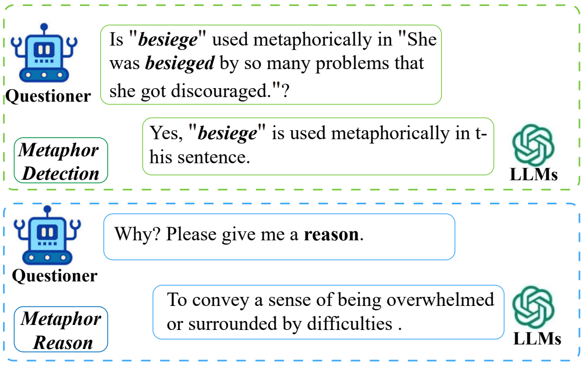
# Merely Judging Metaphor is Not Enough: Research on Reasonable Metaphor Detection

**Puli Chen[1#], Cheng Yang[1#], Qingbao Huang[1,2*]**
[1]School of Electrical Engineering, Guangxi University, Nanning, China
[2]Guangxi Key Laboratory of Multimedia Communications and Network Technology
{2312391007, 2212391065}@st.gxu.edu.cn, qbhuang@gxu.edu.cn

## Abstract

Metaphor, as an advanced form of cognition, is challenging to understand their meaning. Current metaphor detection tasks only provide labels (i.e., metaphor or literal) without interpreting how to understand them. In this paper, we improve the metaphor detection task and explore the reason of metaphor. To the best of our knowledge, we are the first work to reason about metaphor using mainstream Large Language Models (LLMs). Specifically, we utilized ChatGPT3.5 to expand the mainstream datasets in current metaphor detection, including VUA ALL, TroFi, and MOH-X. We input the original sentence, target word, and usage (metaphor or literal) into ChatGPT, guiding it to generate corresponding metaphor reason. Then, we designed supervised baseline experiments (e.g., RoBERTa, GPT-2) and zero-shot experiments with LLMs (e.g., LLaMA3). For the results generated by the above experiments, we provided the case study. We devised four methods that include manual evaluation to evaluate the reason performance of the model, and discussed extensively the advantages and disadvantages of these evaluation methods. Our code is available at https://github.com/yc-cy/Metaphorical-Reasoning.

## 1 Introduction



Is "***besiege***" used metaphorically in "She was ***besieged*** by so many problems that she got discouraged."?

Yes, "***besiege***" is used metaphorically in this sentence.

Why? Please give me a **reason**.

To convey a sense of being overwhelmed or surrounded by difficulties .

Figure 1: we added metaphor reason to the current metaphor detection task. Where the green dashed part indicates metaphor detection and the blue dashed part represents metaphor reason. The model needs to give the corresponding reason after judging the metaphor.

Metaphor is essentially a cognitive mechanism that exists in human thinking, used to construct conceptual frameworks (Lakoff and Wehling, 2012). In NLP, metaphor detection refers to determining whether a given target word is used metaphorically, given its context (Lakoff and Johnson, 2008; Choi et al., 2021). Considering an example of a metaphor detection task: "His voice is like heavenly music.". In this sentence, "heavenly music" is used metaphorically. By associating the term "heavenly music" with his voice, it conveys that his voice is extremely melodious. Detecting and understanding metaphor is crucial for downstream NLP tasks, including information extraction (Tsvetkov et al., 2013), sentiment analysis (Cambria et al., 2017), machine translation (Babieno et al., 2022), and seamless human-computer interaction (Rai and Chakraverty, 2020).

Traditional metaphor detection methods include using dependency trees (Le et al., 2020), employing prompt (Su et al., 2020), and multi-task learning (Choi et al., 2021). However, current metaphor detection tasks face generalization issues (see Experiment 1). The aforementioned metaphor detection methods are insufficient, because they only require the model to determine whether the target word is used metaphorically in context. Additional information is needed , such as metaphor reason, seeing Figure 1. In metaphor-related interpretation researches, previous work has only replaced the target word used metaphorically in the context with a literal word. However, this method also do not address the issue of metaphor reason.

To address the aforementioned issues, this paper

[#] These authors contributed equally to this work.
[*] Corresponding author.

*Findings of the Association for Computational Linguistics: EACL 2024*, pages 5850–5860
November 12-16, 2024 ©2024 Association for Computational Linguistics

adds reason to the current mainstream metaphor detection datasets, constructing an reason-based dataset. "Reason" in this context is the elaboration of a cause (Derrida et al., 1983; Feyerabend, 1987). First, we designed prompt 1 (see subsection 3.2), given three inputs to the LLMs: the target word, the original sentence, and the usage. Then, the LLMs reason about the meaning of the target word in the sentence. Inspired by ChatGPT's outstanding performance in zero-shot or few-shot NLP tasks (Meng et al., 2022; Yoo et al., 2021), and considering that manual annotation on crowdsourcing platforms is more costly (0.11 USD per instance (Yoo et al., 2021)) compared to using ChatGPT for reason annotation. Therefore, we used ChatGPT3.5 to generate metaphor reason (usage is given here, distinguished from experiments) and as the original reason data. In terms of experiments, we designed both supervised and zero-shot metaphor detection experiments. Traditional supervised metaphor detection only determines whether the target word is used metaphorically, thus being a binary classification. However, our proposed reason-based metaphor detection task requires the model to provide the reason behind the usage conclusion. Therefore, the supervised experiment uses the text generation model GPT-2. For the zero-shot experiment, we use LLMs for prediction, including current mainstream LLMs such as Gemma, ChatGPT series, and LLaMA. For the results of the aforementioned experiments, we provided two automatic evaluation methods and a manual evaluation method.

