A Theory Guided Scaffolding Instruction Framework for LLM-Enabled Metaphor Reasoning

Yuan Tian[1,2], Nan Xu[1,3]*, Wenji Mao[1,2]*
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Beijing Wenge Technology Co., Ltd
{tianyuan2021,xunan2015,wenji.mao}@ia.ac.cn

Abstract

Metaphor detection is a challenging task in figurative language processing, which aims to distinguish between metaphorical and literal expressions in text. Existing methods tackle metaphor detection via training or fine-tuning discriminative models on labeled data. However, these approaches struggle to explain the underlying reasoning process behind the metaphorical/literal judgment. Recently, large language models (LLMs) have shown promise in language reasoning tasks. Although promising, LLM-based methods for metaphor detection and reasoning are still faced with the challenging issue of bringing the explainable concepts for metaphor reasoning and their linguistic manifestation. To fill this gap, we propose a novel **T**heory guided **S**caffolding **I**nstruction (TSI) framework that instructs an LLM to infer the underlying reasoning process of metaphor detection guided by *metaphor theories* for the first time. Our work is inspired by a pedagogical strategy called *scaffolding instruction*, which encourages educators to provide questioning and support as scaffolding so as to assist learners in constructing the understanding of pedagogical goals step by step. We first construct a metaphor knowledge graph grounded in metaphor theory, which serves as the instructional structure to obtain a series of scaffolding questions, directing the LLM to incrementally generate the reasoning process for metaphor understanding through dialogue interactions. During this theory guided instruction process, we explore the LLM's mastery boundary and provide the relevant knowledge as scaffolding support when the question is beyond the LLM's capability. Experimental results verify that our method significantly outperforms both the LLM-based reasoning methods and the SOTA methods in metaphor detection, indicating the facilitation of metaphor and instruction theories in guiding LLM-enabled reasoning process.

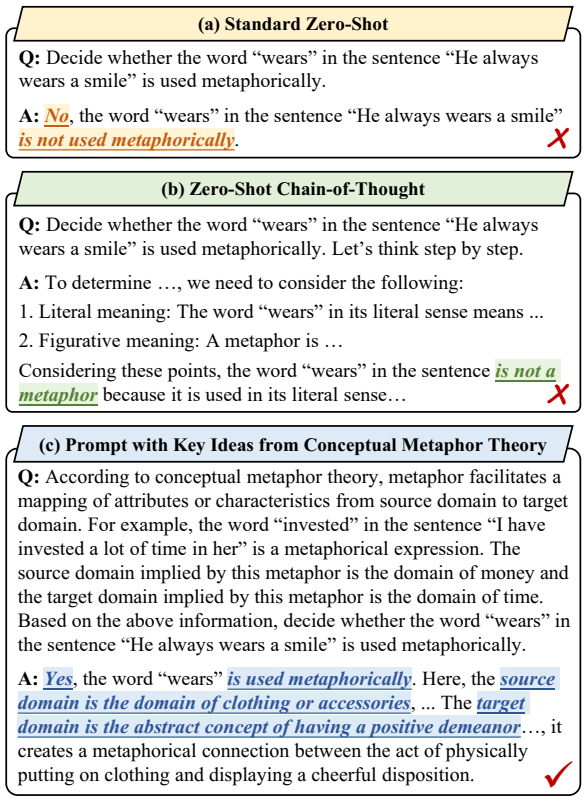

Figure 1: Example inputs and outputs generated by GPT-3.5 utilizing standard zero-shot method, zero-shot chain-of-thought method, and prompt with key ideas from conceptual metaphor theory for metaphor reasoning.

1 Introduction

Metaphor is not just a figurative expression but a pervasive phenomenon in human thought, perception, and reasoning (Lakoff and Johnson, 1980). Merriam-Webster Dictionary defines metaphor as "a figure of speech in which a word or phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness or analogy between them". Metaphor detection as a fundamental research task in natural language processing can benefit a variety of other tasks, which require the understanding of implicit semantics, such as ma-

*Corresponding author

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7735–7752

chine translation (Mao et al., 2018), sentiment and emotion analysis (Mao and Li, 2021), and conversational dialogue (Sun et al., 2023).

Existing studies on metaphor detection establish supervised learning methods based on labeled data (Rohanian et al., 2020; Le et al., 2020; Mao and Li, 2021). Recently, some studies employ metaphor theories to help design models for metaphor detection, resulting in significant performance gains (Zhang and Liu, 2022; Ge et al., 2022; Tian et al., 2023). Although these methods are inspired by different metaphor theories, they essentially train or fine-tune discriminative models to learn a decision boundary between metaphorical and literal samples, lacking the ability to explain the underlying reasoning process of metaphorical/literal judgment.

Recently, large language models (LLMs) have shown promise in generating reasoning processes with natural language across various tasks (Kojima et al., 2022; Wang et al., 2023a; Zhang et al., 2023), which demonstrates their potential to provide the underlying reasoning process for metaphor understanding. Although promising, LLM-based methods for metaphor detection and reasoning are still faced with the challenging issue of bringing the explainable concepts for metaphor reasoning and their linguistic manifestation. Figure 1 gives examples of LLM-based metaphor detection methods. Figure 1 (a) and (b) show that LLMs fail to detect the metaphor; even when employing zero-shot chain-of-thought method (Kojima et al., 2022). This limitation may be attributed to the fact that metaphors are not merely linguistic phenomena but associated with human thought and reasoning process (Lakoff and Johnson, 1980). Thus metaphors inherently convey intricate and implicit meanings that require a deeper level of understanding, making it more difficult for LLMs to comprehend. To better understand and detect metaphors, metaphor theories that provide well-established frameworks can serve as shortcuts to help effective metaphor reasoning and understanding. Figure 1 (c) provides an LLM with the prompt incorporated with the key ideas from conceptual metaphor theory (CMT) (Lakoff and Johnson, 1980), the key metaphor theory in cognitive linguistics. The result shows that, with the help of CMT theory, the LLM can provide a correct answer with an explanation of corresponding reasoning process.

In this paper, we tackle the challenging issue of LLM-based metaphor reasoning. To better release the capability of LLMs for metaphor reasoning and understanding, we make the first attempt to incorporate well-founded metaphor theories in such process. Further, to bridge the abstract conceptual representations of metaphor theories and the detailed LLM reasoning process, we draw inspiration from a pedagogical strategy called scaffolding instruction (Bruner, 1974; Vygotsky, 1978), which emphasizes that an educator continually explores a learner's mastery boundary through dialogues and designs questions and support as scaffolding, to assist learners in constructing the understanding of pedagogical goals step by step. To achieve this, we propose a novel **T**heory guided **S**caffolding **I**nstruction framework (TSI) for LLM-enabled metaphor reasoning. Specifically, we rely on three representative metaphor theories, selection preference violation (SPV) (Wilks, 1975, 1978), metaphor identification procedure (MIP) (Pragglejaz Group, 2007), and conceptual metaphor theory (CMT) (Lakoff and Johnson, 1980). In our computational construct, we first formalize a metaphor theory by converting it into a metaphor knowledge graph (KG), which functions as the instructional structure to construct scaffolding questions that help LLM's metaphor reasoning process. During dialogue interactions, our framework automatically detects the mastery boundary of LLM and provides relevant knowledge as scaffolding support when the question is beyond the LLM's capability.

Our main contributions are as follows:

- To release the metaphor reasoning capability of LLMs, we make the first attempt to propose a theory guided framework TSI and develop the conceptual representations by converting metaphor theories into knowledge graphs for scaffolding instruction.
- The scaffolding instruction process in TSI constructs prompts for scaffolding questions based on metaphor KGs to facilitate the step-by-step reasoning, and provides scaffolding support by dynamically assessing the LLM's mastery boundary.
- Experimental results on two datasets verify that our method achieves significant performance gains compared to LLM-based methods in complex reasoning and the SOTA methods in metaphor detection.

2 Related Work

Metaphor Theories As a fundamental and profound figurative phenomenon in linguistics and cognition, metaphor has been theorized across various disciplines in linguistics, philosophy and psychology (Plato, 1901; Kittay, 1990; Aristotle, 1995). However, not all of these theories are appropriate for computational formalization. Among them, three theories are favorable in existing computational research on metaphor, including selectional preference violation (SPV) (Wilks, 1975), metaphor identification procedure (MIP) (Pragglejaz Group, 2007), and conceptual metaphor theory (CMT) (Lakoff and Johnson, 1980).

