant, this design greatly streamlines the inference process and leads to significant practical efficiency gains.

5.5 Case Study

The case study provides examples of the use of in-context prompts for language models at different scales. Figure 3 presents relevant queries and in-context prompts, showcasing the predicted entities and corresponding scores from different models. Among the models (RoBERTa-Large, RoBERTa-Base, and DeBERTaV3 (He et al., 2023), with sequentially decreasing parameter sizes), it is evident that RoBERTa-Large produces the most accurate predictions, while DeBERTaV3 demonstrates the weakest performance. Notably, RoBERTa-Large assigns the highest scores to correct answers, unlike RoBERTa-Base and DeBERTaV3, which attribute lower scores to their predictions. This pattern suggests that models with larger parameters possess superior in-context learning capabilities.

To further validate this observation, we performed tests with the large language model Llama3, which did not yield the most optimal prediction outcomes. Analysis of Llama's predictions indicates that its responses are correct under the open-world knowledge assumption, as demonstrated by *Barack Obama*'s historical residencies in *New York City* before his presidency and in *Washington D.C.* during his tenure. This case underscores the issue of knowledge distribution bias discussed in Section 4.3 and confirms the need for the knowledge distribution alignment we propose, addressing discrepancies between historical data and current model understanding.

Figure 4 provides the distribution of the model's predicted entities given 0 to 4 in-context demonstrations. We can observe that when given 0 or 1 demonstration, the model struggles to output accurate entities, and the scores for each entity are very low. Starting from 2 demonstrations, the model is able to output the correct entities with higher scores. Moreover, as the number of demonstrations increases, the scores for the correct entities also increase and tend to stabilize.

6 Conclusion

This paper tackles the issue of inductive reasoning within KGs under few-shot scenarios. To tackle this, we introduce an efficient reasoning method that employs a task-aware language model, activating the in-context learning capabilities of the LM and reducing the data discrepancies between structured tasks and LMs. Additionally, we have reconsidered the problem of inductive reasoning in KGs from a generative perspective. Structurally, we have enabled few-shot entities to perceive distant entities, considerably broadening their perceptual domain. Experimental results demonstrate significant improvements in both the accuracy and efficiency of inference.

Limitations

The main limitation of this work lies in the substantial GPU memory consumption caused by the preconstructed Contextual Representation Pool during inference, especially when testing on large-scale knowledge graphs or using larger language models. Therefore, it is necessary to design a more flexible and scalable feature scheduling mechanism to reduce memory usage during the inference process.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2023YFF0905503, National Natural Science Foundation of China under Grants No.62472188.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models. In *Proceedings of The Eleventh International Conference on Learning Representations, ICLR 2023*.

Tu Ao, Yanhua Yu, Yuling Wang, Yang Deng, Zirui Guo, Liang Pang, Pinghui Wang, Tat-Seng Chua, Xiao Zhang, and Zhen Cai. 2025. Lightprof: A lightweight reasoning framework for large language model on knowledge graph. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 23424–23432. AAAI Press.

Jinheon Baek, Dong Bok Lee, and Sung Ju Hwang. 2020. Learning to extrapolate knowledge: Transductive few-shot out-of-graph link prediction. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.

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*, pages 2787–2795.

Jiahang Cao, Jinyuan Fang, Zaiqiao Meng, and Shangsong Liang. 2024. Knowledge graph embedding: A survey from the perspective of representation spaces. *ACM Comput. Surv.*, 56(6):159:1–159:42.

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

Mingyang Chen, Wen Zhang, Wei Zhang, Qiang Chen, and Huajun Chen. 2019. Meta relational learning for few-shot link prediction in knowledge graphs. 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*, pages 4216–4225.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186.

Ling Ding, Lei Huang, Zhizhi Yu, Di Jin, and Dongxiao He. 2025. Towards global-topology relation graph for inductive knowledge graph completion. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 11581–11589. AAAI Press.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Yuxia Geng, Jiaoyan Chen, Jeff Z. Pan, Mingyang Chen, Song Jiang, Wen Zhang, and Huajun Chen. 2023. Relational message passing for fully inductive knowledge graph completion. In *Proceedings of 39th IEEE International Conference on Data Engineering, ICDE 2023*, pages 1221–1233.

