Rethinking Knowledge Graph Reasoning via Bi-stage Tuning for Inference Speedup and Antiphrasis Evaluation

Anonymous submission

Abstract

Leveraging pretrained language models (PLM) for knowledge graph reasoning is an exciting emerging venue. Prefix-Tuning has further addressed the optimization challenges of PLMs by efficiently tuning a small portion of additional parameters. However, adopting Prefix-Tuning to evolve knowledge graph reasoning is still in its infancy. This paper proposes Bi-Link, a novel bi-stage Prefix-Tuning approach to address the space-time learning challenges of reasoning. Bi-Link is a space-time tradeoff process that improves inference speed by pre-activating entities and reasoning relations with precomputed representations in sequential training phases. At inference time, Bi-Link leverages pre-activated representations from a memory bank and swiftly reasons relational texts to reduce latency. We also introduce a novel antiphrasis evaluation protocol to demonstrate relational models' transferability towards unexpected relations. The protocol posits that optimally transferable models should effectively predict entities suitable for antiphrases, even with semantic shifts brought by nonliteral rhetorical probes. The experiments show that Bi-Link achieves competitive graph reasoning results across benchmark datasets, and the proposed antiphrasis evaluation opens a path for further exploration of unusual relations.

Keywords: representation learning, knowledge graph reasoning, inference speedup, antiphrasis test.

1. Introduction

Knowledge graphs (KGs) are structured databases that encapsulate facts, where entities are represented as nodes and their relations as edges (Xiong et al., 2017). A fundamental problem in evolving KGs is to predict missing facts by reasoning with existing facts, a task known as knowledge graph reasoning (Hwang et al., 2021). Within this context, machine learning models for KG reasoning generally subscribe to one of two paradigms: transductive and inductive (Ji et al., 2021). In the transductive paradigm, models predict missing relations among recognized entities, whereas in the inductive paradigm, models extrapolate toward unseen entities or open-domain relations (Daza et al., 2021a). Pre-trained language models (PLMs) (Vaswani et al., 2017) have emerged as potent tools for KG reasoning with their strong reasoning capabilities (Jiang et al., 2023; Yang et al., 2022) in high-performance inference. However, this comes at the cost of computational efficiency, largely due to the quadratic time complexity inherent in the transformer's self-attention mechanism (Keles et al., 2023). Additionally, the generalizing and transfering abilities of these models, especially when faced with novel relations, remains an open question (Oh et al., 2022).

This paper aims to address the imperative challenges limiting the efficient deployment of PLMs in knowledge graph reasoning. Our exploration is anchored around three research questions:

- **RQ1**: How can we accelerate the inference of PLM-based knowledge graph reasoning?
- **RQ2**: What methodology can facilitate the collection of antiphrasis relations, a family of out-of-distribution samples that embody contradictory scenarios?
- **RQ3**: Which evaluation strategy can gauge the zero-shot performance of knowledge graph reasoning models for antiphrasis relations?

To address the first research question, we draw upon previous findings that database designers distinguish between intra-entity attributes and inter-entity relations as distinct constructs (Weber, 1996) and that some PLMs can encode certain factual knowledge before finetuning (Petroni et al., 2019). We hypothesize that, while intra-entity attributes, such as the hierarchical relation between a department and a school, can be pre-computed or pre-activated, inter-entity relations require online reasoning. We introduce a *bi-stage prefix tuning* approach, Bi-Link, which employs two sets of prefixes (Li and Liang, 2021) to facilitate both precomputation and relational reasoning. Specifically, an initial set of prefixes guides a static PLM to serve as an entity encoder, compressing pre-activated entities in a *memory bank*. We name this phase *entity pre-activation*. The subsequent stage utilizes relation prefixes, enabling the PLM to amalgamate the entity with the relation, thus fostering efficient query tensor formulation through prepending learnable prefixes in the attention mechanism. Leveraging a contrastive training objective (Chen et al., 2020), we train the prefixes of both stages end-to-end. Our method strives for a reduced latency while preserving prediction accuracy.

To address the multifaceted issue of model transferability, we propose an innovative *antiphrasis* evaluation protocol inspired by logical inference probing (Tenney et al., 2019). Antiphrasis is a rhetorical device (Blanco, 2015) that employs words

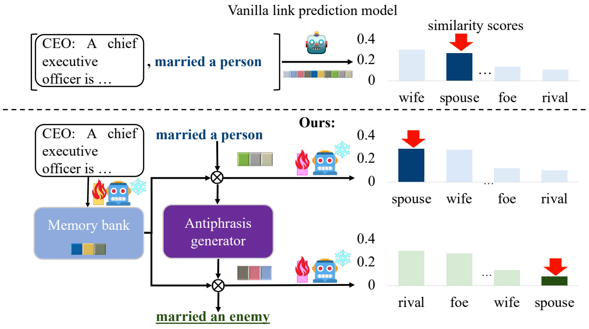


Figure 1: Antiphrasis evaluation for out-of-distribution generalization on graphs. The model leverages similarity scores between the vector encoding the entity and the relation to reason about tail entities. We sample (CEO, married a person, spouse) from the Knowledge Graph (KG) as a reference, while (CEO, married an enemy, ,) serves as an antiphrasis to evaluate transferability, as indicated by the different ranks of the tail entity, "spouse".

or phrases to express meanings diametrically opposed to their literal interpretations. For example, changing "marry a person" to "marry an enemy" reflects a shift from would denote a transition from the conventional notion of "marriage" to a more complex idea of "conscious betrayal" in the realm of commercial rivalry (de Scudéry, 1975).

As shown in Figure 1, we introduce an antiphrasis relation generator to collect out-of-distribution relations. This serves to scrutinize whether the model genuinely understands complex relational semantics or merely memorizes patterns from the training data. Based on the closed-world assumption (Reiter, 1980), we posit that a fully transferable model should be able to predict entities that are better suited for antiphrasis scenarios, even if they diverge significantly from the original entities within the knowledge graph. Importantly, we consider a performance decline for these antiphrasis relations as an indicator of enhanced transferability, as it implies successful generalization beyond the training set. Our contributions are threefold:

- We propose **Bi-Link**, a novel bi-stage prefix tuning approach for knowledge graph reasoning which addresses space-time tradeoff by learning separate prefixes for entities and relations.

- To evaluate transferability on relations, we design an antiphrasis evaluation protocol to collect antiphrasis relations. We propose a new test strategy that considers performance drop as improvement on generalization.

- Our empirical study shows that Bi-Link can effectively reduce latency across different reasoning setups. The approach can address antiphrasis relations with inference speedup modifications.

2. Related work

Inductive knowledge graph reasoning Inductive knowledge representation learning is challenging because the validation and test sets may include entities not seen during training. DKRL (Xie et al., 2016) first attempts to address this issue by using a convolutional network to encode entities leveraging text descriptions. Another solution is training entity encoders with graph neural networks, as in GraphSAGE (Hamilton et al., 2017). However, this line of work has limitations as they require a fixed set of attributes before training (Daza et al., 2021b). Aggregating neighbourhood information through a GNN is one way to encode entities (Hamaguchi et al., 2017). However, these approaches require unseen entities to be surrounded by known entities as neighbours and fail to handle entirely new graphs (Markowitz et al., 2022). KG-BERT (Yao et al., 2019) was the first approach to use pretrained LMs for knowledge base tasks. However, its inference time goes quadratically with the number of entities. To address this cost, MLMLM (Clouatre et al., 2021) optimizes the inference time with a look-up table. Following works (Wang et al., 2021b; Markowitz et al., 2022) optimize learning time scalability with structural objectives. In this paper, we improve the inference performance of inductive KG reasoning tasks with a bi-stage Prefix-Tuning framework.