Overall, our contributions are summarized below:

1. To the best of our knowledge, we are the first to propose a LLMs-based metaphor reason task that requires the model not only to determine whether the target word is a metaphorical usage, but also to reason about that result.

2. We constructed the corresponding metaphor reason dataset using ChatGPT3.5.

3. We are the first to explore various methods for evaluating metaphor reason, including traditional automatic evaluation, evaluation based on fine-tuning models with entailment datasets, evaluation based on ChatGPT, and manual evaluation. We provide a detailed analysis of the advantages and disadvantages of these four methods.

## 2 Related Work

### 2.1 Metaphor Detection

Current metaphor detection methods can be divided into two major categories: supervised and unsupervised. In the supervised direction, some studies Song et al. (2021); Feng and Ma (2022) focus on extracting subject-verb-object (SVO) relations from dependency trees to aid in metaphor detection. Song et al. (2021) processes the SVO outputs in the text through combination, averaging, and maximization to further capture the association between structural semantics, while Feng and Ma (2022) uses a BERT Decoder to generate the start and end positions of SVO based on the context. In recent work, Li et al. (2023) integrates FrameNet to detect metaphors through explicit learning. In the unsupervised direction, Shutova et al. (2016) uses visual features for metaphor detection, comparing the cosine similarity between single word embeddings and phrase embeddings, and judging as a metaphor phrase when it falls below a certain threshold. Unlike (Shutova et al., 2016), Li et al. (2013); Bollegala and Shutova (2013) use big data-driven approaches to determine candidate source domains for metaphors. Recently, Wachowiak and Gromann (2023) has started using GPT-3 to detect metaphorical language in given sentences and target domains without any preset domains and to predict the source domain of the metaphor. Goren and Strapparava (2024) evaluated the performance of GPT-3.5 in a zero-sample setting through word-level metaphor detection, while Chandra et al. (2024) detected religious metaphors (e.g., Bhagavad Gita and the Holy Bible) using LLMs.

### 2.2 Prompt Learning

Unlike traditional fine-tuning methods, prompt learning aims to guide LLMs to generate specific content without fine-tuning. In this task, LLMs act as few-shot or zero-shot learners. Past research on prompt learning typically falls into two categories: generating annotations and generating samples. Ye et al. (2022); Meng et al. (2022) employed the method of adding polarity labels in prompts to guide the model to generate content related to the specified inclination. Wang et al. (2021) proposed a method combining human annotation and LLM annotation to reduce costs. Yoo et al. (2021) designed a template to guide the model to annotate or generate samples by introducing instances of different

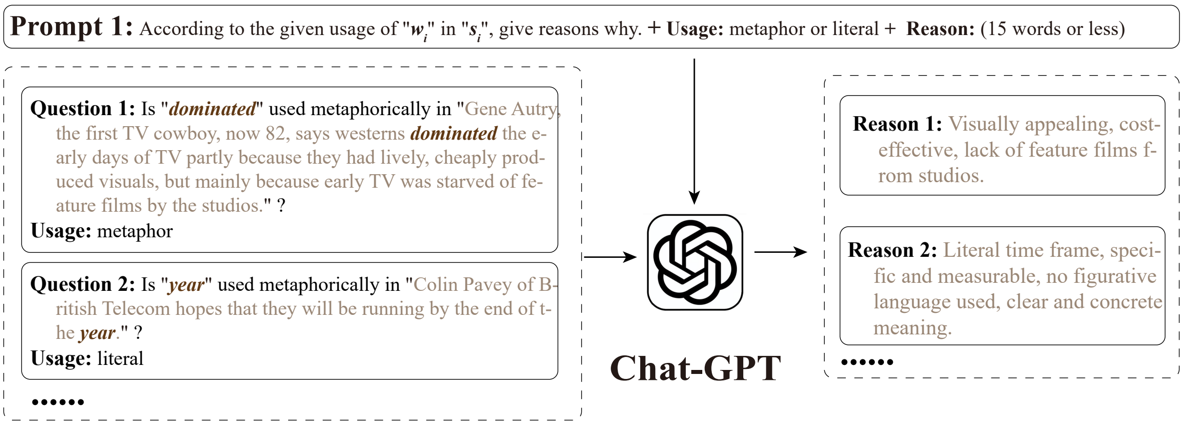

Prompt 1: According to the given usage of "$w_i$" in "$s_i$", give reasons why. + Usage: metaphor or literal + Reason: (15 words or less)
Question 1: Is "dominated" used metaphorically in "Gene Autry, the first TV cowboy, now 82, says westerns dominated the early days of TV partly because they had lively, cheaply produced visuals, but mainly because early TV was starved of feature films by the studios." ?
Usage: metaphor
Question 2: Is "year" used metaphorically in "Colin Pavey of British Telecom hopes that they will be running by the end of the year." ?
Usage: literal
••••••
Chat-GPT
Reason 1: Visually appealing, cost-effective, lack of feature films from studios.
Reason 2: Literal time frame, specific and measurable, no figurative language used, clear and concrete meaning.
••••••