Takuo Hamaguchi, Hidekazu Oiwa, Masashi Shimbo, and Yuji Matsumoto. 2017. Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, page 1802–1808.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023*.

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 *The Tenth International Conference on Learning Representations, ICLR 2022*.

Pengcheng Jiang, Shivam Agarwal, Bowen Jin, Xuan Wang, Jimeng Sun, and Jiawei Han. 2023. Text augmented open knowledge graph completion via pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11161–11180.

Zhiyi Jiang, Jianliang Gao, and Xinqi Lv. 2021. Metap: Meta pattern learning for one-shot knowledge graph completion. In *Proceedings of The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021*, pages 2232–2236.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of International Conference on Machine Learning, ICML 2023*, volume 202, pages 15696–15707.

Jaejun Lee, Chanyoung Chung, and Joyce Jiyoung Whang. 2023. Ingram: Inductive knowledge graph embedding via relation graphs. In *Proceedings of International Conference on Machine Learning, ICML 2023*, volume 202, pages 18796–18809.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Yang Liu, Xiaobin Tian, Zequn Sun, and Wei Hu. 2024. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Xin Lv, Yankai Lin, Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. Do pretrained 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.

Sijie Mai, Shuangjia Zheng, Yuedong Yang, and Haifeng Hu. 2021. Communicative message passing for inductive relation reasoning. In *Proceedings of Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 4294–4302.

Guanglin Niu, Yang Li, Chengguang Tang, Ruiying Geng, Jian Dai, Qiao Liu, Hao Wang, Jian Sun, Fei Huang, and Luo Si. 2021. Relational learning with gated and attentive neighbor aggregator for few-shot knowledge graph completion. In *Proceedings of The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021*, pages 213–222.

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

Kunxun Qi, Jianfeng Du, and Hai Wan. 2023a. Learning from both structural and textual knowledge for inductive knowledge graph completion. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Kunxun Qi, Jianfeng Du, and Hai Wan. 2023b. Learning from both structural and textual knowledge for inductive knowledge graph completion. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Xiaoshuai Song, Keqing He, and Pei Wang. 2023. Large language models meet open-world intent discovery and recognition: An evaluation of chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 10291–10304.

Komal K. Teru, Etienne G. Denis, and William L. Hamilton. 2020. Inductive relation prediction by subgraph reasoning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119, pages 9448–9457.

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 *Proceedings of The Web Conference 2021, WWW 2021*, pages 1737–1748.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022. 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*, pages 4281–4294.

Peifeng Wang, Jialong Han, Chenliang Li, and Rong Pan. 2019. Logic attention based neighborhood aggregation for inductive knowledge graph embedding. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 7152–7159.

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.

Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2018. One-shot relational learning for knowledge graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 1980–1990.

Cheng Yan, Feng Zhao, Xiaohui Tao, and Xiaofeng Zhu. 2025. Multi-view few-shot reasoning for emerging entities in knowledge graphs. *IEEE Trans. Big Data*, 11(3):1321–1333.

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

Chuxu Zhang, Huaxiu Yao, Chao Huang, Meng Jiang, Zhenhui Li, and Nitesh V. Chawla. 2020. Few-shot knowledge graph completion. In *Proceedings of The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 3041–3048.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Wen Zhang, and Huajun Chen. 2024. Making large language models perform better in knowledge graph completion. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 233–242. ACM.

Feng Zhao, Cheng Yan, Hai Jin, and Lifang He. 2023a. Bayeskgr: Bayesian few-shot learning for knowledge graph reasoning. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 22(6):160:1–160:21.

Ruilin Zhao, Feng Zhao, Liang Hu, and Guandong Xu. 2024. Graph reasoning transformers for knowledge-aware question answering. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Vancouver, Canada*, pages 19652–19660. AAAI Press.

Ruilin Zhao, Feng Zhao, and Hong Zhang. 2025. Correcting on graph: Faithful semantic parsing over knowledge graphs with large language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 5364–5376. Association for Computational Linguistics.

Zicheng Zhao, Linhao Luo, Shirui Pan, Quoc Viet Hung Nguyen, and Chen Gong. 2023b. Towards few-shot inductive link prediction on knowledge graphs: A relational anonymous walk-guided neural process approach. In *Proceedings of Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023*, volume 14171, pages 515–532.