Consistency Rating of Semantic Transparency: an Evaluation Method for Metaphor Competence in Idiom Understanding Tasks

Hui Gao*, **Jing Zhang***, **Peng Zhang**†, **Chang Yang**
College of Intelligence and Computing, Tianjin University, Tianjin, China
{hui_gao, pzhang}@tju.edu.cn

Abstract

Idioms condense complex semantics into fixed phrases, and their meaning is often not directly connected to the literal meaning of their constituent words, making idiom comprehension a test of metaphor competence. Metaphor, as a cognitive process in human beings, has not yet found an effective evaluation method to assess the metaphor competence of LLMs (Large Language Models). In this paper, we propose a method to evaluate the metaphor competence of LLMs for the idiom understanding task: the **C**onsistency **R**ating of **S**emantic **T**ransparency (**CR-ST**). This strategy assesses the difficulty of understanding idioms through two dimensions: overall semantic transparency and constituent semantic transparency, aiming to gauge LLMs' mastery of metaphor competence. Subsequently, we introduce a prompt mechanism-**P**araphrase **A**ugmentation **S**trategy with **S**elf-checking (**PASS**), based on human language logic, which guides the model to enhance its metaphor competence by explicitly generating idiom paraphrases. We conducted a baseline evaluation of seven LLMs on the CINLID and ChID datasets and analyzed the effectiveness of PASS on different subsets of semantic transparency. The experimental results demonstrate that LLMs can achieve performance comparable to PLMs (Pre-trained Language Models) without additional training, and PASS has a positive effect on the metaphor competence of LLMs.

1 Introduction

Idioms are expressions whose meanings are not deducible from the literal meanings of the individual words. They often originate from ancient stories or customary usage. Compared to ordinary text, the meanings of idioms are deeper and often hidden within the literal meanings of the words, referred to as metaphorical meanings (Hu, 2023). For instance,

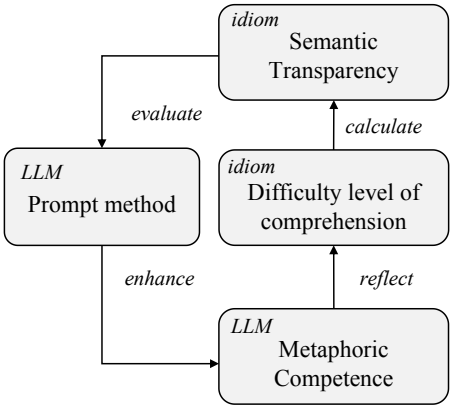


Figure 1: The goal of this work is to evaluate the metaphor competence of through an evaluation of idiom transparency consistency. The idiom with lower transparency require higher metaphorical abilities from LLMs. Thus, we can reflect the metaphor competence of LLMs by evaluating their performance on idioms of varying difficulty.

consider the idiom "望梅止渴 (wàng méi zhǐ kě)". Its literal meaning is *"To quench thirst by thinking of plums."*, while its metaphorical meaning, *"Comfort yourself with unrealistic fantasies,"* stems from a historical allusion. Understanding idioms requires humans to possess metaphor competence — the ability to move from literal to metaphorical meanings, which poses a significant challenge for LLMs (Orifjonovich, 2023; Julich-Warpakowski and Pérez Sobrino, 2023).

Metaphor competence represents a sophisticated aspect of human language cognition, and idiom comprehension serves as a criterion for assessing whether LLMs possess this ability. In cognitive linguistics, Lakoff and Johnson (2008); Shenghuan (2022) argue that the deep cognitive motivation behind the structure and expressive function of idioms lies in the metaphorical nature of human thought, which allows humans to transition from literal to metaphorical meanings. Research in applied psy-

*These authors contributed equally.
†Corresponding author: Peng Zhang

Proceedings of the 31st International Conference on Computational Linguistics, pages 10460–10471
January 19–24, 2025.

chology indicates that human idiom comprehension involves dynamic processing mechanisms influenced by factors such as idiom difficulty, the user's cultural background, and contextual information (Fang Yuanyuan and Xinchun, 2023).