Figure 2: We input the original sentence $s_i$, target word $w_i$, and usage into ChatGPT3.5. Guided by **Prompt 1**, ChatGPT3.5 generates the corresponding reason.

tasks. Lang et al. (2022) designed a joint training framework for GPT-3 and BERT for annotation in classification tasks. Recently, Khattak et al. (2023) improved the coherence between visual and verbal representations through prompt learning.

# 3 Dataset Construction

## 3.1 Metaphor Datasets

**VUA ALL:** VUA ALL has been applied to the shared task of metaphor detection (Leong et al., 2018, 2020), annotating all real-meaning words (including adjective, verb and noun) in a sentence.
**TroFi:** TroFi (Birke and Sarkar, 2006) is a dataset focused on verb metaphor detection. The dataset consists of 3717 samples.
**MOH-X:** The MOH (Mohammad et al., 2016) dataset focuses on verb metaphor detection and consists of 1639 sentences extracted from WordNet. MOH-X(Shutova et al., 2016) is a subset of the MOH dataset.

## 3.2 Data Reason Generation

We utilized ChatGPT3.5 to generate reason for traditional metaphor datasets VUA ALL, MOH-X, and TroFi. The reason generation process is shown in Figure 2.

We processed the original metaphor datasets. The processed datasets include three parts: the original sentence $s_i$, the target word $w_i$, and the usage (metaphor or literal). Based on the metaphor reason task of this paper, we carefully designed a prompt 1 suitable for this task. We used the processed metaphor datasets and prompt 1 as inputs to ChatGPT3.5, guiding it to explain the input data and output the reason. Considering that ChatGPT 3.5 has been provided with the usage instructions, the reasoning produced in this section will serve as the foundational reference data. For the expanded VUA ALL, MOH-X, and TroFi datasets, we divided them into training, testing, and validation sets in a ratio of 0.7, 0.15, and 0.15. Table 1 shows the statistics of the metaphor reason data after the expansion.

## 3.3 Manual Evaluation

We employ three volunteers with a background in metaphor to examine and screen the multiple reasons to determine if they accurately and completely convey the meaning of the original data sample. The task of each volunteer is to verify: 1) whether the metaphor reason of ChatGPT3.5 is contextualized; 2) whether the reason is complete and correct. After verification, the volunteer will choose the most appropriate one of the reasons as the final output of the model.

| Usage | VUA ALL_R | TroFi_R | MOH-X_R |
|---|---|---|---|
| **metaphor** | 11721 | 1607 | 314 |
| **literal** | 68153 | 2130 | 333 |
| **total** | 79874 | 3737 | 647 |

Table 1: Metaphor reason dataset statistics. Among them, we inform ChatGPT3.5 about the usage, including "metaphor" and "literal". "**R**" stands for "**Reason**".

# 4 Experiment

## 4.1 Baseline Models

**BERT**: BERT(Devlin et al., 2018) used a bidirectional Transformer encoder and came in two versions: base and large. **RoBERTa**: Unlike BERT,

| Model | VUA ALL | | | TroFi | | | MOH-X | | | V to T | | | V to M | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| **BERT-b** | 0.675 | 0.679 | 0.677 | 0.624 | 0.775 | 0.691 | 0.676 | 0.820 | 0.741 | 0.522 | 0.709 | 0.601 | 0.712 | 0.773 | **0.741** |
| **X-RoB** | 0.668 | 0.690 | 0.703 | 0.644 | 0.730 | 0.687 | 0.683 | 0.803 | 0.767 | 0.530 | 0.743 | 0.611 | 0.709 | 0.722 | 0.715 |
| **RoB-b** | 0.706 | 0.687 | 0.697 | 0.638 | 0.762 | **0.695** | 0.749 | 0.833 | 0.789 | 0.533 | 0.705 | 0.607 | 0.723 | 0.713 | 0.718 |
| **DeB-b** | 0.722 | 0.694 | **0.707** | 0.593 | 0.811 | 0.685 | 0.753 | 0.853 | **0.800** | 0.528 | 0.755 | **0.622** | 0.702 | 0.753 | 0.727 |
| **GPT2-b** | 0.642 | 0.551 | 0.593 | 0.709 | 0.571 | **0.632** | 0.618 | 0.700 | 0.656 | 0.534 | 0.651 | 0.586 | 0.609 | 0.540 | 0.572 |
| **GPT2-l** | 0.700 | 0.592 | **0.641** | 0.716 | 0.555 | 0.625 | 0.665 | 0.767 | **0.712** | 0.519 | 0.711 | **0.600** | 0.704 | 0.633 | **0.667** |

Table 2: Performance of supervised models on binary metaphor detection. The selected models include BERT-base (**BERT-b**), XLM-RoBERTa (**X-RoB**), RoBERTa-base (**RoB-b**), DeBERTa-base (**DeB-b**), GPT2-base (**GPT2-b**), and GPT2-large (**GPT2-l**). The metaphor detection datasets include VUA ALL, TroFi, and MOH-X, evaluated using within-dataset detection. Additionally, we conducted cross-dataset detection, where "**V to T**" indicates training on VUA ALL and testing on TroFi, and "**V to M**" means training on VUA ALL and testing on MOH-X. The evaluation metrics include Precision (**P**), Recall (**R**), and F1 score (**F1**), with F1 score being the core metric.