Contrastive representation learning Inspired by the contrastive noise estimation principle (Gutmann and Hyvärinen, 2010), CPC (Oord et al., 2018) and SimCLR (Chen et al., 2020) learn robust contrastive representations for varied data types. SimCSE (Gao et al., 2021) adapts the method from vision to textual similarity by simplifying previous contrastive sentence embedding methods using dropout (Srivastava et al., 2014) to generate contrastive samples. PromptBERT (Jiang et al., 2022) further improves the results with hard prompts which are essentially handcrafted relation templates. However, prompt tuning (Lester et al., 2021) and Prefix Tuning (Liu et al., 2021) achieved parameter-efficient prompt tuning on text generation and classification by embedding soft prompts in the text input or prefixes in the attention heads. Ongoing research (Tam et al., 2022) aims to determine the most effective way of using the Bi-Encoder paradigm for this task. This work introduces a bi-stage prefix tuning method to reduce inference latency while balancing space-time trade-off. Our method facilitates combining large language models (LLMs) and knowledge graphs and improves practical usability.

Antiphrasis and sarcasm detection. Antiphrasis (Dupriez, 1991) is a rhetorical device where the intention is expressed through the opposite of what

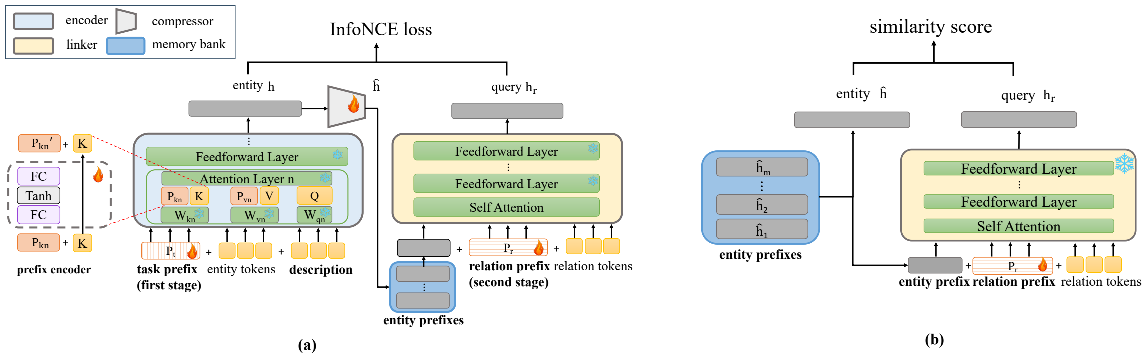


Figure 2: An overview of the bi-stage Prefix-Tuning framework, Bi-Link. In the first Prefix-Tuning stage, entity prefixes p_t prompt a static LLM, encoding entity context as a vector h, and then fusing it with relation tokens as a query vector h_r. In the second stage, relation prefixes p_r prompt the same LLM for another round of Prefix-Tuning for inference speedup as shown in (a). During inference, Bi-Link only needs the second stage with pre-computed entities stored in a memory bank, as shown in (b). Bi-Link is adaptable for both BERT and GPT models.

is stated, as a particular form of multi-word expressions (MWE), as shown in Table 4. It is closely related to ironic and sarcastic language, widely used in informal language. Those figures of speech still challenge social media understanding applications (Joshi et al., 2017; Kannangara, 2018; Küçük and Can, 2020). However, recent advances in language modelling have led to significant improvements in detecting them. For example, (Misra and Arora, 2019) learns salient features from word embeddings to detect sarcasm. In contrast, (Potamias et al., 2020) trained a recurrent transformer model to accomplish the same task. Recently, (Zhang et al., 2023) combined BERT (Devlin et al., 2018) with a graph attention mechanism to incorporate structured information. In this paper, we introduce antiphrases as probes to evaluate relation models learnt from KGs, demonstrating our model adaptivity to distinguish similar, yet different, relations correctly.

3. Method

Preliminary Given a knowledge graph $\mathcal{G} = (V, R, E)$ with $|V|$ observed entities, $|R|$ relation types, and edged $E = \{(e, r, t)\}, e, t \in V$, the inductive knowledge graph reasoning task is to infer a missing edge (h, r, t) from an emerging entity $h \notin V$ to an existing entity $t \in V$.

Bi-stage Relation Reasoning Figure 2 exhibits an overview of bi-stage prefix tuning for knowledge graph reasoning tasks[1]. The method leverages different prefixes to learn intra-entity and relation representations separately at two stages. Since the entity representations can be precomputed, this novel framework reduces inference latency.

As shown in Figure 2 (a), we tune task-related prefixes by applying vanilla Prefix-Tuning (Li and Liang, 2021) in the first stage, feeding the entity context to a static LLM to obtain representation h. This representation is then encoded with relation tokens r as query h_r to link the tail entity t. The contrastive learning process maximise the following conditional probability,

$$p_t' = \underset{p_t \in \mathbb{R}^{l \times 2 \times c \times d}}{\arg\max} \ \mathbb{E}[\log P(t \mid p_t, r, h)] \tag{1}$$

where p_t represents entity prefixes, l is the number of layers, c is the prefix length, and d is the hidden dimension. h corresponds to tokens for the head entity. The prefix encoder contains two fully connected layers with Tanh nonlinearity. The prefixes p_t are mapped as $p_k n$ and $p_v n$, efficiently interacting with entity context via self-attention mechanism (Vaswani et al., 2017).

In the second stage, we learn another set of relation prefixes p_r to prompt the same LLM to encode relation tokens r and compressed entity representation $\hat{h}$, maximizing the log-likelihood as follows,

$$p_r' = \underset{p_r \in \mathbb{R}^{l \times 2 \times z \times d}}{\arg\max} \ \mathbb{E}\left[\log P\left(t \mid p_r, r, \hat{h}\right)\right] \tag{2}$$

where p_r represents relation prefixes, $\hat{h}$ is the compressed representation selected by a multi-layer perceptron similar with (Lee et al., 2022).

Both stages are tuned with supervised contrastive loss (Chen et al., 2020), pulling together labeled representations h_r and t. The training objectives can be written as follows,

$$L = L_1 + L_2 \tag{3}$$

$$\mathcal{L}_1 = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} [\log e^{s(h,r,t_i)} - \log \sum_{j \in \mathcal{B}_{\setminus i}} e^{s(h,r,t_j)}] \tag{4}$$

$$\mathcal{L}_2 = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} [\log e^{s(\hat{h},r,t_i)} - \log \sum_{j \in \mathcal{B}_{\setminus i}} e^{s(\hat{h},r,t_j)}] \tag{5}$$

[1] The code is available at `https://anonymous.4open.science/r/Bi-Link-757/`.

where $\mathcal{B}$ denotes the current batch, t_i is the positive tail while t_j is the negative tail candidates

During inference, as shown in Figure 2, Bi-Link only needs to run the second reasoning process by calling a pre-activated representation from a memory bank, encoding it with the queried relation embedding as a query vector h_r. Bi-Link can further retrieve answers with this query vector.