The SPV theory suggests that the disparity between the *context* of a word within a sentence and its frequently used contexts is an indicator of this word's metaphorical usage. The MIP theory aims to standardize the metaphor annotation process, emphasizing that a metaphor is identified if the contextual meaning of the word differs from its basic *meaning*. Unlike SPV and MIP, which utilize indicative clues based on linguistic features of contexts or word meanings to identify metaphors, CMT goes beyond linguistic analysis and proposes a cognitive basis for metaphor understanding, suggesting that metaphor facilitates a mapping of attributes from *source domain* to *target domain* in human cognition. In this paper, we focus on releasing the metaphor reasoning capability of LLMs with the guidance of these three metaphor theories.

Computational Work on Metaphor without Theories Computational studies on metaphor focus on three main tasks, including metaphor detection, metaphor interpretation, and metaphor generation. Most of the work, especially the early work, established methods for computational metaphor processing without consideration of metaphor theories.

Metaphor detection, which attracts more research attention compared to the other two tasks, aims to identify an expression in text as metaphorical or literal. Early researchers utilize machine learning based methods to detect metaphors (Tsvetkov et al., 2014; Shutova et al., 2016; Bulat et al., 2017). Given that understanding metaphors requires knowledge beyond the context, some studies employ external knowledge resources for this task (Rohanian et al., 2020; Wan et al., 2021; Li et al., 2023). Benefiting from multitask learning, other studies learn shared information between metaphor detection and related tasks (Le et al., 2020; Mao and Li, 2021; Zhang and Liu, 2023; Badathala et al., 2023). Moreover, some studies (Lin et al., 2021; Feng and Ma, 2022) explore data augmentation methods to alleviate the problem of insufficient data for this task.

Metaphor interpretation involves explaining the implicit meaning conveyed by a metaphorical expression, which is a more challenging task than metaphor detection. Existing studies without using metaphor theories often formulate this task as a paraphrasing task to generate literal substitute paraphrase for a metaphorical word/phrase (Shutova, 2010; Shutova et al., 2012; Zayed et al., 2020) or identify the literal interpretation for a metaphorical expression in a candidate sentence set (Bizzoni and Lappin, 2018; Chakrabarty et al., 2022).

Metaphor generation focuses on generating metaphorical expressions from literal ones. Some studies generate a metaphorical word to replace the literal one using the end-to-end generation framework (Yu and Wan, 2019; Chakrabarty et al., 2021). Other studies generate the metaphorical sentence based on syntactic patterns (Brooks and Youssef, 2020) or a given Chinese noun (Li et al., 2022).

Computational Work on Metaphor with Theories Unlike the above research only focusing on computational perspectives for metaphor processing, recently, some researchers (Stowe et al., 2021b; Ge et al., 2022) have incorporated well-established metaphor theories into computational work on metaphor and achieved promising results.

Some studies (Su et al., 2021; Song et al., 2021; Choi et al., 2021; Zhang and Liu, 2022; Wang et al., 2023b) adopt SPV and MIP in network design for metaphor detection. Inspired by CMT, Ge et al. (2022) propose a model that generates plausible source and target concepts to help metaphor detection. Tian et al. (2023) further propose an attribute Siamese network to learn similar attributes between the source and target domains for this task.

In metaphor interpretation, some research considers the source domain implied by the metaphor in CMT as the interpretation and develops various methods to identify these domains, such as the unsupervised method (Shutova et al., 2017), deep learning based method (Rosen, 2018) and LLM-based method (Wachowiak and Gromann, 2023). Other research extracts attributes that link source and target domains as interpretation of metaphors (Su et al., 2017; Rai et al., 2019; Su et al., 2020).

In metaphor generation, Stowe et al. (2021b)

make the first attempt to generate metaphors with conceptual mappings grounded in CMT and develop the lexical model and seq-to-seq model for metaphor generation. In addition, Stowe et al. (2021a) focus on comparing free and controlled metaphor generation based on CMT.

Although previous methods achieve promising results, they only apply metaphor theories for network design or model the elements in these theories, lacking the ability to explicitly reflect the inherent reasoning processes grounded in theories. Recently, large language models (LLMs) have shown potential for reasoning and can generate reasoning processes in natural language, while how to develop the LLM's capability on metaphor reasoning remains unexplored. Thus, we focus on releasing LLM's ability on metaphor reasoning and understanding with the guidance of metaphor theories.

Complex Reasoning with LLMs In recent years, LLMs have shown multiple emergent abilities (Wei et al., 2022a), leading the shift of paradigm in natural language processing from fine-tuning to in-context learning. However, LLMs still exhibit limitations when they tackle complex reasoning tasks. To mitigate this gap, Wei et al. (2022b) propose chain-of-thought (CoT) prompting with task-specific exemplars of reasoning processes. After that, some studies design automatic CoT prompting methods encouraging LLMs to generate reasoning processes in a zero-shot manner (Kojima et al., 2022; Wang et al., 2023a; Zhang et al., 2023). Other studies explore enhancement methods to improve the reasoning ability of current CoT methods, such as refinement (Madaan et al., 2023), question decomposition (Zhou et al., 2023), voting (Wang et al., 2023c), ranking (Khalifa et al., 2023) and using external knowledge (Zhao et al., 2023). Cognitive linguists argue that metaphors are not just linguistic expressions but fundamental to human thought and cognition (Lakoff and Johnson, 1980), thus understanding metaphors requires complex reasoning processes in the human brain. Although the above research has explored many complex reasoning tasks, LLM-enabled metaphor reasoning is still a challenging task. Therefore, in this paper, we make the first attempt to design methods to unleash the potential of LLM on metaphor reasoning.

3 Problem Definition

Formally, $\mathcal{D}_{te}$=$\{(s_k, w_k, l_k)\}_{k=1}^{N_{te}}$ is the test dataset with N_{te} instances, where s_k is a sentence, w_k is a word within s_k, and l_k is the label (metaphorical or literal) for w_k. Our goal of metaphor detection with LLM is to predict the label of the word within a sentence in $\mathcal{D}_{te}$ without any training data.

4 Method

We propose a novel **T**heory guided **S**caffolding **I**nstruction framework (TSI) for LLM-enabled metaphor reasoning. Under the guidance of a metaphor theory, our framework instructs LLM to explicitly give a reasoning process with questioning and support as scaffolding to determine whether the word within a sentence is metaphorical or literal. Figure 2 shows the overall architecture of our framework, which contains four primary components: (1) *Metaphor Knowledge Graph Construction*, which represents a simplified metaphor theory and computational aspects in a graphical format; (2) *Scaffolding Question Construction*, which constructs a sequence of scaffolding questions for an LLM with metaphor knowledge graph as the instructional reference; (3) *Scaffolding Support based on Mastery Level of LLM*, which automatically detects the current capability of LLM (i.e. mastery level) and provides knowledge as support if necessary; and (4) *Classification*, which categories the word within the sentence as metaphorical or literal, based on the comparison between the structure of knowledge graph constructed from LLM's reasoning processes and the structure of metaphor knowledge graph.