RoBERTa removed the NSP (Next Sentence Prediction) task during pre-training, which means it no longer determined whether two sentences were adjacent. **DeBERTa**: DeBERTa(He et al., 2020), an improvement over BERT, was introduced in 2020. DeBERTa incorporated enhanced decoding mechanisms and disentangled attention mechanisms. **ChatGPT**: The ChatGPT series includes language models developed by OpenAI, such as GPT-2 and ChatGPT3.5, which can be accessed via API calls. **LLaMA**: LLaMA is a series of large language models developed by Meta (Facebook), including 8B and 70B models. LLaMA 3-70B, one of the significant models, has 70 billion parameters, and its weights can be requested from the official website[1]. **Gemma**: a natural language processing model developed by OpenAI, based on the Transformer architecture, similar to the GPT.

### 4.2 Experimental Design

**Experiment 1:** We designed a supervised experiment with binary classification, employing two types of model architectures. The first type is Masked Language Models (MLM), including BERT-base, RoBERTa-base, and DeBERTa-base. The second type is Causal Language Models (CLM), including the base and large versions of GPT-2. For the $i$-th sample $n_i \in N$, target word $w_i$, and context $s_i$, MLM has:

$$\hat{y}_i = MLM(s_i, w_i)[0]$$

where $\hat{y}_i$ is the output predicted by the MLM model at the corresponding CLS position. For CLM, similarly:

$$\hat{y}(t = t_0) = CLM(s_i, w_i, t < t_0)$$

where $\hat{y}(t = t_0)$ represents the $t_0$-th token generated by the CLM.

Compared to MLM models, CLM models first need to determine whether the target word is a metaphor and then provide an reason based on this judgment. We first performed experiments on VUA ALL, TroFi, and MOH-X. Subsequently, we performed experiments across datasets (i.e., transfers from VUA ALL to TroFi and from VUA ALL to MOH-X). The binary classification experiments aim to explore the generalization performance of the model on a metaphor detection task.

**Experiment 2:** We also conducted binary classification zero-shot metaphor detection experiments on other prompt-based LLMs. These models include Gemma7B, Llama3-8B, Llama3-70B, and ChatGPT3.5. The experiments were also carried out on the VUA ALL, TroFi, and MOH-X datasets. The prompt 2 we designed for the LLMs is as follows:

*Determine whether "$w_i$" is used metaphorically or literally in "$s_i$" and give reasons why.*
***Usage:*** *(metaphor or literal, judge by models)*
***Reason:*** *(15 words or less)*

The main difference between prompt 2 and prompt 1 lies in whether specific usage instructions are provided. In prompt 1, we clearly presented the specific usage based on previous labels and guide

[1] https://github.com/meta-llama/llama3

| Model | VUA ALL | | | TroFi | | | MOH-X | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| **LLaMA3-8B** | 0.214 | 0.455 | 0.291 | 0.580 | 0.409 | 0.479 | 0.914 | 0.430 | 0.584 |
| **Gemma-7B** | 0.160 | **0.997** | 0.276 | 0.443 | **0.976** | 0.609 | 0.474 | **0.993** | 0.642 |
| **ChatGPT3.5** | 0.292 | 0.443 | 0.352 | **0.589** | 0.446 | 0.508 | **0.948** | 0.487 | 0.643 |
| **LLaMA3-70B** | **0.341** | 0.622 | **0.440** | 0.566 | 0.763 | **0.650** | 0.908 | 0.792 | **0.846** |

Table 3: Metaphorical binary classification detection performance of LLMs in other prompt classes. The selected LLMs include LLaMA3-8B, Gemma7B, LLaMA3-70B, and ChatGPT 3.5. The metaphor detection datasets include VUA ALL, TroFi, and MOH-X. The evaluation metrics include precision (**P**), recall (**R**) and composite metric (**F1**), where **F1** is the core metric.

ChatGPT3.5 to reason under the known label conditions. In prompt 2, for the $i$-th sample $n_i \in N$, which includes the target word $w_i$ in the context $s_i$, we required the model to independently judge the usage of the word and provide corresponding reasoning.

## 5 Implementation

For the baseline in the experiments, the epoch of MLM is set to 25, the learning rate is initialized to 3e-5, and the metaphor weight is given to 5. The epoch of CLM is set to 50. Both MLM and CLM are initialized using the weight parameter of Huggingface library. The hidden layer unit of the classifier is set according to the size of the model, which is set to 768 for the base model and 1024 for the large model. The experiments are run on a cloud server with a single A100 80G GPU.