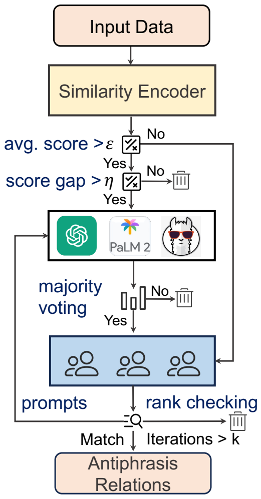
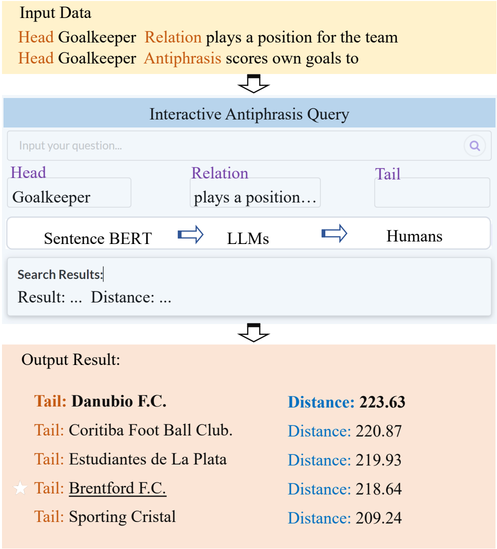


Figure 3: Antiphrasis collection protocol has three steps. The first step measures the similarity score between antiphrasis and tails with a similarity encoder. A large enough score gap might indicate antiphrasis and encourage the candidate to go through the rest of checking. The second step employs LLMs to detect antiphrasis with a majority voting. An agreement for antiphrasis will allow the candidate to be passed to human examiners.

Antiphrasis evaluation Inspired by linguistic probes(Chen and Gao, 2022), we propose a novel evaluation protocol measuring the transferability of KG reasoning models in handling novel triples and identifying dissonant relations. In this method, we modify relation spans with antiphrasis, thereby changing the original triples significantly.

As shown in Figure 3, our semi-automatic process for collecting antiphrasis relations includes three steps. First, we measure the semantic gap, indicated by similarity score, between candidate antiphrasis and tails, e.g., "CEO married an enemy", "spouse" and "wife". As antiphrasis brings a huge semantic shift, the similarity score of the original tail, "spouse", will drop dramatically, with the similarity score gap as a clue. After similarity filtering, we further check these rhetorical relations with the help of LLMs via majority voting. Finally, an agreement from LLMs will encourage the candidate to enter the human examination step where human markers assess the coherence of the antiphrasis within the triple context to ensure data quality and detect mode collapse (Huang et al., 2020). An example of our collection process is shown in Figure 4.



Input Data

Head Goalkeeper Relation plays a position for the team

Head Goalkeeper Antiphrasis scores own goals to

Interactive Antiphrasis Query

Input your question...

Head Goalkeeper

Relation plays a position...

Tail

Sentence BERT ⇨ LLMs ⇨ Humans

Search Results:

Result: ... Distance: ...

Output Result:

Tail: Danubio F.C. **Distance: 223.63**

Tail: Coritiba Foot Ball Club. Distance: 220.87

Tail: Estudiantes de La Plata Distance: 219.93

☆ Tail: Brentford F.C. Distance: 218.64

Tail: Sporting Cristal Distance: 209.24

Figure 4: Antiphrasis collection platform. The rank drop of the labeled tail entity, marked by a star, identifies the presence of antiphrasis. The top-ranked tail to the new context is bolded.

Formally, the antiphrasis collection problem is defined as finding the best antiphrasis relation r_A that maximizes the conditional likelihood of the triple given a head entity h and tail entities $\mathcal{T}$, denoted as

$$P(y|\, h, r_A, \mathcal{T}) \tag{6}$$

where y is the label for antiphrasis identification, and $\mathcal{T} = [t_0, \ldots, t_n]$ is the candidate set for ranked entities.

The antiphrasis probing evaluation measures the discrepancy between the latest and original reasoning probabilities. The corresponding score gap s can be written as follows,

$$s = P(t|h, r_A) - \frac{1}{k}\sum_{i=1}^{k} P(t_i|h, r_A) \tag{7}$$

where t is the original labeled tail entity, and t_i belongs to the top-k retrieved tail entities. Please refer to Appendix A.1 for more implementation details of the antiphrasis collection.

4. Experimental Setup

4.1. Datasets and Baselines

We evaluate Bi-Link on three representative knowledge graphs with transductive and inductive setups. Table 1 shows the statistics of these datasets. WN18RR (Bordes et al., 2013) is a curated subgraph of WordNet (Miller, 1995), a knowledge graph of lexical relations among English words. FB15k-237 (Toutanova et al., 2015) is a subgraph of Freebase (Bollacker et al., 2008), a knowledge base containing diverse facts. Wikidata5m (Wang

Dataset	WN18RR		FB15k-237		Wikidata5m	
	Transductive	Inductive	Transductive	Inductive	Transductive	Inductive
$\|V\| + \|R\|$	40,943+11	7553+9	14,541+237	6683+219	4, 594k+822	4, 579k+822
#training	86,835	7,940	272,115	27,203	20,614k	20,496k
#validation	3,034	1,394	17, 535	1,416	5,163	6,699
#test	3,134	1,429	20, 466	1,424	5,163	6,894

Table 1: Statistics of datasets. $|V|$, $|R|$, and # denote the numbers of entities, relations, and triples, respectively. In the inductive setups, test entities are unseen, and disconnected from the training graphs.

	FB15k-237				WN18RR				Wikidata5m			
	PRR	MRR	Hit@1	Hit@10	PRR	MRR	Hit@1	Hit@10	PRR	MRR	Hit@1	Hit@10
TransE (Bordes et al., 2013)		27.9	19.8	44.1		22.3	1.3	53.1		25.3	17.0	39.1
DistMult (Yang et al., 2014)		24.1	15.5	41.9		42.5	19.8	49.1		25.7	20.9	33.4
ComplEx (Trouillon et al., 2016)		27.1	18.4	44.7		44.6	41.0	50.2		28.1	22.8	37.3
RotatE (Sun et al., 2019)		30.4	21.6	47.9		47.2	42.8	56.5		29.0	23.4	39.0
DKRL-BERT (Xie et al., 2016)		14.4	8.4	26.3		13.9	4.8	16.9		16.0	12.0	22.9
MLMLM (Clouatre et al., 2021)		25.9	18.7	40.3		50.2	43.9	61.1		22.3	20.1	26.4
KEPLER (Wang et al., 2021b)		13.9	9.2	28.4		43.2	40.7	52.6		15.4	10.5	24.4
BLP (Daza et al., 2021a)		19.5	11.3	36.3		28.5	13.5	58.0		31.9	25.7	38.5
RAILD (Gesese et al., 2022)		21.6	12.7	39.7		29.1	13.6	59.9		31.4	26.8	37.9
SimKGC (Wang et al., 2022)		**33.6**	**24.9**	**51.1**		**66.6**	**58.7**	**80.0**		**35.8**	**31.3**	**44.1**
Adapted Prompt-Tuning (Ours)	8.3	3.4	1.0	5.5	16.9	13.3	4.6	18.2	19.7	6.6	1.9	15.2
Adapted Prefix-Tuning (Ours)	88.3	29.1	20.8	48.4	85.8	55.2	48.9	72.9	89.3	31.9	28.4	38.9
Bi-Link (Ours)	**91.9**	30.8	21.2	50.6	**91.1**	56.8	50.9	79.6	**92.4**	32.2	30.1	40.2

Table 2: Performance comparison on three transductive reasoning datasets in terms of MRR (%), Hit@1 (%) and Hit@10 (%). Our work adapts Prompt-Tuning and Prefix-Tuning as efficient tuning baselines by learning for each relation a dedicated prompt or prefix, respectively. We report the performance retention ratio, PRR (%), between the efficient tunings and the baseline model SimKGC. The best results are in bold, while the second-best are underlined.

et al., 2021b) is a large-scale KG constructed from Wikidata (Vrandečić and Krötzsch, 2014) and Wikipedia (Lehmann et al., 2015). These KGs have different scales with the number of nodes varying from ∼15k to ∼5M. We compare the performance of our principal method, Bi-Link, with static representation methods and transformer-based methods (Yao et al., 2019; Clouatre et al., 2021). We provide implementation details of Bi-Link and antiphrasis evaluation in Appendix A.1 and A.2.