In recent years, LLMs represented by GPT (Liu et al., 2023) and Llama (Insuasti et al., 2023) have sparked a new wave of technological innovation, significantly enhancing various NLP tasks such as text classification, question answering, and natural language generation (Shi et al., 2023; Tan et al., 2023; Qin et al., 2024; Singhal et al., 2023). However, based on cognitive linguistic theories, how to evaluate and enhance the metaphor competence of LLMs remains an unexplored issue. This work, based on cognitive linguistic theory and combined with prompt techniques, investigates the two points:

(1) Evaluating Metaphor Competence: Semantic Transparency. We measure the metaphor competence of LLMs by evaluating the difficulty level of idiom comprehension, specifically calculating the semantic transparency of idioms. Traditional methods of semantic transparency typically rely on semantic similarity to gauge transparency; however, this approach only considers the similarity between embeddings of idioms, overlooking the intrinsic relationship between an idiom's semantics and the meanings of its components. Hence, we propose a method that integrates Overall Semantic Transparency (OST) and Constituent Semantic Transparency (CST) to evaluate the semantic transparency of each idiom comprehensively. Furthermore, in tasks involving idiom reading comprehension and matching where multiple idioms are included in the examples, consistency rating method is applied to each sample. That is, when there is a significant difference in the difficulty levels among multiple idioms within a single sample, it becomes harder to understand, requiring the LLM to possess stronger metaphor competence.

(2) Enhancing Metaphor Competence: Paraphrase Augmentation. In the process of human learning of idioms, definition information aids in comprehending semantics thoroughly and mastering usage. Given that LLMs inherently possess extensive knowledge repositories, this work utilizes prompt techniques to guide models in fully exploring their internal information and generating definitions for idioms. This explicit definition information helps the model understand the process of mapping idioms from literal meanings to metaphorical meanings, thereby enhancing the model's ability to comprehend idiomatic metaphors. Additionally, we introduce a self-check mechanism where the model reflects on the candidate answers it generates, further bolstering its autonomous judgment capabilities.

The research idea of our work is shown in Figure 1, and the main contributions are as follows:

- We propose a **Consistency Rating of Semantic Transparency (CR-ST)** to measure the difficulty degree of understanding idioms, and evaluate the metaphor competence of LLMs.
- Based on the prompt technology of LLMs, we construct a **Paraphrase Augmentation Strategy with Self-checking (PASS)** to improve the metaphor competence of LLMs.
- On CINLID and ChID datasets, we evaluate the idiom understanding results of seven LLMs to prove the rationality of CR-ST and PASS.

2 Related Work

Metaphor and semantic transparency are crucial concepts in cognitive linguistics that have significant implications for artificial intelligence, particularly in natural language understanding and generation. According to Lakoff and Johnson (1980), metaphors are not merely linguistic expressions but fundamental to human cognition, enabling the mapping of literal meanings to figurative meanings. Semantic transparency refers to the extent to which the meaning of an idiom or phrase can be inferred from its constituent parts. High semantic transparency indicates that the meaning is easily deduced from the component words, while low transparency means the idiom's meaning is less apparent.

The current mainstream methods for calculating semantic transparency include both manual evaluation and evaluations based on PLMs (Pre-trained Language Models). Manual evaluation involves human annotators rating the transparency of idioms, which can provide high-quality, nuanced assessments but is time-consuming and subjective. On the other hand, PLM-based evaluations leverage models like BERT and GPT-3 to compute semantic similarity between idioms and their component words (Liu et al., 2019a). These automated methods are scalable and consistent but may struggle with capturing the deeper, more nuanced aspects of

idiomatic meaning that humans can intuitively understand (Shwartz et al., 2019). Recent advances in the field have sought to combine these approaches to leverage the strengths of both manual and automated evaluations (Schuster et al., 2020).

In the context of Chinese language understanding, recent progress in large language models (LLMs) such as ChatGLM and LLAMA3 has been notable. These models have shown significant improvements in handling idiomatic and metaphorical language, yet challenges remain. For instance, while models like ChatGLM3-6B can outperform traditional PLMs in certain tasks, their performance can vary significantly across different idioms, particularly those with low semantic transparency. Additionally, the inclusion of metaphorical meanings and semantic transparency in LLMs remains a complex issue, as these models often introduce noise and struggle with idioms' nuanced interpretations (Zhang et al., 2023; Kuhn and Farquhar, 2023; Elazar et al., 2021). Further research is needed to enhance LLMs' capabilities in this area, potentially through the integration of cognitive linguistic theories and the development of specialized evaluation metrics.

3 Consistency Rating of Semantic Transparency

We propose a semantic transparency algorithm consistent with cognitive linguistics, and introduce a consistency rating to calculate the semantic transparency of samples containing multiple idioms.

3.1 Semantic Transparency

The semantic transparency is often used to measure the difficulty of understanding idioms. In our work, the semantic transparency of an idiom consists of two parts: the Overall Semantic Transparency (OST) and the Constituent Semantic Transparency (CST). OST is the extent to which the figurative meaning of the idiom is similar to its literal meaning. CST is the extent to which the constituent retains its meaning in the figurative meaning of the idiom.

For OST, we use semantic vectors in distributional semantics to represent the figurative meanings of idioms. This method captures the figurative meaning by considering the context in which the idiom is used. To obtain the literal meaning, we calculate the mean of the word semantic vectors, treating the literal meaning as a simple combination of these vectors and disregarding the usual contextual nuances. We then measure the similarity between the figurative and literal meanings using cosine similarity. Refer to the **Function** `OST` in lines 1-4 of Algorithm 1 for implementation.