## 6 Experimental Analysis

**Experiment 1:** The experimental results are shown in Table 2. The results of the MLM experiments indicate that current supervised methods based on binary metaphor detection have achieved high performance. Specifically, the F1 score on MOH-X even reaches 0.8. Theoretically VUA ALL fine-tuning has better model learning ability. However, despite the fact that the VUA ALL training samples are 30 times larger than the TroFi training samples and contain more types, generalization experiments ("**V to T**" and "**V to M**") show significant decreases in performance on the metrics (e.g., on F1 and V to T, -9% on BERT and -8.8% on RoBERTa and -6.3% on DeBERTa). Therefore, there are issues with the current binary metaphor detection task in terms of generalization. In contrast, the classification with reason generation task based on CLM, designed by us, enhances the model's generalization performance to some extent (e.g., on F1 and VUA ALL to TroFi, -4.6% on GPT2-b and -2.5% on GPT2-l).

Furthermore, compared to the single binary classification of MLM, CLM with reason shows a decrease in performance on all three datasets (e.g., on F1, GPT2-b 0.593 vs. RoB-b 0.697 on VUA ALL). The experimental results indicate that the task of adding reason significantly increases the difficulty for models.

**Experiment 2:** Experimental results are shown in Table 3. Compared to the supervised methods in Table 2, all LLMs except LLaMA3-70B exhibit significant performance declines (e.g., on F1, DeB-b 0.707 vs. ChatGPT 0.352 on VUA ALL). Even on the more extensive and complex VUA ALL dataset, LLaMA3-70B's performance is still unsatisfactory (e.g., LLaMA3-70B 0.440 vs. DeB-b 0.707). These results indicate that current LLMs still have shortcomings in metaphor detection tasks. Furthermore, different LLMs exhibit varying judgment strategies in metaphor tasks. For example, Gemma-7B has a higher accuracy in detecting metaphorical samples but performs poorly in detecting literal samples, reflected in its high recall and low precision. In contrast, LLaMA and ChatGPT show relatively balanced performance in both aspects. Although Gemma outperforms ChatGPT on the TroFi dataset, its performance significantly drops on the VUA ALL and MOH-X datasets. This is mainly due to its very low precision (e.g., Gemma 0.160 vs. ChatGPT 0.292 on VUA ALL and Gemma 0.474 vs. ChatGPT 0.948 on MOH-X).

| Model | VUA ALL | | | | TroFi | | | | MOH-X | | | | VUAverb | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | B1 | R1 | RL | M | B1 | R1 | RL | M | B1 | R1 | RL | M | B1 | R1 | RL | M |
| **GPT2-large** | **0.27** | **0.37** | **0.33** | **0.25** | **0.19** | **0.29** | **0.25** | **0.18** | 0.21 | 0.31 | 0.27 | 0.19 | 0.264 | 0.37 | 0.329 | 0.25 |
| **ChatGPT3.5** | 0.22 | 0.34 | 0.3 | 0.21 | 0.17 | **0.29** | **0.25** | 0.16 | **0.23** | **0.38** | **0.31** | **0.23** | 0.22 | 0.341 | 0.296 | 0.207 |
| **LLaMA3-70B** | 0.21 | 0.32 | 0.28 | 0.18 | 0.14 | 0.23 | 0.2 | 0.12 | 0.21 | 0.34 | 0.29 | 0.19 | 0.199 | 0.304 | 0.262 | 0.174 |

Table 4: Comparison of the reason performance of the supervised learning method GPT2-large with the LLMs methods ChatGPT3.5 and LLaMA3-70B. The evaluation metrics include BLEU-1 (**B1**), ROUGE-1 (**R1**), ROUGE-L (**RL**), and METEOR (**M**). All three metrics use the n-gram matching mechanism. Where "**1**" indicates the exact match of a single word, and "**L**" considers longer texts.

## 7 Automatic Evaluation Experiment

In comparison to the reason generated from the original dataset in section 3.2, we used automatic evaluation methods to assess the supervised method GPT2-large, as well as the LLMs methods ChatGPT3.5 and LLaMA3-70B. We employed three evaluation metrics: BLEU, METEOR, and ROUGE. All three metrics use n-gram matching mechanisms but differ slightly in the factors they consider. Specifically, BLEU and ROUGE emphasize precision and recall, respectively, while METEOR additionally takes into account synonyms and stems.

The automatic evaluation results are shown in Table 4. From the table, we can see that the supervised fine-tuned GPT2-large model achieves the highest scores on both VUA ALL and TroFi, indicating that supervised methods generate answers more similar to the original dataset compared to the zero-shot reason of LLMs. Furthermore, from VUA ALL to TroFi, and then to MOH-X, the performance gap between LLMs methods (i.e., ChatGPT3.5 and LLaMA3) and the supervised method GPT2 gradually narrows. By the time of MOH-X, the zero-shot ChatGPT3.5 method has already surpassed the supervised GPT2 method in all respects. Compared to TroFi and MOH-X, VUA ALL contains more metaphor categories and richer samples. Therefore, GPT2 fine-tuned on VUA ALL performs better on the test set, demonstrating that increasing the number of samples helps improve the quality of metaphor reason generated by supervised methods.