4.2. Evaluation Metrics

Following the standard evaluation protocol in (Wang et al., 2021b), we measure the reasoning performance with mean reciprocal rank (MRR) and Hits@k scores (Radev et al., 2002). MRR computes the average reciprocal ranks of the labeled entities, while Hits@k calculates the retrieval accuracy when the labeled entity ranks among the top-k. We report the performance retention ratio (PRR) (Xue and Aletras, 2023) as the ratio between predictive performance compared to the upper bound baseline.

According to the closed world assumption (CWA) (Reiter, 1980), triples described by antiphrases should also be in the same KG if they exist. The correct answer should be very different from the original tail due to semantic discrepancy. Therefore, we use retrieval performance drop to estimate the transferability of antiphrasis. The lower retrieval accuracy reveals higher transferability, and hence, down is up!

	Wikidata5m Inductive					
	PRR	Size	MRR	Hit@1	Hit@10	Latency
SentenceTransformer		0.46×	22.2	0.0	57.1	0.4×
DKRL-BERT		0.57×	32.2	9.7	72.0	0.4×
MLMLM		0.49×	28.4	22.6	34.8	> 9.9×
KEPLER		0.55×	35.1	15.4	71.9	1.8×
BLP		0.55×	47.8	24.1	87.1	1.7×
RAILD		0.31×	45.5	22.0	84.9	1.1×
SimKGC		1.00×	**60.1**	**39.4**	**92.4**	1.0×
Prompt-Tuning	10.75	**0.02**×	6.6	1.9	15.2	7.7×
Prefix-Tuning	86.33	0.14×	53.9	30.5	84.9	9.0×
Bi-Link (Ours)	**95.52**	0.14×	59.1	35.7	90.2	**0.3**×

Table 3: Results on the Wikidata5m inductive setup. Performance retention ratio (%), trainable parameter size, test latency are shown as ratio to the baseline SimKGC.

4.3. Results

Link prediction The performance of knowledge representations in transductive and inductive reasoning settings are presented in Table 2 and Table 3 respectively. Bi-Link achieves over 90% in performance retention ratio (PRR) compared to the finetuned baseline, SimKGC, while accelerating inference by three times. Bi-Link shows strong performance to texted-based models, effectively improves KG reasoning speed across four datasets. Specifically, Bi-Link achieves 92.4% PRR compared with Prompt-Tuning's 19.7% PRR. The comparison demonstrates that Bi-Link provide more effecient solutions than PEFT baselines for learning entity and relation representations in separate attention spaces. Table 5 shows slightly lower results obtained by GPTs.

Original triple: (Brentford F.C. , has a top scorer playing, **Forward**)
Antiphrasis: (Brentford F.C. , was scored own goals by, ...)
Entity description: Brentford F.C. is a football club in the London Borough of Hounslow, that plays in Football League One.
Original tail: Forwards are the players who play nearest to the opponents' goal, and are aiming for scoring goals.

Method	Predictive tails
SimKGC	**Forward** : Forwards are the players who play nearest to the opponents' goal, and are aiming for scoring goals. Midfielder : A midfielder is generally positioned on the field between their team's defense and forwards. Defender : A defender is an outfield player whose primary role is to prevent the opposition from attacking.
Bi-Link GPT	Goalkeeper : Goalkeeper, often shortened to keeper or goalie, is one of the major positions of football. Defender : A defender is an outfield player whose primary role is to prevent the opposition from attacking. **Forward** : Forwards are the players who play nearest to the opponents' goal, and are aiming for scoring goals.

Table 4: An error example of antiphrasis evaluation, with the original relation "has a top scorer playing" replaced by an antiphrasis relation "was scored own goals by". Huge semantic misalignment makes the original tail "Forward" unreasonable, so it should be ranked down in replacement for entities that suit the rhetoric expression.

	FB15k-237		WN18RR	
	MRR	Hit@10	MRR	Hit@10
SimKGC BERT	33.6	51.1	66.6	80.0
Contrastive GPT (Ours)	27.3	43.6	64.8	88.1
Bi-Link BERT (Ours)	30.8	50.6	56.8	79.6
Bi-Link GPT (Ours)	22.9	35.4	59.7	77.6

Table 5: Comparison between GPT-2 and BERT base.

Error Analysis Table 5 compares GPT and BERT trained with constrastive loss and bi-stage Prefix-Tuning. GPTs show slightly weaker results on both FB15k237 and WN18RR. We analyse the distribution of relevance scores of error samples. The relevance score is computed with the cosine similarity between the wrongly predicted sample and the labeled sample. GPT shows a lower mean and higher variance than BERT, meaning the model might have a greater potential to predict highly relevant entities. Bi-Link shows highly relevant recalled entities using Prefix-Tuning. Table 6 shows an example where BERT predicts wrongly but GPT predicts correctly. This might reveals a better semantics understanding to the perturbed relations from GPTs. We show the difficulty of error cases is at the same level in Appendices B Figure B.

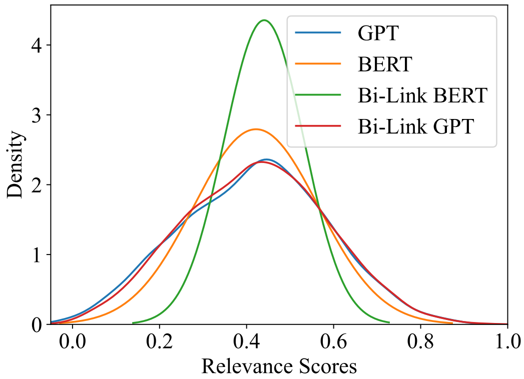


Figure 5: Different semantic distributions of predicted tails by GPT and BERT on WN18RR. GPT excels in predicting most relevant tails. Bi-Link BERT show a high mean relevance score, indicating the model might predict tails more semantically related to labeled entities.

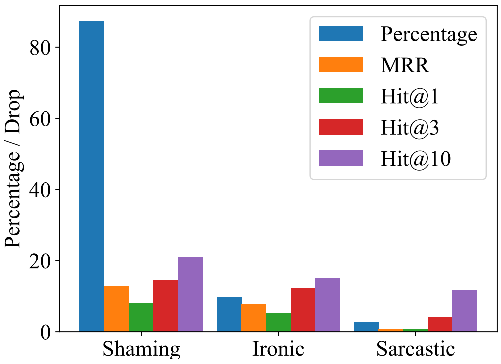


Figure 6: Performance drop of Bi-Link in antiphrasis relations on FB15k237 data. The performance drop indicates the model might successfully detect shaming relations due to their semantic gaps with normal relations.

	FB15k-237 Transductive			
	MRR↓	Hit@1↓	Hit@3↓	Hit@10↓
SentenceTransformer	1.8	0.0	2.5	5.3
DKRL-BERT	2.1	0.0	3.6	6.0
BLP	4.2	3.1	4.9	7.5
SimKGC	**13.6**	**9.5**	**15.3**	**21.9**
Prompt Tuning	1.9	0.0	3.1	5.7
Prefix Tuning	9.8	6.7	12.1	17.5
Bi-Link (Ours)	12.0	7.6	13.9	20.1

Table 6: Performance drop of models in antiphrasis relation evaluation. Higher performance drop is better.