4.1 Metaphor Knowledge Graph Construction

To facilitate theory guidance and computational construction for LLM-enabled metaphor reasoning, we manually simplify the metaphor theory and construct a metaphor knowledge graph (KG), which covers the essential information in both the theory and computational aspects. On the theory side, metaphor KG represents the fundamental concepts and the relation between concepts identified in the metaphor theory. On the computational side, metaphor KG also represents the core linguistic expressions of the word, sentence, and POS information as well as their connections. Specifically, the metaphor KG is formulated as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{F}\}$, where $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{F}$ represent sets of entities, relations, and facts, respectively. There are two groups of entities in our metaphor KG. One group consists of all the entities of concepts $\{e_k^c\}_{k=1}^{N_c}$ mentioned

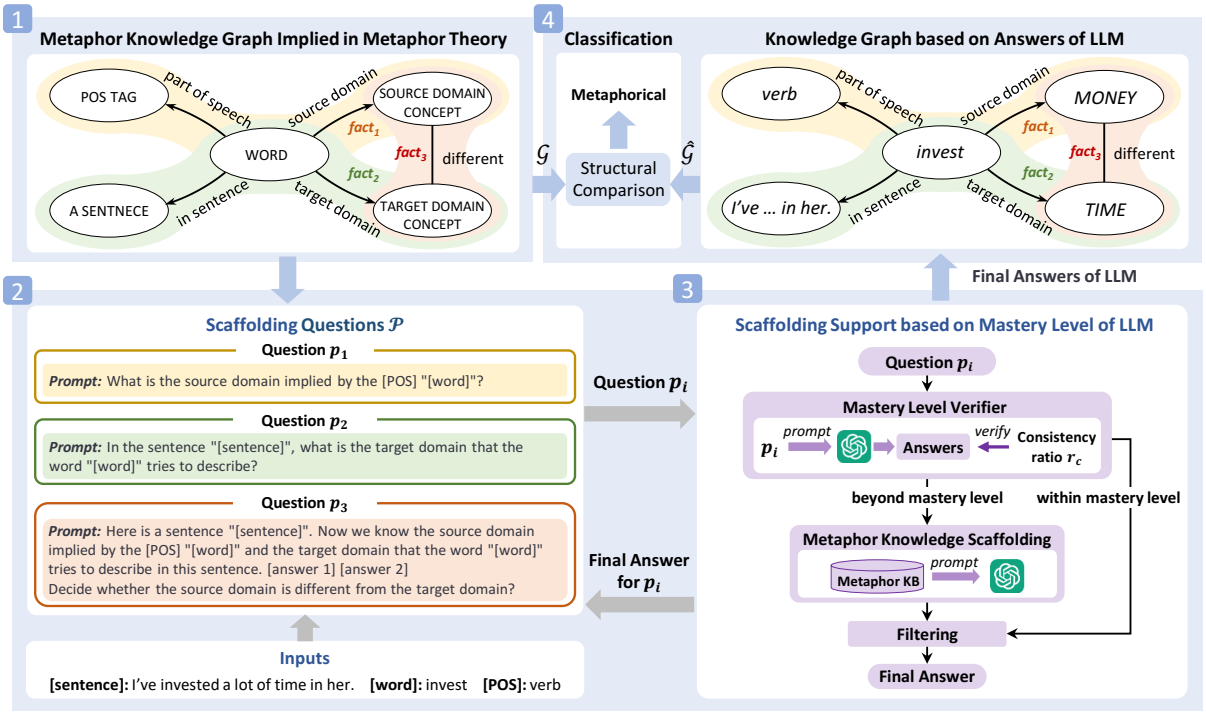

Figure 2: Overall architecture of our proposed framework TSI guided by conceptual metaphor theory for LLM-enabled metaphor reasoning. *[answer i]* denotes the input slot of the final answer for the scaffolding question p_i.

in metaphor theory or used for computational usage, and the other group consists of all the entities of attributes $\{e_k^a\}_{k=1}^{N_a}$ associated with the concepts in metaphor KG. The relation could be a directed relation r^d or an undirected relation r^u between entities. The i-th fact $f_i \in \mathcal{F}$ is denoted as a triple (e^h, r, e^t), where e^h and e^t denote head and tail entities, respectively. We apply our framework to three metaphor theories widely used in computational metaphor processing, including selectional preference violation (SPV) (Wilks, 1975, 1978), metaphor identification procedure (MIP) (Praggle-jaz Group, 2007), and conceptual metaphor theory (CMT) (Lakoff and Johnson, 1980). The details of metaphor KGs constructed based on the simplification of SPV, MIP and CMT in our methods are shown in Appendix B.

4.2 Scaffolding Question Construction

Questioning is a commonly used approach in the pedagogical strategy scaffolding instruction (Zheng et al., 2019; Laili and Siswono, 2020), which helps ascertain the learner's current level of understanding, stimulate critical thinking and encourage deeper understanding. Inspired by this, taking a sentence and a word within the sentence as initial information, we construct a sequence of questions as scaffolding with the metaphor KG as the reference guiding LLM to obtain metaphor reasoning processes. Each scaffolding question induces a reasoning process based on a fact in the metaphor KG. We have two types of reasoning processes, which are as follows:

- *Forward Reasoning*: In a fact (e.g. $fact_1$ and $fact_2$ in Figure 2), given an instantiated head entity of concept (along with its associated attribute entities, if applicable) and a direct relation in the metaphor KG, the LLM needs to determine the unknown instantiated tail entity;
- *Relation Reasoning*: In a fact (e.g. $fact_3$ in Figure 2), given two instantiated entities of concepts (along with their associated attribute entities, if applicable) in the metaphor KG, the LLM needs to infer whether the undirected relation between these two entities exists.

For the forward reasoning, we use a Wh-question to construct the scaffolding question. For the relation reasoning, we use the template "Decide whether ..." to construct the scaffolding question. We construct a sequence of scaffolding question prompts $\mathcal{P} = [p_1, p_2, \ldots, p_K]$ according to the sequence of facts $\mathcal{F} = [f_1, f_2, \ldots, f_K]$ within the metaphor KG. Details of these scaffolding question prompts for LLMs based on CMT, MIP and SPV, respectively, are illustrated in Appendix B.

4.3 Scaffolding Support based on Mastery Level of LLM

Scaffolding instruction also emphasizes the importance of providing support to learners when they engage in tasks beyond their *mastery levels*, which means what a learner can do independently (Wood et al., 1976; Vygotsky, 1978; Harris and Pressley, 1991). Inspired by this, we devise an approach to automatically detect the mastery level of LLM and provide knowledge as scaffolding support if the question is beyond the current capability of LLM. Figure 2 shows the overall architecture of this approach, which is a pipeline framework, including a mastery level verifier, a metaphor knowledge scaffolding module, and a filtering module. Appendix B gives the pseudocode of scaffolding support based on mastery level of LLM.

Mastery Level Verifier When humans are confused or lack confidence in answering a question, they tend to provide inconsistent responses when they are asked the same question multiple times (Schaeffer and Presser, 2003), as they may simply guess. Inspired by this observation, we propose the mastery level verifier by examining the consistency of the answers that the LLM provides to estimate the LLM's mastery level. For the i-th fact, given the scaffolding question prompt p_i, the set of inputs $\mathcal{X}_i = \{x_{(i,1)}, x_{(i,2)}, \ldots, x_{(i,N_i)}\}$ for this question, and an LLM denoted as $\mathcal{M}$, our framework generates the answer a_i formulated as

$$a_i = \mathcal{M}(p_i \| \mathcal{X}_i), \tag{1}$$

where N_i is the number of inputs, and we perform Eq. (1) for N_o times with temperature t_p, yielding a set of answers $\mathcal{A}_i = \{a_{(i,1)}, a_{(i,2)}, \ldots, a_{(i,N_o)}\}$. $r_c \in (0.5, 1]$ is the consistency ratio and $n_{min} = \lfloor N_o \times r_c + 0.5 \rfloor$ is the minimum number of consistent answers, where $\lfloor \cdot \rfloor$ denotes the floor function. The specific designs of our verifier for different reasoning processes are as follows:

- *Forward Reasoning:* We use T5 (Raffel et al., 2020) to calculate a semantic textual similarity score between every answer pair in $\mathcal{A}_i$, and obtain the similarity matrix $\mathbf{S}^{N_o \times N_o}$. If there exists a submatrix $\mathbf{B} = \mathbf{S}[j_1, j_2, \ldots, j_n; j_1, j_2, \ldots, j_n]$ and $s \geq s_{min}$ for $\forall s \in \mathbf{B}$, where $\{j_1, j_2, \ldots, j_n\}$ represents the indices of the n selected answers, $n \geq n_{min}$, and s_{min} is the semantic similarity threshold, we obtain the set of consistent answers $\mathcal{A}_i^c = \{a_{(i,j_1)}, a_{(i,j_2)}, \ldots, a_{(i,j_n)}\}$.
- *Relation Reasoning:* We first project the answers into "Yes", "No", or "Uncertain" with rules. If there exists a subset of answers $\mathcal{A}_i^c = \{a_{(i,j_1)}, a_{(i,j_2)}, \ldots, a_{(i,j_n)}\}$ and all elements in $\mathcal{A}_i^c$ can be projected into the same answer ("Yes" or "No"), where $\{j_1, j_2, \ldots, j_n\}$ represents the indices of the n selected answers, and $n \geq n_{min}$, we regard $\mathcal{A}_i^c$ as the set of consistent answers.