## 8 Evaluation Experiment of Text-embedded Fine-tuning Model

The fine-tuning model evaluation method first trains the RoBERTa-large model on entailment datasets. Then, it evaluates the results generated by GPT2-large, ChatGPT3.5, and LLaMA3. The advantage of this method is that, compared to direct vocabulary distribution calculation in automatic evaluation methods, the fine-tuned model often contains certain high-dimensional semantic information. In this experiment, we used two entailment datasets: *Semantic Textual Similarity Benchmark (**STS-B**)* and *Sentences Involving Compositional Knowledge (**SICK**)*. The STS-B is a similarity and paraphrase dataset, consisting of sentence pairs extracted from news headlines, video titles, image captions, and natural language inference data, each annotated by humans. The SICK dataset is used for compositional distributional semantics. It includes a large number of sentence pairs that exhibit rich lexical, syntactic, and semantic phenomena. Each pair of sentences is annotated with two dimensions: relatedness and entailment. We only consider relatedness, with scores ranging from 1 to 5.

The experimental results are shown in Table 5. We found that the evaluation method based on fine-tuned models yielded similar results to the automatic evaluation methods (see Table 3). Firstly, compared to STS-B, we observed that the model fine-tuned on SICK provided relatively higher evaluation results. Moreover, whether using STS-B or SICK, the evaluation results for literal usage were always higher than those for metaphorical usage across all three datasets. This indicates that both supervised and LLMs methods have a better understanding of literal usage compared to metaphorical usage. Secondly, as the sample size decreases (i.e., from VUA ALL to MOH-X), the evaluation results of supervised methods are gradually surpassed by LLMs methods (e.g., on STS-B, ChatGPT3.5 2.56 vs. GPT2-large 2.67 on VUA ALL). This further confirms that the sample size of the dataset somewhat affects the quality of metaphor reason generated by supervised methods.

| Model | VUA ALL | | | | | | TroFi | | | | | | MOH-X | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | STS-B | | | SICK | | | STS-B | | | SICK | | | STS-B | | | SICK | | |
| | Met | Lit | Wtd | Met | Lit | Wtd | Met | Lit | Wtd | Met | Lit | Wtd | Met | Lit | Wtd | Met | Lit | Wtd |
| **GPT2-large** | **2.25** | **2.74** | **2.67** | 2.97 | **3.39** | **3.33** | **2.26** | 2.39 | 2.33 | **2.90** | 3.05 | 2.99 | **2.61** | 2.51 | 2.56 | 3.18 | 3.27 | 3.23 |
| **ChatGPT3.5** | 2.21 | 2.62 | 2.56 | **2.99** | 3.28 | 3.24 | 2.14 | **2.58** | **2.39** | 2.89 | **3.24** | **3.09** | 2.51 | **3.18** | **2.87** | **3.19** | **3.64** | **3.43** |
| **LLaMA3-70B** | 2.05 | 2.52 | 2.44 | 2.90 | 3.29 | 3.22 | 1.99 | 2.18 | 2.10 | 2.79 | 3.03 | 2.93 | 2.48 | 3.02 | 2.77 | 3.09 | 3.56 | 3.34 |

Table 5: Training the RoBERTa-large model on the datasets **STS-B** and **SICK**, we utilize RoBERTa-large to evaluate the reason performance of GPT2-large, ChatGPT3.5, and LLaMA3-70B on the metaphor datasets VUA ALL, TroFi, and MOH-X, with a scale of 1-5. Where "**Met**" denotes metaphor, "**Lit**" denotes literal, and "**Wtd**" denotes weighted sum (i.e., weighted by the percentage of metaphor samples).

| Model | VUA ALL | | | TroFi | | | MOH-X | | | VUAverb | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Met | Lit | Wtd | Met | Lit | Wtd | Met | Lit | Wtd | Met | Lit | Wtd |
| **GPT2-large** | 3.66 | 4.00 | 3.95 | 3.69 | 3.80 | 3.75 | 3.62 | 3.74 | 3.69 | 3.67 | 3.95 | 3.89 |
| **ChatGPT3.5** | 3.76 | **4.02** | **3.98** | 3.81 | **4.10** | **3.97** | 3.66 | **4.67** | 4.20 | 3.72 | **4.03** | **3.96** |
| **LLaMA3-70B** | **3.87** | 3.93 | 3.92 | **3.92** | 3.88 | 3.90 | **3.91** | 4.49 | **4.22** | **3.85** | 3.83 | 3.83 |

Table 6: Evaluation of ChatGPT's inference performance on GPT2-large, ChatGPT3.5, and LLaMA3-70B on the metaphorical datasets VUA ALL, TroFi, and MOH-X, with a scale of 1-5. Where "**Met**" denotes **Metaphor**, "**Lit**" stands **Literal**, and "**Wtd**" indicates a **Weighted** (i.e., weighted by the proportion of metaphor samples).