Antiphrasis evaluation We probe relational models with antiphrasis relations on the transductive link prediction setting of FB15k237 dataset to assess out-of-distribution reasoning abilities. We show the adaptivity estimated by performance drop in Table 6 and Figure B with the following findings. The distribution of antihprases is imbalanced across different categories, with a predominant on shaming antiphrases, such as "marry an enemy" or "cheat to win". In general, antiphrasis probes only change the retrieval performance by less than 20% on Hit@10, meaning the relational models cannot understand these novel relations very well. Probing accuracy on frozen SentenceTransformer remains unchanged, indicating antiphrasis relational reason-

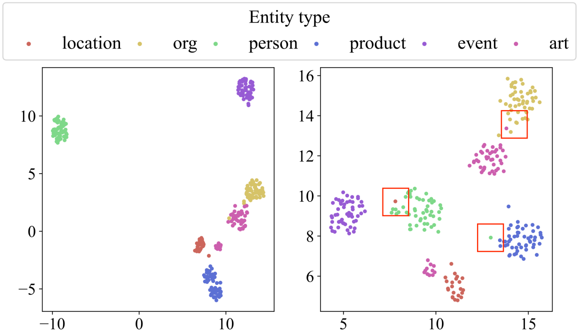


Figure 7: Entity representations learnt by Bi-Link (a) and SentenceTransformer (b) from Wikidata5m

ing does not benefit much from masked language modeling or sentence similarity training tasks. By contrast, Bi-Link and SimKGC show a significant drop in antiphrasis probes after learning knowledge graph reasoning models, which demonstrates enhanced understanding between relation concepts and text description. Antiphrases of the shaming class cause a significant performance drop with over 20% to the shaming class, suggesting more subtle reasoning steps. At the same time, the models are less sensitive to the sarcastic or ironic class. This comparison demonstrates certain contradictory word pairs may have more co-occurrence during training. The models can, therefore, identify such antiphrasis relations through compositionality. Antiphrasis will contribute to bias evaluation. For example, a relational model consistently relates "cheats to win" with "Iron Man 3" and "Spider-Man 2", indicating bias toward movie sequels. As shown in Figure B, Bi-Link's struggles with ironic and sarcastic relations underscore the difficulty of this new generalization task. More antiphrasis relation results are presented in Appendix D.

Ablation study To investigate the contribution of each component, we compare Bi-Link with different variants in terms of retrieval performance on the transductive setup of Wikidata5m. Specifically, we modify Bi-Link by removing the first stage prefix tuning (w/o first PT), the second stage prefix tuning (w/o second PT), and removing relation texts (w/o relation texts), respectively. We show their results in Table 7 with the following findings.

The improvements confirm that entity-related and relation prefixes encode adequate information for discovering missing links. Compared with the baselines, SimKGC and SentenceTransformer, Bi-Link reduces inference latency while achieving a high-performance retention ratio. Compared with w/o first stage Prefix-Tuning (PT), the inference latency is significantly reduced by 30 times, demonstrating that prefixes of the first stage play an essential role in promoting inference speed.

	Wikidata5m Transductive					
	PRR	MRR	H@1	H@3	H@10	Latency
SentenceTransformer		6.4	0.0	9.4	18.2	0.4×
SimKGC		**35.8**	**31.3**	**37.6**	**44.1**	1.0×
w/o first stage PT	83.19	29.7	25.9	31.6	36.6	9.0×
w/o second stage PT	31.58	10.3	0.9	14.2	25.1	**0.2**×
w/o relation texts	78.12	28.0	23.9	29.7	34.8	0.3×
Bi-Link	**92.39**	32.2	30.1	34.7	40.2	0.3×

Table 7: An ablation study on the Wikidata5m transductive. PRR and latency are compared to SimKGC.

The PRR of Bi-Link is, whereas the PRR of w/o second stage PT averages only, demonstrating prefixes of the second stage are vital for relational reasoning.

Visualisation Figure 7 compares the entity representations of Bi-Link with the frozen baseline method on Wikidata5m. We plot the scatter plot with UMAP (McInnes et al., 2018). Despite being trained on individual instances, the representations naturally form clusters corresponding to meaningful entity types. Each cluster has around 50 samples. While there are outliers, particularly among art-related entities due to their lexical overlap with other types, the overall representation quality is better than the frozen model, underscoring the effectiveness of our approach.

4.4. Discussion

Prefix-Tuning provides more effective KG reasoning in a learnt attention space. We show the superiority of selecting prefixes over prompts as the learnable modules in Figure 9. The number of learnable parameters for both methods increases linearly with the prompt length, but their slopes reconfirm their efficiency differences. Despite more parameters to tune, prefix tuning yields better performance as it can adapt knowledge graph reasoning in the attention mechanism. This performance gap highlights that learning prompts in the input space are insufficient for adapting a static PLM to knowledge graph reasoning, emphasizing the necessity for deeper prompting, which Bi-Link can offer. The results of inductive reasoning further show the effectiveness of separately modelling entity and relation representations. Relational prefix tuning allows for concentration on relations and efficient extraction of representative features from relational descriptions.

Optimising prefix length reduces overfitting. Prefixes and prefix encoders are tunable modules in the proposed method. Figure 8 shows the relation between prefix length and performance.

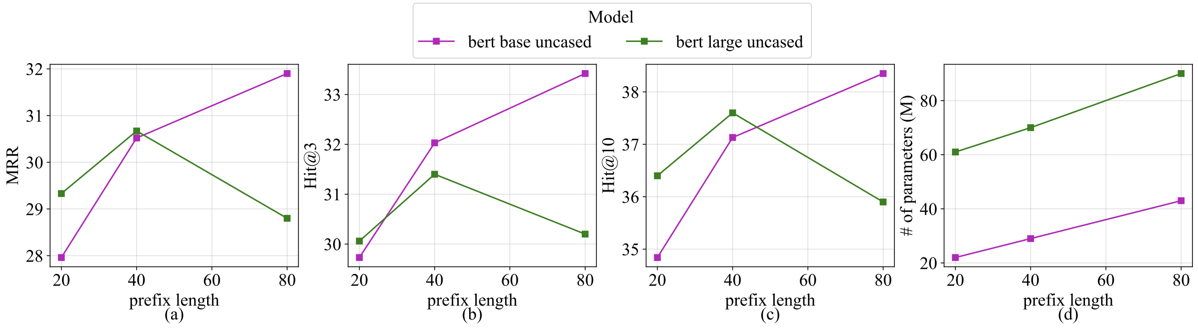


Figure 8: Relation between prefix length and model generalization on Wikidata5m transductive reasoning dataset. Shorter prefixes lead to underfitted expressiveness, while overly long prefixes lead to overfitting on training texts.

Compared with the best prefix length, longer prefixes correspond to a more delicate relation model for complex relations, but lengthy prefixes can become over-parameterized, resulting in overfitting. By contrast, overly short prefixes may lead to underfitted performance. Notably, larger backbones are prone to over-parameterization during prompt tuning, as the turning points arrive earlier.

Find right representations for compression. The compression rate (CR) is defined as the ratio between the sequence length of entity description before and after compression as, $CR = seq\_len(h)/seq\_len(\hat{h})$. Figure 10 shows the effects of compression rate on inference performance on Wikidata5m. Our experiments uncover that the prefixes play a dual role in this context. It serves as instructive guidance for compression while also playing a second role as a data filtering function to remove noisy data during prefix tuning. A compression rate larger than two will cause collapse to Bi-Link, indicating memory expense for inference speedup.