If there exist consistent answers $\mathcal{A}_i^c$, we regard the question falls within the mastery level of LLM; otherwise, the question is beyond the LLM's ability.

Metaphor Knowledge Scaffolding If the scaffolding question p_i surpasses the mastery level of the LLM, this module provides a scaffolding knowledge prompt b_i from the metaphor knowledge base (KB) to help the LLM answer this question. The detail of metaphor KB is shown in Appendix B. Given the scaffolding question prompt p_i, inputs $\mathcal{X}_i = \{x_{(i,1)}, x_{(i,2)}, \ldots, x_{(i,N_i)}\}$ for this question, scaffolding knowledge prompt b_i, and the LLM $\mathcal{M}$, the metaphor knowledge scaffolding module generates the answer a_i^s, which is formulated as

$$a_i^s = \mathcal{M}(p_i \| \mathcal{X}_i \| b_i). \tag{2}$$

We perform Eq. (2) for N_k times with temperature t_p resulting in a set of outputs $\mathcal{A}_i^s = \{a_{(i,1)}^s, a_{(i,2)}^s, \ldots, a_{(i,N_k)}^s\}$.

Filtering We randomly select one answer from the remaining answers of $\mathcal{A}_i^s$ after denoising or $\mathcal{A}_i^c$ as the final answer a_i^o.

4.4 Classification

Based on all the final reasoning answers of the LLM, we can construct an instantiated knowledge graph $\hat{\mathcal{G}}$ for the original input sentence and a word within the sentence. If the structure of $\hat{\mathcal{G}}$ is identical to that of $\mathcal{G}$, we label this instance as metaphorical; otherwise, we label it as literal.

5 Experiments

5.1 Datasets

We conducted experiments on two publicly available metaphor datasets: (1) **MOH-X** (Mohammad et al., 2016) comprises 647 sentences, where only a single verb is annotated as metaphorical or literal in each sentence; (2) **TroFi** (Birke and Sarkar, 2006) is another metaphor detection dataset, collected from 1987-1989 Wall Street Journal Corpus.

Dataset	#Instance	%Met.	Avg. L	#Samp
MOH-X	647	48.7	8.0	300
TroFi	3737	43.5	28.3	300

Table 1: Statistics of datasets. **#Instance** represents the number of all the instances. **%Met.** represents the percentage of metaphorical instances. **Avg. L** denotes the average length of instances. **#Samp** denotes the number of randomly sampled instances utilized for evaluation.

We sampled 300 balanced instances from MOH-X and 300 balanced instances from TroFi for testing. Table 1 shows the statistics of datasets.

5.2 Baseline Methods

We use representative methods for LLM-based complex reasoning as baselines, including the standard zero-shot method, automatic chain-of-thought methods, and chain-of-thought enhancement methods, which are as follows: (1) **Standard zero-shot** (Wei et al., 2021) prompts the LLM to provide answers directly in a zero-shot manner; (2) **Zero-shot CoT** (Kojima et al., 2022) concatenates the question prompt with a simple trigger sentence (i.e. "*Let's think step by step*") encouraging LLM to generate the reasoning process in a zero-shot manner; (3) **Plan-and-solve** (Wang et al., 2023a) designs a prompt aimed at guiding LLM to solve the problem by breaking the task into smaller subtasks, and then implementing the devised plan; (4) **Self-refine** (Madaan et al., 2023) improves initial outputs from LLM through iterative feedback and refinement; (5) **Self-consistency** (Wang et al., 2023c) selects the most consistent answer through majority voting among sampled reasoning chains; (6) **Least-to-most** (Zhou et al., 2023) initially decomposes the question into sub-questions in a top-down manner and then addresses each sub-question.

We also use the SOTA methods for metaphor detection as strong baselines, which are as follows: (1) **MelBERT** (Choi et al., 2021) leverages contextualized information inspired by MIP and SPV for metaphor detection; (2) **MisNet** (Zhang and Liu, 2022) which incorporates MIP and SPV into their linguistics enhanced network; (3) **AdMul** (Zhang and Liu, 2023) is the SOTA method for metaphor detection, which employs a multi-task learning framework to transfer knowledge from basic sense discrimination to metaphor detection via adversarial training. We use these methods trained on a large metaphor dataset VUA All (Leong et al., 2018) to perform zero-shot transfer on our datasets.

5.3 Implementation Details

We employ the accuracy and macro-average F1 for evaluation and report the mean and standard deviation of 3 runs in our experiments. The OpenAI GPT-3.5 (gpt-3.5-turbo-0613)[1] serves as the LLM in our experiments. The details of rules to filter answers in our method are shown in Appendix B. The prompt design for the baselines and other implementation details are shown in Appendix C.[2]

5.4 Main Results

Comparison with Baselines Table 2 shows the comparative results between our methods and baselines. Our methods guided by MIP and CMT achieve significant improvements over all the representative LLM-based methods in complex reasoning, which verifies the effectiveness of our scaffolding instruction framework for LLM-enabled metaphor reasoning. Least-to-most, tackling problems through sub-questions, outperforms other baselines on MOH-X. Zero-shot CoT, encouraging LLM to think step by step, achieves better results than other baselines on TroFi. Both of these methods guide LLM to tackle problems step by step, indicating the necessity of deep and sequential thinking in metaphor reasoning.

Compared with SOTA methods in metaphor detection, our method guided by CMT outperforms them across all datasets, and our method guided by MIP also achieves better performances on TroFi. These results further verify the effectiveness of our method. Although MelBERT and MisNet design their networks under the guidance of metaphor theories and fine-tune pre-trained models for metaphor detection, they rely on capturing surface-level clues in these theories and fail to capture the underlying metaphor reasoning process reflected in these theories. This limitation results in our method guided by MIP and CMT achieving significant improvements over MelBERT and MisNet.

Comparison between Metaphor Theories The experimental results in Table 2 indicate that our method guided by SPV performs worse compared to our methods guided by MIP and CMT. One possible reason is that as some metaphorical usages become prevalent over time, SPV, relying on data fre-

[1] https://platform.openai.com/docs/models/gpt-3-5

[2] Our codes are available at https://github.com/TIAN-viola/TSI.

Method	MOH-X		TroFi	
	F1	Acc.	F1	Acc.
Methods in LLM-based complex reasoning				
Standard zero-shot (Wei et al., 2021)	66.43 ± 1.45	69.11 ± 1.29	58.58 ± 1.01	61.22 ± 0.95
Zero-shot CoT (Kojima et al., 2022)	70.97 ± 0.37	71.22 ± 0.32	*64.18 ± 1.07*	*64.78 ± 0.95*
Plan-and-Solve (Wang et al., 2023a)	54.27 ± 2.73	57.89 ± 2.01	58.13 ± 2.29	58.33 ± 2.23
Self-refine (Madaan et al., 2023)	64.06 ± 0.37	67.67 ± 0.27	57.27 ± 0.22	60.89 ± 0.16
Self-consistency (Wang et al., 2023c)	70.65 ± 1.58	72.22 ± 1.50	60.30 ± 0.76	61.56 ± 0.68
Least-to-most (Zhou et al., 2023)	*74.43 ± 1.27*	*75.00 ± 1.25*	63.88 ± 3.16	64.11 ± 3.03
Methods in metaphor detection (zero-shot transfer)				
MelBERT (Choi et al., 2021)	77.88 ± 0.83	77.89 ± 0.83	*62.36 ± 1.51*	*62.89 ± 1.29*
MisNet (Zhang and Liu, 2022)	77.08 ± 1.12	77.11 ± 1.13	62.01 ± 0.64	62.67 ± 0.54
AdMul (Zhang and Liu, 2023)	*79.74 ± 0.44*	*79.89 ± 0.42*	60.54 ± 1.43	62.67 ± 0.98
Our methods guided by different metaphor theories				
TSI (SPV)	74.22 ± 1.95	74.33 ± 1.96	51.93 ± 1.60	56.22 ± 1.03
TSI (MIP)	79.39 ± 0.35	79.44 ± 0.31	65.60 ± 1.08	65.89 ± 1.03
TSI (CMT)	**82.59 ± 2.22**	**82.93 ± 1.94**	**66.07 ± 1.11**	**66.89 ± 1.13**

Table 2: Comparison between our methods and baselines. The best results are highlighted in bold font. The best results for baselines in LLM-based complex reasoning and those in metaphor detection are underlined