## 9 ChatGPT Evaluation Experiment

The ChatGPT evaluation method aims to use ChatGPT to score the "**similarity**" between the original dataset output and the prediction model output. Its advantage lies in the fact that ChatGPT has been trained on a broader and higher-quality dataset, enabling it to understand more complex semantic information and relationships. For the $i$-th sample $n_i \in N$, target word $w_i$, and context $s_i$, the prompt for ChatGPT is designed as follows:

*Discuss the use of "$w_i$" in "$s_i$".*
***Answer:*** $a_i$
***Prediction:*** $a_{i,j}$
*Rate the prediction based on the answer (1 to 5).*
***Output:***

Where $a_i$ is the correct answer for the $i$-th sample $n_i$, and $a_{i,j}$ is the prediction of the $j$-th model for the $i$-th sample. Similar to automatic evaluation and fine-tuned model evaluation, ChatGPT will score the predictions based on their similarity on a scale of 1 to 5. The experimental results are shown in Table 6. Unlike the automatic evaluation method and the fine-tuned model evaluation method, the GPT2-large model, which originally represented the supervised approach, showed superior performance on the VUA ALL evaluation compared to LLMs. However, the ChatGPT evaluation results were the opposite. There was even a significant discrepancy in the reason of metaphor results (e.g., GPT2-large 3.69 vs. LLaMA 3.92 on TroFi). This indicates that similarity in lexical distribution (automatic evaluation) and possessing certain semantic information (fine-tuned model evaluation) do not necessarily determine the quality of generated metaphor reason. This also reveals the shortcomings of automatic evaluation methods and fine-tuned model evaluation methods in judging the similarity of metaphor reason. Additionally, similar to the previous evaluation methods, the ChatGPT

| Model | VUA ALL | TroFi | MOH-X |
|---|---|---|---|
| **GPT2-large** | 3.48 | 3.38 | 3.65 |
| **ChatGPT3.5** | 3.83 | **3.97** | 3.88 |
| **LLaMA3-70B** | **3.90** | 3.85 | **3.94** |

Table 7: Manual evaluation results. We invite three volunteers to evaluate the reasoning performance of GPT2-large, ChatGPT3.5, and LLaMA3-70B on the metaphorical datasets VUA ALL, TroFi, and MOH-X. The evaluation metric is "**similarity**" with a scale of 1-5.

| Dataset | Sentence | Reason |
|---|---|---|
| **VUA ALL (destroy)** | **1.** Israeli warplanes destroyed a PLO base in a Druse - held hilltop above Beirut in Israel 's deepest air raid into Lebanon this year. (Usage:literal)<br>**2.** Public reaction to this melodrama nearly destroyed the administration and solidified the belief that no moderates existed in Iran. (Usage:metaphor) | **Reason 1.** Military objective, strategic necessity, neutralizing threat, eliminating enemy presence.<br>**Reason 2.** Intense backlash, loss of credibility, polarized public opinion. |
| **TroFit (absorb)** | **1.** Equally important, his Socialist government absorbed and then , in effect, destroyed the French Communist Party. (Usage:literal)<br>**2.** In a small company, the CEO and his people become so absorbed in the Eleven that it dominates the existence of the company. (Usage:metaphor) | **Reason 1.** To eliminate competition and consolidate power within the Socialist government.<br>**Reason 2.** Because they are fully focused and engrossed in the project, to the point of obsession. |
| **MOH-X (besiege)** | **1.** The Turks besieged Vienna. (Usage:literal)<br>**2.** She was besieged by so many problems that she got discouraged. (Usage:metaphor) | **Reason 1.** Historical event, literal military action, and time period context.<br>**Reason 2.** To convey a sense of being overwhelmed or surrounded by difficulties . |

Table 8: Partial reason data. We selected some data from VUA ALL, TroFit, and MOH-X for analysis, including the original sentence and the corresponding reason.

evaluation method also confirmed the importance of data quantity and the phenomenon that models perform better in explaining literal usage than metaphorical usage (e.g., on Wtd GPT2-large 3.95 vs. ChatGPT 3.98 on VUA ALL).

## 10 Manual Evaluation Experiment

Finally, we used a manual method for "**similarity**" evaluation. Compared to the previous three methods, manual evaluation can comprehensively consider aspects such as semantics, fluency, and logic, thus providing more accurate results. To reduce subjectivity, we invited three volunteers to independently evaluate the correct answers and the model-predicted answers, with a scoring range of 1-5. The results from the three volunteers were then averaged to provide the final evaluation score for each prediction.

The experiment results are detailed in Table 7. It can be observed that the supervised method GPT2-large achieved relatively good scores in the first three evaluation methods. However, it performed poorly in the manual evaluation (e.g., GPT2-large 3.48 vs. LLaMA3 3.90 on VUA ALL). On the one hand, combining the results of the binary classification supervised experiment and the zero-shot experiment, the supervised method often performs better in usage judgment (e.g., GPT2-large 0.647 vs. LLaMA3 0.440 on VUA ALL). Therefore, the first three evaluation methods might be influenced by preconceived notions.