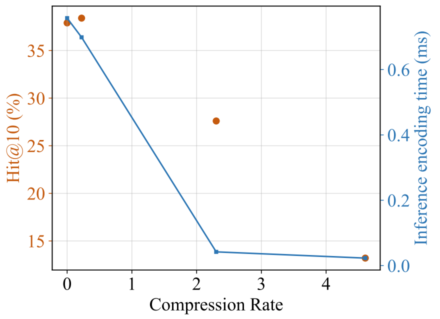


Figure 10: Compression rate on inference performance. The inference time shows relation encoding only.

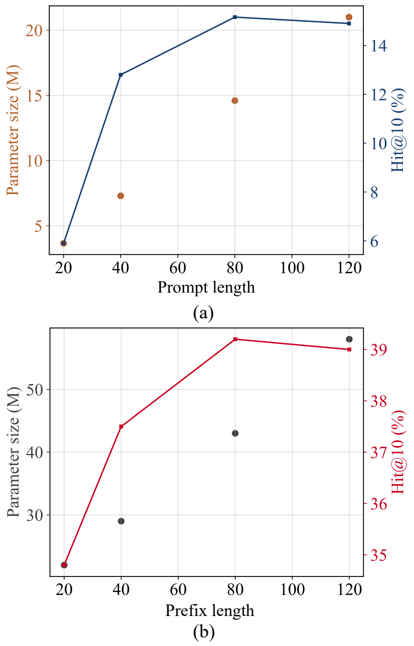


Figure 9: Comparison between prompts and prefixes

5. Conclusion

This paper proposes Bi-Link, a novel bi-stage prefix tuning method for text-based knowledge graph reasoning. Bi-Link optimizes inference speed and generalization through a two-stage process: pre-activating entities in the first stage and reasoning about relations in the second stage. Our experiments show competitive knowledge graph reasoning results with significant inference speed improvements.

Future Work Antiphrasis evaluation gives a new dimension to LLM evaluation. Because modern GPTs perform well in following instructions, future work can explore causal inference to GPTs to improve zero-shot abilities in novel relation understanding.

Ethics Statement This research adheres to the ethical standards outlined in the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct (Association et al., 2016), ensuring that all datasets employed were acquired through legitimate means and utilized in compliance with their respective licenses.

6. References

American Psychological Association et al. 2016. Ethical principles of psychologists and code of conduct.

Carmen Mellado Blanco. 2015. Antiphrasis-based comparative constructional idioms in spanish. *Journal of Social Sciences*, 11(3):111.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Zeming Chen and Qiyue Gao. 2022. Probing linguistic information for logical inference in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10509–10517.

Louis Clouatre, Philippe Trempe, Amal Zouaq, and Sarath Chandar. 2021. MLMLM: Link prediction with mean likelihood masked language model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4321–4331, Online. Association for Computational Linguistics.

Tom Dalzell. 2014. English-language idioms.

Mark Davies. 2015. English corpora.

Daniel Daza, Michael Cochez, and Paul Groth. 2021a. Inductive entity representations from text via link prediction. In *Proceedings of the Web Conference 2021*, pages 798–808.

Daniel Daza, Michael Cochez, and Paul Groth. 2021b. Inductive entity representations from text via link prediction. In *Proceedings of the Web Conference 2021*, WWW '21, page 798–808, New York, NY, USA. Association for Computing Machinery.

Madeleine de Scudéry. 1975. *Clelia, an Excellent New Romance the Whole Work in Five Parts, Dedicated to Mademoiselle de Longueville*. London:: Printed and are to be sold by H. Herringman, D. Newman, T. Cockerel ....

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bernard Marie Dupriez. 1991. *A dictionary of literary devices: Gradus, AZ*. University of Toronto Press.

Mikhail Galkin, Xinyu Yuan, Hesham Mostafa, Jian Tang, and Zhaocheng Zhu. 2023. Towards foundation models for knowledge graph reasoning. *ArXiv*, abs/2310.04562.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Genet Asefa Gesese, Harald Sack, and Mehwish Alam. 2022. Raild: Towards leveraging relation features for inductive link prediction in knowledge graphs. *arXiv preprint arXiv:2211.11407*.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Takuo Hamaguchi, Hidekazu Oiwa, Masashi Shimbo, and Yuji Matsumoto. 2017. Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach. *arXiv preprint arXiv:1706.05674*.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. 2020. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. In *International Conference on Computational Linguistics*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: on symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514.

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. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11161–11180, Toronto, Canada. Association for Computational Linguistics.

Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2022. Promptbert: Improving bert sentence embeddings with prompts. *arXiv preprint arXiv:2201.04337*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Sandeepa Kannangara. 2018. Mining twitter for fine-grained political opinion polarity classification, ideology detection and sarcasm detection. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 751–752.

Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. 2023. On the computational complexity of self-attention. In *International Conference on Algorithmic Learning Theory*, pages 597–619. PMLR.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Jooyoung Lee, Seyoon Jeong, and Munchurl Kim. 2022. Selective compression learning of latent representations for variable-rate image compression. *Advances in Neural Information Processing Systems*, 35:13146–13157.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

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

Elan Markowitz, Keshav Balasubramanian, Mehrnoosh Mirtaheri, Murali Annavaram, Aram Galstyan, and Greg Ver Steeg. 2022. Statik: Structure and text for inductive knowledge graph completion. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 604–615.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

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

Rishabh Misra and Prahal Arora. 2019. Sarcasm detection using hybrid neural network. *arXiv preprint arXiv:1908.07414*.

Byungkook Oh, Seungmin Seo, Jimin Hwang, Dongho Lee, and Kyong-Ho Lee. 2022. Open-world knowledge graph completion for unseen entities and relations via attentive feature aggregation. *Information sciences*, 586:468–484.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. *arXiv preprint arXiv:1905.00563*.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32:17309–17320.

Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *LREC*. Citeseer.

Raymond Reiter. 1980. A logic for default reasoning. *Artificial intelligence*, 13(1-2):81–132.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Weng Lam Tam, Xiao Liu, Kaixuan Ji, Lilong Xue, Xingjian Zhang, Yuxiao Dong, Jiahua Liu, Maodi Hu, and Jie Tang. 2022. Parameter-efficient prompt tuning makes generalized and calibrated neural text retrievers. *arXiv preprint arXiv:2207.07087*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

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*, pages 1499–1509.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. 2021a. Structure-augmented text representation learning for efficient knowledge graph completion. In *Proceedings of the Web Conference 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)*, pages 4281–4294, Dublin, Ireland. Association for Computational Linguistics.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Ron Weber. 1996. Are attributes entities? a study of database designers' memory structures. *Information Systems Research*, 7(2):137–162.

Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. DeepPath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 564–573, Copenhagen, Denmark. Association for Computational Linguistics.

Huiyin Xue and Nikolaos Aletras. 2023. Pit one against many: Leveraging attention-head embeddings for parameter-efficient multi-head attention. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10355–10373, Singapore. Association for Computational Linguistics.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement subgraph reasoning for fake news detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2253–2262.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Yazhou Zhang, Dan Ma, Prayag Tiwari, Chen Zhang, Mehedi Masud, Mohammad Shorfuzzaman, and Dawei Song. 2023. Stance-level sarcasm detection with bert and stance-centered graph attention networks. *ACM Trans. Internet Technol.*, 23(2).