Variant	MOH-X		TroFi	
	F1	Acc.	F1	Acc.
TSI (SPV)	**74.22 ± 1.95**	**74.33 ± 1.96**	**51.93 ± 1.60**	**56.22 ± 1.03**
w/o Metaphor Knowledge Scaffolding	73.08 ± 1.30	73.22 ± 1.34	50.94 ± 1.54	55.56 ± 1.26
w/o Mastery Level Verifier	72.30 ± 1.58	72.44 ± 1.59	50.13 ± 1.17	55.11 ± 0.96
w/o Metaphor Knowledge Graph	69.11 ± 0.25	69.44 ± 0.31	51.78 ± 1.00	53.33 ± 0.72
TSI (MIP)	**79.39 ± 0.35**	**79.44 ± 0.31**	**65.60 ± 1.08**	**65.89 ± 1.03**
w/o Metaphor Knowledge Scaffolding	70.36 ± 1.30	71.33 ± 1.19	65.40 ± 0.91	65.78 ± 0.63
w/o Mastery Level Verifier	69.86 ± 0.72	71.00 ± 0.82	65.16 ± 0.40	65.22 ± 0.42
w/o Metaphor Knowledge Graph	69.11 ± 0.25	69.44 ± 0.31	58.69 ± 0.87	58.78 ± 0.83
TSI (CMT)	**82.59 ± 2.22**	**82.93 ± 1.94**	**66.07 ± 1.11**	**66.89 ± 1.13**
w/o Metaphor Knowledge Scaffolding	79.11 ± 0.69	79.11 ± 0.68	64.28 ± 0.15	65.33 ± 0.27
w/o Mastery Level Verifier	79.07 ± 0.53	79.11 ± 0.57	63.43 ± 1.05	64.67 ± 0.98
w/o Metaphor Knowledge Graph	75.55 ± 1.77	76.11 ± 1.73	62.50 ± 2.32	62.63 ± 2.10

Table 3: Experimental results of ablation study.

quency rather than semantic essence, tend to lead our method to wrong results on such metaphorical usages. In addition, our method using MIP performs better than our SPV guided method. This improvement seems to benefit from the fact that understanding contextual meaning is a strength of LLM, which is an important step in our method guided by MIP. In contrast to SPV and MIP, which analyze metaphor at the linguistic level, CMT mainly leverages source and target domains to explain implicit comparisons in metaphors reflected in human cognition, which seems to contribute to the superior performances of our method guided by CMT across all datasets. These results also indicate the potential of the comprehensive perspectives in metaphor theories, especially CMT, for enhancing computational work on metaphor.

5.5 Ablation Study

We conduct the ablation study to evaluate the impact of components in our methods, using the following variants: (a) **w/o Metaphor Knowledge Scaffolding** removes the *metaphor knowledge scaffolding* module in our method from our methods; (b) **w/o Mastery Level Verifier** ablates *mastery level verifier* and its subsequent *metaphor knowledge scaffolding* module; (c) **w/o Metaphor Knowledge Graph** removes the metaphor KG and presents LLM with a single question prompt. This prompt comprises a description of key ideas in the metaphor theory and a direct query about whether the word in the sentence is metaphorical, templates of which are shown in Appendix C.

Table 3 shows the results of our ablation study.

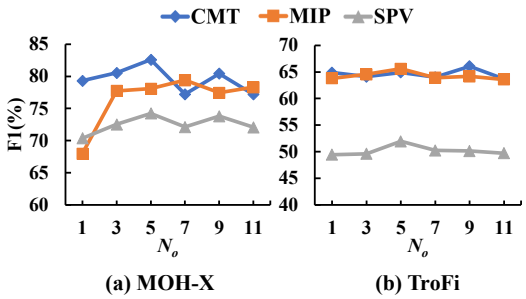

Figure 3: Results of our methods with different numbers of answers generated by LLM on MOH-X and TroFi.

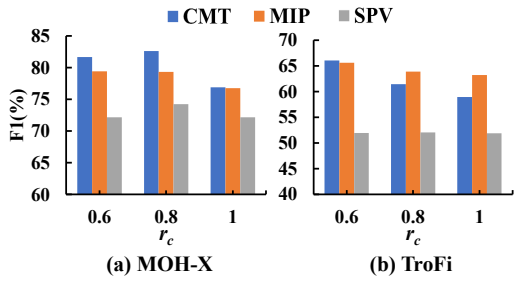

Figure 4: Results of our methods with different consistency ratios on MOH-X and TroFi.

Removing the metaphor knowledge scaffolding module reduces the performance, thus verifying the effectiveness of the knowledge base we provide in our method. Furthermore, we remove the mastery level verifier from our methods, resulting in significant performance declines across all the datasets and theories. These variants demonstrate the effectiveness of the approach to provide scaffolding support based on detecting the mastery level of LLM in our method. In addition, to evaluate the effectiveness of the metaphor KG representation, we directly ablate it and present LLM with the theory description and a query to ask whether the word in the sentence is metaphorical. This variant uses the in-context learning ability of LLM to learn metaphor theory, which leads to sharp drops in performance across all the datasets, indicating that LLM can learn a metaphor theory better through the representation of metaphor knowledge graphs rather than theory descriptions.

5.6 Hyper-parameter Analysis

Number of Answers To explore the impact of the number of answers generated by LLM (denoted as N_o) in our methods, we experiment on varying N_o from 1 to 11. The experimental results in Figure 3 show that increasing N_o initially enhances the performance, followed by a plateau or marginal decline of the performance.

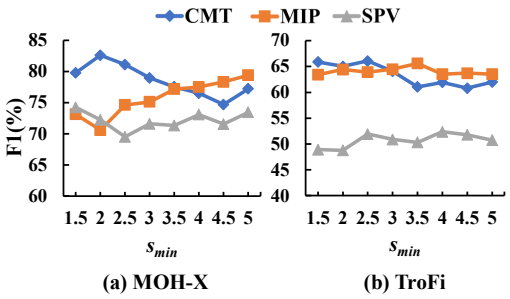

Figure 5: Results of our methods with different semantic similarity thresholds on MOH-X and TroFi.

Consistency Ratio To analyze the impact of the consistency ratio r_c in our mastery level verifier, we conduct experiments varying r_c from 0.6 to 1. The experimental results in Figure 4 indicate that a high consistency ratio could bring performance drops for the reason that an appropriate ratio allows LLM to make a slight number of incorrect answers.

Semantic Similarity Threshold To evaluate the influence of semantic similarity threshold s_{min} in our mastery level verifier, we conduct experiments on varying the threshold from 1.5 to 5.0, where a higher score indicates greater semantic similarity. From the experimental results in Figure 5, we can see that our method guided by CMT achieves the optimal performance at lower thresholds, while our method guided by MIP excels with larger thresholds. This phenomenon might be attributed to the variance in the quality of knowledge presented in the metaphor knowledge scaffolding module. The knowledge derived from MIP seems to be more comprehensible for LLM in comparison to the knowledge derived from CMT.

6 Conclusion

In this paper, we propose a theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. Inspired by a pedagogical strategy *scaffolding instruction*, our framework instructs an LLM to infer the metaphor reasoning process using questioning and support as scaffolding guided by a metaphor theory. Experimental results show that our framework outperforms the previous LLM-based complex reasoning methods and the SOTA methods for metaphor detection, verifying the effectiveness of our proposed framework and indicating the facilitation of metaphor and instruction theories in LLM-enabled reasoning.

Limitations

Our work has some limitations. Firstly, we are unable to evaluate our framework on large-scale test sets due to the costs of experimentation with LLMs, thus we obtain balanced randomly sampled test datasets for the evaluation of this work. In addition, we simplify metaphor theories by focusing on fundamental concepts, their relations and corresponding knowledge in our method. However, the metaphor theories that we use, especially CMT, contain rich contexts, which are worth further exploration in LLM-enabled metaphor reasoning. Compared to the current pretrained model based work on metaphor detection, our work can be viewed as opening a new path to the theory-directed, instruction-oriented LLM-based metaphor understanding and reasoning. To this end, we hope our work can promote future research on further enriching knowledge representation based on the comprehensive perspectives in metaphor theories, forming a more in-depth understanding of metaphors and incorporating it into the LLM-enabled reasoning process through novel methodologies, by peer researchers and ourselves.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grants #72293575 and #62206287. We thank the anonymous reviewers for the valuable comments.