## 11 Case Study

As shown in Table 8, we have selected data from the VUAPOS, TroFi, and MOH-X datasets, covering both literal and metaphor usage. Taking the verb “absorb” as an example, we provide the corresponding example sentences and their usage. In previous datasets, the usage of “absorb” was only labeled without explaining its specific reason. For example, in the sentence “Equally important, his Socialist government absorbed and then , in effect, destroyed the French Communist Party.”, if the LLM reasons “absorb” as a literal use, it might further infer that “absorb” means “To eliminate......the Socialist government.” In addition, LLM’s analysis of the metaphorical use of “absorb” suggests that the verb in the sentence refers to people devoting themselves to the project to the point of obsession. Similarly, for the verbs “besiege” and “destroy”, LLM was able to infer the reason of the target word in different sentences.

## 12 Conclusion

As an advanced approach, current textual metaphor detection tasks typically provide usage labels. However, this method falls short in understanding metaphors. Metaphor research requires not only detection but also in-depth reasoning. To our knowledge, we are the first to utilize LLMs for metaphor reason to better understand metaphorical meanings. Firstly, we expanded existing traditional metaphor datasets using the ChatGPT and provided results for metaphor reason. Secondly, we designed supervised experiments demonstrating that current metaphor detection models have poor generalization performance, while incorporating metaphorical reasons significantly improves the models’ generalization ability. Additionally, to evaluate LLMs performance in metaphor reason, we proposed four novel evaluation methods. Overall, our experimental results indicate that LLMs can reduce resource consumption for data reasoning (compared to crowdsourcing) and show significant room for improvement.

## Limitations

This paper builds on metaphor detection to propose the task of metaphor reason, further investigating the meanings conveyed by metaphors. Currently, we have only expanded traditional metaphor datasets, but some metaphors may have become literal over time. In future work, we aim to construct novel and high-quality interpretable metaphor datasets.

## Ethics Statement

In this study, we strictly adhere to academic and research ethical guidelines, emphasizing transparency and openness of information. We meticulously and clearly cite all publicly available data sources used, fully respecting the contributions of original researchers and data providers in the metaphor identification field. Throughout the research process, we maintain a respectful attitude towards the work of others, avoiding any form of malicious criticism or plagiarism. Our research methods and practices strictly adhere to the principles of academic integrity, aiming to ensure the full acknowledgment and respect of previous work. At every stage of the research, we uphold the requirements of academic ethics, committed to ensuring the authenticity, transparency, and fairness of the research.

## Acknowledgments

We would like to thank all reviewers for their insightful comments and suggestions to help improve the paper. This work was supported by the National Natural Science Foundation of China (62276072), Guangxi Natural Science Foundation Key Project (Application No. 2024JJD170001), Guangxi Natural Science Foundation (2022GXNSFAA035627), the Open Research Fund of Guangxi Key Laboratory of Multimedia Communications and Network Technology, and Innovation Project of Guangxi Graduate Education (JGY2023016).

## References

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. Miss roberta wilde: Metaphor identification using masked language model with wiktionary lexical definitions. *Applied Sciences*, 12(4):2081.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European chapter of the association for computational linguistics*, pages 329–336.

Danushka Bollegala and Ekaterina Shutova. 2013. Metaphor interpretation using paraphrases extracted from the web. *PloS one*, 8(9):e74304.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Rohitash Chandra, Abhishek Tiwari, Naman Jain, and Sushrut Badhe. 2024. Large language models for metaphor detection: Bhagavad gita and sermon on the mount. *IEEE Access*.

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. *arXiv preprint arXiv:2104.13615*.

Jacques Derrida, Catherine Porter, and Edward P Morris. 1983. The principle of reason: The university in the eyes of its pupils. *diacritics*, 13(3):3–20.

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*.

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 Association for Computational Linguistics: EMNLP 2022*, pages 656–667.

Paul Feyerabend. 1987. *Farewell to reason*. Verso.

Gamze Goren and Carlo Strapparava. 2024. Context matters: Enhancing metaphor recognition in proverbs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3825–3830.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

George Lakoff and Elisabeth Wehling. 2012. *The little blue book: The essential guide to thinking and talking democratic*. Simon and Schuster.

Hunter Lang, Monica N Agrawal, Yoon Kim, and David Sontag. 2022. Co-training improves prompt-based learning for large language models. In *International Conference on Machine Learning*, pages 11985–12003. PMLR.

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*, volume 34, pages 8139–8146.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 vua and toefl metaphor detection shared task. In *Proceedings of the second workshop on figurative language processing*, pages 18–29.

Chee Wee 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.

Hongsong Li, Kenny Q Zhu, and Haixun Wang. 2013. Data-driven metaphor recognition and explanation. *Transactions of the Association for Computational Linguistics*, 1:379–390.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loïc Barrault. 2023. Framebert: Conceptual metaphor detection with frame embedding learning. *arXiv preprint arXiv:2302.04834*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the fifth joint conference on lexical and computational semantics*, pages 23–33.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 160–170.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. 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 4240–4251.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the second workshop on figurative language processing*, pages 30–39.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51.

Lennart Wachowiak and Dagmar Gromann. 2023. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*.