Appendices

A. Implementation details

A.1. Bi-Link

For a fair comparison, we employ BERT-base-uncased model as the language model backbone[2]. We pad a fixed number of neighbor entity names to encode structural knowledge to the entity description. This section discusses model selection regarding shared hyperparameters, prefix length, and compression rate. The careful choice of these parameters can significantly improve the model adaptivity and yield robust representation for downstream tasks.

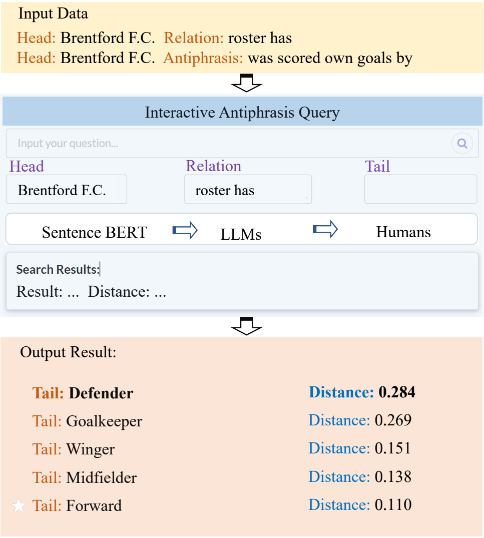
The hyperparameters shared across all datasets are shown in Table A. We use a large learning rate 1×10^{-3} for small-scale knowledge bases, such as WN18RR and FB15k-237, to avoid overfitting. For the larger knowledge base, Wikidata5m, we use a learning rate of 1×10^{-4}. We report both transductive and inductive results under the filtered setting, which ignores the scores of all known true triples of the training and validation sets. This common practice improves the ranks of tail entities of potentially new triples by removing tail entities of known triples.

Hyperparameter	Value
batch size	1024
entity document length	50
compression rate	0.8
task prefix length	8
relation prefix length	40
number of GPUs	1
memory (GB)	128
contrastive temperature	20.0
gradient clip	10.0
warmup steps	400
dropout	0.1
weight decay	0.0001
additive margin	0.02
pooling for BERT	mean
pooling for GPT	last

Table A: Hyperparameters for Bi-Link models

A.2. Antiphrasis collection platform

In this section, we elucidate the semi-automatic process used for the collection of antiphrases, aiming at evaluating the out-of-distribution extrapolation abilities of knowledge graph reasoning models. Antiphrases, capable of altering original semantics and representing unforeseen relation scenarios, serve as exciting probes for relation models learnt from knowledge bases. Figure 4 shows the platform we developed to semi-automatically collect antiphrases. As shown in Figure A, the process of collecting antiphrasis relations starts with replacing original relations with idiomatic expressions gleaned from resources such as Wiki English Idioms (Dalzell, 2014) and English-Corpora (Davies, 2015). Then we encode these rhetorical terms as vectors with a SentenceTransformer. The platform then assesses semantic similarities between the head entity, the candidate antiphrasis, and the tail entity. Following (Pezeshkpour et al., 2019), to ensure we only probe a pre-trained model without updating its parameters, we freeze the parameters of the sentence encoder during antiphrasis collection and evaluation. We set the score gap threshold as 0.4 for FB15k237 and Wikidata5m knowledge graphs. When utilizing LLMs for antiphrasis identification, we typically prompt the models to “classify the text as either an antiphrasis or a neutral sentence.” Replacing the antiphrasis with its synonym (i.e. antonym) or its hyponyms, e.g., ironic, immoral, or sarcastic, can also yield valid prompts. Ensemble these prompts can reduce the variance of predictions. Here, we utilize both zero-shot and few-shot hard prompts to guide LLMs to detect antiphrases and form a voting. Finally, human evaluators are recruited to judge the sense and appropriateness of antiphrases in the context from two aspects: i) human annotators detect potential collapse of the representations. ii) human markers assess the coherence of the antiphrases within the context.


Input Data
Head: Brentford F.C. Relation: roster has
Head: Brentford F.C. Antiphrasis: was scored own goals by

Brentford F.C.
roster has

Search Results:
Result: ... Distance: ...

Output Result:

Tail: Defender **Distance: 0.284**
Tail: Goalkeeper Distance: 0.269
Tail: Winger Distance: 0.151
Tail: Midfielder Distance: 0.138
☆ Tail: Forward Distance: 0.110

Figure A: Semi-automatic antiphrasis collection platform.

[2] BERT-base-cased shows lower results than BERT-base-uncased in our experiments.

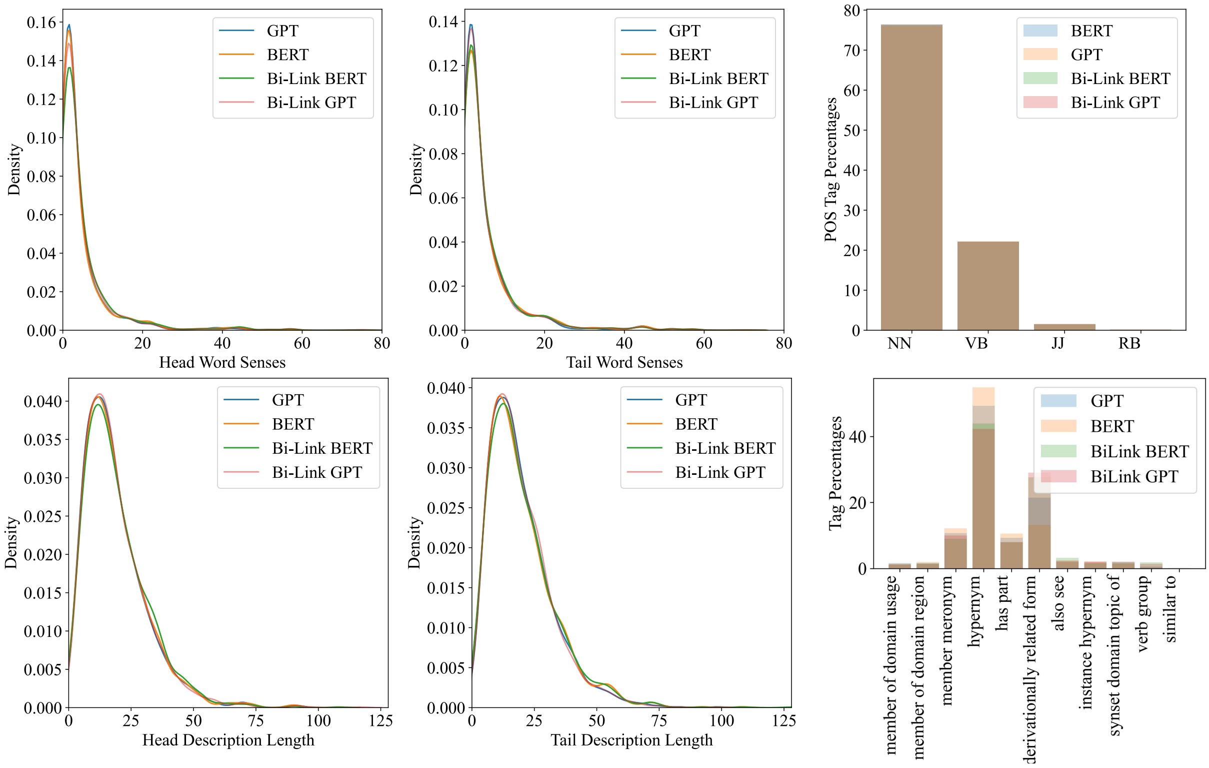


Figure B: Distribution of wrongly predictive triples in terms of word senses, part-of-speech tags and relations. Error samples made by differently tuned models show the same level of difficulty.