References

Aristotle. 1995. *Treatise on Rhetoric*. Prometheus Books.

Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. In *Findings of the Annual Meeting of the Association for Computational Linguistics*, pages 388–401.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55.

Jennifer Brooks and Abdou Youssef. 2020. Discriminative pattern mining for natural language metaphor generation. In *Proceedings of the IEEE International Conference on Big Data*, pages 4276–4283.

Jerome Bruner. 1974. *Toward a theory of instruction*. Harvard University Press.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–528.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4250–4261.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1763–1773.

Huawen Feng and Qianli Ma. 2022. It's better to teach fishing than giving a fish: An auto-augmented structure-aware generative model for metaphor detection. In *Findings of the Conference on Empirical Methods in Natural Language Processing*, pages 656–667.

Mengshi Ge, Rui Mao, and Erik Cambria. 2022. Explainable metaphor identification inspired by conceptual metaphor theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10681–10689.

Karen Harris and Michael Pressley. 1991. The nature of cognitive strategy instruction: Interactive strategy construction. *Exceptional Children*, 57(5):392–404.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. GRACE: Discriminator-guided chain-of-thought reasoning. In *Findings of the Conference on Empirical Methods in Natural Language Processing*, pages 15299–15328.

Eva Feder Kittay. 1990. *Metaphor: Its cognitive force and linguistic structure*. Oxford University Press.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 1–25.

Nurul Laili and Tatag Yuli Eko Siswono. 2020. Giving questions as scaffolding to help student in constructing proof. *MUST: Journal of Mathematics Education, Science and Technology*, 5(2):143–155.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago press.

Duong Le, My Thai, and Thien Nguyen. 2020. Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8139–8146.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022. CM-Gen: A neural framework for Chinese metaphor generation with explicit context modelling. In *Proceedings of the International Conference on Computational Linguistics*, pages 6468–6479.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023. FrameBERT: Conceptual metaphor detection with frame embedding learning. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563.

Zhenxi Lin, Qianli Ma, Jiangyue Yan, and Jieyu Chen. 2021. CATE: A contrastive pre-trained model for metaphor detection with semi-supervised learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3888–3898.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. SELF-REFINE: Iterative refinement with self-feedback. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 1–54.

Rui Mao and Xiao Li. 2021. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13534–13542.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1222–1231.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Plato. 1901. *The Republic*. Macmillan.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Sunny Rai, Shampa Chakraverty, Devendra K Tayal, Divyanshu Sharma, and Ayush Garg. 2019. Understanding metaphors using emotions. *New Generation Computing*, 37:5–27.

Omid Rohanian, Marek Rei, Shiva Taslimipoor, and Le An Ha. 2020. Verbal multiword expressions for identification of metaphor. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2890–2895.

Zachary Rosen. 2018. Computationally constructed concepts: A machine learning approach to metaphor interpretation using usage-based construction grammatical cues. In *Proceedings of the Workshop on Figurative Language Processing*, pages 102–109.

Nora Cate Schaeffer and Stanley Presser. 2003. The science of asking questions. *Annual review of sociology*, 29(1):65–88.

Ekaterina Shutova. 2010. Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 160–170.

Ekaterina Shutova, Lin Sun, Elkin Darío Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. 2017. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, 43(1):71–123.

Ekaterina Shutova, Tim Van de Cruys, and Anna Korhonen. 2012. Unsupervised metaphor paraphrasing using a vector space model. In *Proceedings of the International Conference on Computational Linguistics*, pages 1121–1130.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4240–4251.

Kevin Stowe, Nils Beck, and Iryna Gurevych. 2021a. Exploring metaphoric paraphrase generation. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 323–336.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021b. Metaphor generation with conceptual mappings. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 6724–6736.

Chang Su, Shuman Huang, and Yijiang Chen. 2017. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300–311.

Chang Su, Ying Peng, Shuman Huang, and Yijiang Chen. 2020. A metaphor comprehension method based on culture-related hierarchical semantic model. *Neural Processing Letters*, 51:2807–2826.

Chang Su, Kechun Wu, and Yijiang Chen. 2021. Enhanced metaphor detection via incorporation of external knowledge based on linguistic theories. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1280–1287.

Weiwei Sun, Shuyu Guo, Shuo Zhang, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2023. Metaphorical user simulators for evaluating task-oriented dialogue systems. *ACM Transactions on Information Systems*, 42(1):1–29.

Yuan Tian, Nan Xu, Wenji Mao, and Daniel Zeng. 2023. Modeling conceptual attribute likeness and domain inconsistency for metaphor detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7736–7752.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 248–258.

Lev Semyonovich Vygotsky. 1978. *Mind in Society: The development of higher psychological processes*. Harvard University Press.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? Identifying metaphor mappings with generative language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1018–1032.

Hai Wan, Jinxia Lin, Jianfeng Du, Dawei Shen, and Manrong Zhang. 2021. Enhancing metaphor detection by gloss-based interpretations. In *Findings of the Annual Meeting of the Association for Computational Linguistics*, pages 1971–1981.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2609–2634.

Shun Wang, Yucheng Li, Chenghua Lin, Loic Barrault, and Frank Guerin. 2023b. Metaphor detection with effective context denoising. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 1404–1409.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations*, pages 1–24.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. In *Proceedings of the International Conference on Learning Representations*, pages 1–46.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*, pages 1–30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 24824–24837.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, 6(1):53–74.

Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223.

David Wood, Jerome Seymour Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100.

Zhiwei Yu and Xiaojun Wan. 2019. How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 861–871.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2020. Figure me out: A gold standard dataset for metaphor interpretation. In *Proceedings of the Language Resources and Evaluation Conference*, pages 5810–5819.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced Siamese network. In *Proceedings of the International Conference on Computational Linguistics*, pages 4149–4159.

Shenglong Zhang and Ying Liu. 2023. Adversarial multi-task learning for end-to-end metaphor detection. In *Findings of the Annual Meeting of the Association for Computational Linguistics*, pages 1483–1497.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *Proceedings of the International Conference on Learning Representations*, pages 1–32.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-Edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5823–5840.

Wenshu Zheng, Chenglin Wang, et al. 2019. Teachers' questioning to scaffold students' critical thinking. *Academic Journal of Humanities & Social Sciences*, 2(2):107–110.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *Proceedings of the International Conference on Learning Representations*, pages 1–61.

A Licenses of Scientific Artifacts

The OpenAI API for gpt-3.5-turbo-0613 is available at `https://openai.com/api/`. The license of T5 (Raffel et al., 2020) is Apache-2.0 and the license of MOH-X (Mohammad et al., 2016) is available at `https://saifmohammad.com/WebPages/SentimentEmotionLabeledData.html`. The license of TroFi (Birke and Sarkar, 2006) is available at `https://www2.cs.sfu.ca/~anoop/students/jbirke/LICENSE.html`.

B Method Details

Metaphor Knowledge Graph Construction To facilitate theory guidance and computational construct for LLM-based metaphor reasoning, we make the compromise and simplify the metaphor theory, retaining the fundamental concepts and their relations in our metaphor knowledge graph representations. Specifically, Figure 6 illustrates the details of metaphor knowledge graphs constructed based on the computational elements and relations, as well as the simplified theories, including CMT, MIP, and SPV in our methods.

Prompt Design for Scaffolding Questions Table 4 summarizes a list of template prompts for scaffolding questions $\mathcal{P}$ in our methods using different metaphor theories.

Scaffolding Support based on Mastery Level of LLM Table 5 shows the rules in our methods to project the answers of LLM for the relation reasoning process into "Yes", "No", or "Uncertain" in scaffolding support based on the mastery level module. If there are any candidate phrases in Table 5 that appear in the answer after converting all the uppercase characters in the answer string into lowercase characters, we project this answer into the corresponding category. If none of these candidate phrases appear in the answer, we check the answer manually. Table 6 shows details about the template prompts of the metaphor knowledge base in the metaphor knowledge scaffolding module in our methods using different metaphor theories. Algorithm 1 is the pseudocode of scaffolding support based on mastery level of LLM.

Filtering We first convert all the uppercase characters in the answer string into lowercase characters and then filter out the uncertain answers generated from the metaphor knowledge scaffolding module that contains any of the following phrases: "it is not possible to determine", "cannot be determined", "is uncertain" and "difficult to provide a definitive answer".