	FB15k-237		WN18RR		Wiki5m Inductive	
	MRR	Hit@10	MRR	Hit@10	MRR	Hit@10
StAR (Wang et al., 2021a)	26.3	45.2	36.4	64.7	20.6	33.2
ULTRA (Galkin et al., 2023)	32.5	52.8	48.0	61.4	-	-
Bi-Link - text (Test)	5.4	10.4	13.8	25.3	17.1	33.0
Bi-Link - text (Train & Test)	19.3	34.7	37.7	57.2	20.0	44.2

Table B: Performance comparison with SOTA KG reasoning methods with MRR (%), Hit@1 (%) and Hit@10 (%).

B. Further comparison

The result for comparison with more recent state-of-the-art methods is shown in Table B. For fair comparison with more recent methods, we add two more ablation setups by removing text description at development time and at test time only. The performance gap demonstrates the deficiency in understanding structure knowledge of pretrained LLMs. When removing text description only during test time, Bi-Link shows even lower performance indicating the model is vulnerable to discretely corrupted perturbations that remove a big chunk of texts from the original input, leaving it out of the distribution the model was trained on.

C. Limitation

Table B also reveals major limitations of text-based methods like Bi-Link. Bi-Link will not generalize if entity descriptions are out-of-distribution (OOD). Moreover, like most text-based methods, Bi-Link is hard to transfer to cross-lingual KG reasoning because cross-lingual models have different comprehension for two languages, which easily gives rise to fairness-related problems. In this case, methods focusing on structural reasoning will perform better with respect to fairness.

D. Antiphrasis error analysis

This section provides a text analysis of two relation models, SimKGC and Bi-Link, by investigating a range of cases. Table 4 demonstrates that the antiphrasis relations rendered by the two models significantly differ. This is particularly evident in the instance where SimKGC creates a non-existent correlation between "score own goal" and "forwards", even though the latter is often associated with scoring responsibilities in a game. This demonstrates that the baseline fails to detect the huge semantic gap existing in the novel relation because it is literally similar to "score a goal".

Comparatively, both models accurately predict the relationships among entities in the case presented in Table C due to the absence of any significant textual distraction encountered in the previous example. Nevertheless, both relation models fail to accurately predict the relations in the third example provided in Table D. This could be attributed to the fact the head entity, "outfielder", represents a defensive position in the game and is thus unlikely to be associated with "score own goals". Notably, Boston Red Sox garners substantial ranking primarily because various information sources report this team to have higher scores in the league.

The inconsistency also exists in the fourth example showcased in Table E, where Bi-Link demonstrates apparent bias towards the sequel of a movie. The results extracted from Wikidata5m show similar patterns of discordance, suggesting that Bi-Link seems to be better at learning robust features for knowledge graph reasoning. This analysis underscores the performance dip when dealing with antiphrasis relations as compared to normal relations, emphasizing the limitations in terms of generalization in both models.

Original triple: (Richard L, plays a position for the team, **Brentford Football Club**)
Antiphrasis: (Richard L, scores a goal against, ...)
Head description: Richard L is the goal keeper,one of the major positions of Brentford Football Club.
Original tail: Brentford Football Club is a football club in the London Borough of Hounslow, that plays in Football League One.

Method	Predicted tails
SimKGC	Danubio F.C. : Danubio Fútebol Club is an Uruguayan association football club based in Montevideo. Coritiba Foot Ball Club : Coritiba Football Club is a Brazilian football team from Curitiba, Paraná. Sporting Cristal : Club Sporting Cristal is a Peruvian football team. Based in the Rímac District, in the department of Lima.
Bi-Link	Sporting Cristal : Club Sporting Cristal is a Peruvian football team that plays in the Peruvian First Division. UD Las Palmas : Unión Deportiva Las Palmas, is a Spanish football team in the autonomous community of Canary Islands. Estudiantes de La Plata : Club Estudiantes de La Plata, is an Argentine professional sports club based in La Plata.

Table C: An example where both Bi-Link and the baseline model are able to detect the antiphrasis relation and reason correctly by removing the original tail entity, the player's parent club, from the candidate list.

Original triple: (Sonny Siebert, plays a position for the team, **Boston Red Sox**)
Antiphrasis: (Sonny Siebert, scores a goal against, ...)
Head description: Sonny Siebert plays in one of the three defensive positions in his Red Sox, farthest from the batter.
Original tail: Boston Red Sox is a professional baseball team based in Boston, Massachusetts.

Method	Predicted tails
SimKGC	**Boston Red Sox** : The Boston Red Sox is a professional baseball team based in Boston, Massachusetts. Chicago White Sox : The Chicago White Sox is a Major League Baseball team in the south side of Chicago, Illinois. Baltimore Orioles : The Baltimore Orioles are a professional baseball team in Baltimore, Maryland in the United States.
Bi-Link	**Boston Red Sox** : The Boston Red Sox is a professional baseball team based in Boston, Massachusetts Washington Nationals : The Washington Nationals are a professional baseball team based in Washington, D.C. Los Angeles Angels of Anaheim : The Los Angeles Angels of Anaheim is a baseball team based in California, United States.

Table D: A difficult example where both Bi-Link and the baseline model fail to predict divergent results from the original tail entity, the player's parent club. The relation models may misunderstand the shaming antiphrasis relation.

Original triple: (**Batman & Robin**, was nominated for, MTV Movie Award for Best Movie)
Antiphrasis: (... , cheats to win for, MTV Movie Award for Best Movie)
Tail description: This is a following list of the MTV Movie Award winners and nominees for Best Movie.
Original head: Batman & Robin is a 1997 American superhero film, and it is the fourth and final film of Warner Bros.

Method	Predicted tails
SimKGC	The Dark Knight Rises : The Dark Knight Rises is a 2012 British-American superhero film directed by Christopher Nolan. The Matrix : The Matrix is a 1999 American science fiction action film written by The Wachowski Brothers. Man of Steel : Man of Steel is a 2013 American superhero film directed by Zack Snyder and written by David S. Goyer.
Bi-Link	Spider-Man 2 : Spider-Man 2 is a 2004 American superhero film directed by Sam Raimi and written by Alvin Sargent. Thor : Thor is a 2011 American superhero film based on the Marvel Comics character of the same name. Iron Man 3 : Iron Man 3 is a superhero film featuring the Marvel Comics character Iron Man, produced by Marvel Studios.

Table E: An inverse relation modeling example where both relation models successfully predict different entities.

Original triple: (George Benson, marries, **Marriage**)
Antiphrasis: (George Benson, marries an enemy, ...)
Head description: George Benson is a ten-time Grammy Award-winning American musician and singer-songwriter.
Original tail: Marriage is a socially or ritually recognized union or legal contract between spouses.

Method	Evidence
SimKGC	**Marriage** : Marriage is a socially or ritually recognized union or legal contract between spouses. Common-law marriage : Common-law marriage, also known as sui juris marriage, is informal marriage. Civil union : Civil union, also referred to as civil partnership or registered partnership, is a legally recognized partnership.
DuaLink	**Marriage** : Marriage is a socially or ritually recognized union or legal contract between spouses. Civil union : Civil union, also referred to as civil partnership or registered partnership, is a legally recognized partnership. Common-law marriage : Common-law marriage, also known as sui juris marriage, is informal marriage.

Table F: A difficult example where Bi-Link and the baseline fail to predict diverse results from the original tail entity.