C Implementation Details

Table 7 shows the hyperparameter values of our methods in our experiments on different datasets. We use T5-3B[1] to calculate the semantic textual similarity score between the answer pair. The number of sampled reasoning paths for the baseline Self-consistency is 40. Table 8 summarizes a list of prompts for baselines in our experiments. Table 9 illustrates the template prompts for the variant of *w/o Metaphor Knowledge Graph* in the ablation study. If there is no knowledge prompt for a question, we just use the majority voter to replace scaffolding support based on mastery level of LLM module and get the final answer for this question. The system prompt that we use for GPT-3.5 in the experiments of our methods and baselines is as follows.

[1] https://huggingface.co/t5-3b

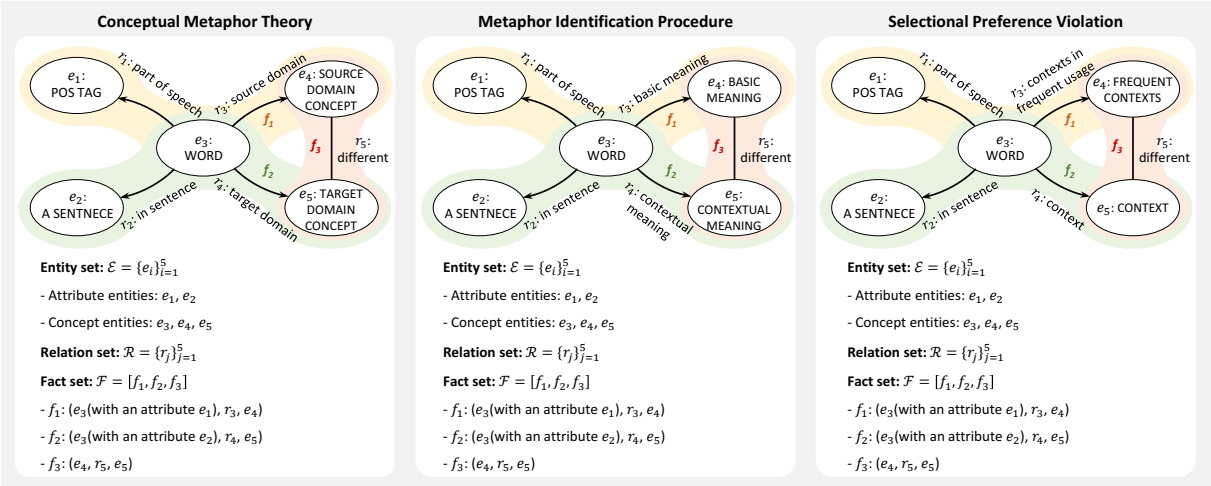

Figure 6: Metaphor knowledge graphs guided by CMT, MIP and SPV.

Theory	Question	Template Prompt
CMT	**Question** p_1	The source domain is a conceptual domain containing concepts that are typically concrete, tangible, and familiar to us. What is the source domain implied by the [POS] "[word]"?
	Question p_2	The target domain is a conceptual domain that contains concepts that are typically vague and abstract. In the sentence "[sentence]", what is the target domain that the word "[word]" tries to describe?
	Question p_3	Here is a sentence "[sentence]". Now we know the source domain implied by the [POS] "[word]" and the target domain that the word "[word]" tries to describe in this sentence. [answer 1] [answer 2] Based on the above information, decide whether the source domain is different from the target domain?
MIP	**Question** p_1	What is the basic meaning of the [POS] "[word]"?
	Question p_2	In the sentence "[sentence]", what is the contextual meaning of the word "[word]"?
	Question p_3	Here is a sentence "[sentence]". Now we know the basic meaning of the [POS] "[word]" and the contextual meaning of the word "[word]". [answer 1] [answer 2] Based on above information, decide whether the contextual meaning of the word "[word]" is different from its basic meaning.
SPV	**Question** p_1	Please provide examples of the frequent usage of the [POS] "[word]".
	Question p_2	Now we know examples of the frequent usage of the [POS] "[word]". [answer 1] In the sentence "[sentence]", decide whether the context of the word "[word]" is different from its contexts in the above sentences.

Table 4: Template prompts for scaffolding questions $\mathcal{P}$. *[sentence]* and *[word]* denote the input slots for the sentence in an instance and the target word within this sentence, respectively. *[POS]* denotes the input slot of the part of speech of the target word. *[answer i]* denotes the input slot for the LLM's final answer to question p_i.

You are ChatGPT, a large language model trained by OpenAI. Answer as concisely as possible.
Knowledge cutoff: 2021-09
Current date: 2023-10-15

Algorithm 1 Scaffolding Support based on Mastery Level of LLM

Require: (1) $\mathcal{M}$: an LLM; (2) $\mathrm{Combs}(\mathcal{L}, k)$: obtains combinations of all elements within a set $\mathcal{L}$ taken k elements at a time.

Input: (1) $\mathcal{A}_i$: set of N_o answers of $\mathcal{M}$ for the i-th scaffolding question prompt p_i; (2) r_c: consistency ratio; (3) s_{min}: semantic similarity threshold; (4) b_i: the scaffolding knowledge prompt for p_i; (5) $\mathcal{X}_i$: inputs for p_i.

Output: a_i^o: final answer for the i-th scaffolding question p_i.

Compute $n_{min} = \lfloor N_o \times r_c + 0.5 \rfloor$ ▷ ***Mastery Level Verifier***
if p_i is for forward reasoning **then**
  Compute $\mathbf{S}^{N_o \times N_o}$ with semantic textual similarity scores between every pair of answers in $\mathcal{A}_i$
  for all $j \leftarrow N_o$ **to** n_{min} **do**
    for all $I_j \in Combs([1, \ldots, N_o], j)$ **do**
      if $\forall s \in \mathbf{S}[I_j, I_j], s \geq s_{min}$ **then**
        $\mathcal{A}_i^c = \mathcal{A}_i[I_j]$, break
    if $\mathcal{A}_i^c$ exists **then** break
else
  Obtain $\mathcal{A}_i^p$ by project answers in $\mathcal{A}_i$ to "Yes", "No" or "Uncertain" with rules
  if number of "Yes" in $\mathcal{A}_i^p \geq n_{min}$ **then**
    Retrieve the index list I of element "Yes" within $\mathcal{A}_i^p$
    $\mathcal{A}_i^c = \mathcal{A}_i[I]$
  else
    if number of "No" in $\mathcal{A}_i^p \geq n_{min}$ **then**
      Retrieve the index list I of element "No" within $\mathcal{A}_i^p$
      $\mathcal{A}_i^c = \mathcal{A}_i[I]$
if $\mathcal{A}_i^c$ does not exists **then** ▷ ***Metaphor Knowledge Scaffolding***
  Compute $a_i^s = \mathcal{M}(p_i \| \mathcal{X}_i \| b_i)$ for N_k times ▷ Eq. (2)
  Obtain $\mathcal{A}_i^s$ by denoising answers in $\{a_{(i,1)}^s, a_{(i,2)}^s, \ldots, a_{(i,N_k)}^s\}$
  Sample an answer a_i^o from $\mathcal{A}_i^s$ ▷ ***Filtering***
else
  Sample an answer a_i^o from $\mathcal{A}_i^c$ ▷ ***Filtering***

Answer	Category	Candidate Phrase
Answer for question q_3 (CMT)	Yes	Yes,; Yes.; is/are (indeed/likely/most likely/conceptually/generally/inherently/potentially/ fundamentally) different; can/might/seems to/appears to/may/could/would (indeed) be (considered) different/distinct; may/appears to/might/can/could/does/would differ; differs from; do not align; are/is not (inherently/conceptually/generally/potentially) the same; the source (domain) and (the) target domains/domain differ
	No	No.; No,; is/are (likely/conceptually/generally/inherently/potentially) the same; can/might/seems to/appears to/may/could/would (indeed) be (considered) the same; is/are (likely/conceptually/generally/inherently/potentially/closely) similar/aligned/related; is/are not indeed/likely/most (likely/conceptually/generally/inherently/potentially/fundamentally) different/disinct
	Uncertain	it's difficult to provide a definitive answer; it is not possible to determine; without further context/information/clarification; need further clarification/context
Answer for question q_3 (MIP)	Yes	Yes,; Yes.; is/are (indeed/likely/most likely/conceptually/generally/inherently/potentially/ fundamentally) different; can/might/seems to/appears to/may/could/would (indeed) be (considered) different/distinct; may/appears to/might/can/could/does/would differ; differs from; do not align; are/is not (inherently/conceptually/generally/potentially) the same; goes beyond the/its basic meaning; diverges from; can/might/seems to/appears to/may/could/would (indeed) deviate
	No	No.; No,; is/are (likely/conceptually/generally/inherently/potentially) the same/equivalent; can/might/seems to/appears to/may/could/would (indeed) be (considered) the same; is/are (likely/conceptually/generally/inherently/potentially/closely) similar/aligned/related; is/are not indeed/likely/most (likely/conceptually/generally/inherently/potentially/fundamentally) different/distinct; corresponds with its/the basic meaning; is used in its/the basic meaning
	Uncertain	it's difficult to provide a definitive answer; it is not possible to determine; without further context/information/clarification; need further clarification/context
Answer for question q_2 (SPV)	Yes	Yes,; Yes.; is/are (indeed/likely/most likely/conceptually/generally/inherently/potentially/ fundamentally) different; can/might/seems to/appears to/may/could/would (indeed) be (considered) different/distinct; may/appears to/might/can/could/does/would differ; differs from; do not align; are/is not (inherently/conceptually/generally/potentially) the same; does suggest a different usage; implies a different context; is used differently in this context
	No	No.; No,; is/are (likely/conceptually/generally/inherently/potentially) the same; can/might/seems to/appears to/may/could/would (indeed) be (considered) the same; is/are (likely/conceptually/generally/inherently/potentially/closely) similar/aligned/related; is/are not indeed/likely/most (likely/conceptually/generally/inherently/potentially/fundamentally) different/distinct; would represent/be a similar usage; can/might/seems to/appears to/may/could/would (indeed) be consistent with its frequent usage; is used in the same context; in a context similar to its frequent usage; falls under the first/second/third example; is related to its usage in the first/second/third example
	Uncertain	it's difficult to provide a definitive answer; it is not possible to determine; without further context/information/clarification; need further clarification/context

Table 5: Rules to project the answers of LLM in the relation reasoning process into "Yes", "No" or "Uncertain".

Theory	Knowledge	Template Prompt
CMT	**Knowledge** b_1	1. The source domain is the concept area from which the metaphor is drawn. 2. The concepts in source domain are typically concrete. 3. The source domain is the domain of experience or concepts that are more concrete, tangible, and familiar to us. It serves as the basis for understanding or talking about a less concrete or abstract concept, which is referred to as the target domain. The source domain provides the metaphorical elements or framework through which we comprehend the target domain. 4. The source domain is a conceptual domain. Conceptual domains are sets of value meanings (presented using a list of concepts or a description of the members of the set) and are used to describe the set of concepts that can be represented within a data element. 5. For example, in the metaphor "I've invested a lot of time in her," the source domain is "money" and we draw upon our understanding of money to make sense of the concept of time in terms of value, efficiency, and spending. Taking the knowledge provided above into account, please answer the following question:
	Knowledge b_2	1. Target domain is used for the concept area to which the metaphor is applied. 2. The concepts in the target domain are typically vague and abstract. 3. The target domain is a conceptual domain. Conceptual domains are sets of value meanings (presented using a list of concepts or a description of the members of the set) and are used to describe the set of concepts that can be represented within a data element. 4. For example, in the metaphor "I've invested a lot of time in her," the target domain is "time", which is being conceptualized in terms of the source domain of money. Taking the knowledge provided above into account, please answer the following question:
MIP	**Knowledge** b_1	The basic meaning can be found in general users' dictionaries. The basic meaning is not necessarily the most frequent meaning of the lexical unit. The basic meaning tends to be - More concrete; what they evoke is easier to imagine, see, hear, feel, smell, and taste. - Related to bodily action. - More precise (as opposed to vague) - Historically older. Taking the knowledge provided above into account, please answer the following question:
	Knowledge b_2	The contextual meaning means the meaning of the lexical unit in context, that is, how the lexical unit applies to an entity, relation, or attribute in the situation evoked by the text. Take into account what comes before and after the lexical unit. The contextual meaning may be conventionalized and will thus be found in a general users' dictionary. It may also be novel or specialized and will thus not be found in a general users' dictionary. Taking the knowledge provided above into account, please answer the following question:
	Knowledge b_3	If the basic meaning of a word contrasts with its contextual meaning, there is a difference as well as comparison between the contextual and a more basic meaning. Taking the knowledge provided above into account, please answer the following question:
SPV	**Knowledge** b_1	The phrase "frequent usage of a word" refers to the regular and common application of a word.
	Knowledge b_2	In the theory of selectional preference violation, Wilks suggests that metaphor represents a violation of combinatory norms in the linguistic context and that metaphorical expressions can be detected via such violation. Taking the knowledge provided above into account, please answer the following question:

Table 6: Prompt design for scaffolding knowledge based on metaphor theories in our methods.

Notation	Our method (CMT)		Our method (MIP)		Our method (SPV)		Description
	MOH-X	TroFi	MOH-X	TroFi	MOH-X	TroFi	
N_o	5	9	7	5	5	5	number of answers generated by LLM for a question
s_{min}	2.0	2.5	5.0	3.5	1.5	2.5	semantic similarity threshold
r_c	0.8	0.6	0.6	0.6	0.8	0.6	consistency ratio in mastery level verifier
t_p	1	1	1	1	1	1	temperature of GPT 3.5
N_k	3	3	3	3	3	3	number of answers of LLM in metaphor knowledge scaffolding

Table 7: Hyper-parameter values in our proposed methods.

Method	Template Prompt
Standard zero-shot	Decide whether the word "[word]" in the sentence "[sentence]" is used metaphorically.
Zero-shot CoT	Decide whether the word "[word]" in the sentence "[sentence]" is used metaphorically. Let's think step by step.
Plan-and-Solve	Decide whether the word "[word]" in the sentence "[sentence]" is used metaphorically. Let's first understand the problem and devise a plan to solve the problem. Then, let's carry out the plan and solve the problem step by step.
Self-refine	*Prompt for Initial Generation:* Decide whether the word "[word]" in the sentence "[sentence]" is used metaphorically. *Prompt for Feedback:* Can you give suggestions to improve the above answer? If the answer is perfect, just return "NONE". *Prompt for Refine:* Okay, let's use this feedback to improve the above answer.
Self-consistency	Decide whether the word "[word]" in the sentence "[sentence]" is used metaphorically.
Least-to-most	*Prompt for Stage 1: Decompose Question into subquestions* Q: Is the word "[word]" in the sentence "[sentence]" used metaphorically? A: Let's break down this problem into subquestions: *Prompt for Stage 2: Sequentially Solve Subquestions* Q: Is the word "[word]" in the sentence "[sentence]" used metaphorically? A: Let's break down this problem into subquestions: [answer] Q: Please answer the above subquestions and give the final answer for the initial question "Is the word '[word]' in the sentence '[sentence]' used metaphorically?"

Table 8: Prompt design for LLM-based baselines. *[sentence]* denotes the input slot for the sentence in an instance from the test dataset and *[word]* denotes the input slot for the target word within this sentence. *[answer]* denotes the input slot for the answer of LLM to the previous question.

Theory	Template Prompt
CMT	According to conceptual metaphor theory, metaphor facilitates a mapping of attributes or characteristics from the source domain to the target domain. The source domain is a conceptual domain containing concepts that are typically concrete, tangible, and familiar to us. The target domain is a conceptual domain containing concepts that are typically vague and abstract. Based on the above information, answer the following question: In the sentence, "[sentence]", decide whether the word "[word]" is used metaphorically.
MIP	According to metaphor identification procedure, a metaphor is identified if the contextual meaning of the target word is different from its more basic meaning. Based on the above information, answer the following question: In the sentence, "[sentence]", decide whether the word "[word]" is used metaphorically.
SPV	According to selectional preference violation, a metaphor is identified by noticing the difference between the context of a target word and its frequently used contexts. Based on the above information, answer the following question: In the sentence, "[sentence]", decide whether the word "[word]" is used metaphorically.

Table 9: Prompt design for the variant of *w/o Metaphor Knowledge Graph*. *[sentence]* and *[word]* denote the input slots for the sentence in an instance and the target word within this sentence, respectively.