For CST, since the idioms are usually symmetrical structures, we split the idiom into two compositions of the same length, and encoding them by semantic vectors. To obtain the retention degree of compositions semantic in the figurative semantic, we use cosine similarity to calculate, and take the mean of the semantic transparency of the two components as the final CST. The above ideas are formalized as the **Function** `CST` in lines 5-12 of Algorithm 1.

3.2 Consistency Rating

Semantic transparency is measured in terms of individual idioms. However, in NLP tasks, since there may be multiple idioms in a sample, we also consider the calculation of sample idiom transparency. One cognitive assumption here is that idioms with similar levels of transparency are easier to understand together, which we refer to as "transparency consistency". For example, the idioms "高高兴兴[1] (gāo gāo xìng xìng)" and "心想事成[2] (xīn xiǎng shì chéng)" form a pair with high consistency, while the pair "高高兴兴 (gāo gāo xìng xìng)" and "阳春白雪 [3](yáng chūn bái xuě)" has lower consistency because the former has higher semantic transparency, whereas the latter has lower semantic transparency.

We propose a **C**onsistency **R**ating of **S**emantic **T**ransparency (**CR-ST**), using variance to measure the difficulty of understanding idioms in a sample (line 20 in Algorithm 1). The smaller the variance, the higher consistency of the idioms transparency in the sample, and the easier the model to understand them. The greater the variance, the less consistent the idioms transparency in the sample, and the model understand them more difficult. Taking ChID dataset as an example, the data distribution after CR-ST is shown in Figure 2.

Next, to evaluate the impact of CR-ST on LLMs performance, we the dataset into four subsets, T_1, T_2, T_3 and T_4, whose CR-ST is from high to low, indicating that idiom comprehension difficulty ranges

[1]"高高兴兴" describes a state of being extremely joyful and cheerful.

[2]"心想事成" expresses the hope or blessing that whatever one desires or aims for will come to fruition.

[3]"阳春白雪" refers to highbrow or refined artistic works that are appreciated by only a few.

Algorithm 1: The algorithm of semantic transparency rating.

Input: The Pre-train encoder $\texttt{Enc}(\cdot)$; The mean function $\texttt{Mean}(\cdot)$; The variance function $\texttt{Var}(\cdot)$;The cosine similarity function $\texttt{Cos}(\cdot)$; The normalization operation $\texttt{Nor}(\cdot)$.

Output: The sample semantic transparency S_{st}.

Data: The the sample sequence $\mathcal{D} = \{S_1, S_2, \cdots, S_M\}$, and each sample contains a idioms sequence: $\mathcal{S}_i = \{idiom_1, idiom_2, \cdots, idiom_N\}$.

Function $\texttt{OST}($ *idiom: str, fm: list* $\rightarrow O_{st}$*: int*$)$**:**
  $idiom = [w_1, w_2, \cdots, w_K]$
  $lm = \texttt{Mean}(\texttt{Enc}(w_1), \texttt{Enc}(w_2), \cdots, \texttt{Enc}(w_K))$ // Encode the literal meaning of $idiom$
  return $\texttt{Cos}(fm, lm)$ // Measure the similarity between figurative and literal meanings
Function $\texttt{CST}($ *idiom: str, fm: list* $\rightarrow C_{st}$*: int*$)$**:**
  $N = len(idiom)$
  $C_1 = [w_1, w_2, \cdots, w_{[N/2]}]$ // Divide $idiom_n$ into two constituent C_1 and C_2
  $C_2 = [w_{[N/2+1]}, w_{[N/2+2]}, \cdots, w_N]$
  for $k \in [1, 2]$ **do**
    $cm_k = \texttt{Enc}(C_k)$
    $C_{(k,st)} = \texttt{Cos}(cm_k, fm)$ // Measure the retention degree of the constituent C_i
  return $\texttt{Mean}(C_{(1,st)}, C_{(2,st)})$
Function $\texttt{Main}$**:**
  for $\mathcal{S} \in \mathcal{D}$ **do**
    Initializing a list I_{st} // store the semantic transparency of the candidate idioms
    for $idiom \in \mathcal{S}$ **do**
      $fm = \texttt{Enc}(idiom)$ // Encode the figurative meaning of $idiom$
      $O_{idiom} = \texttt{OST}(idiom, fm)$ // Calculate the overall semantic transparency
      $C_{idiom} = \texttt{CST}(idiom, fm)$ // Calculate the constituent semantic transparency
    $S_{st} = \texttt{Mean}(\texttt{Var}(\boldsymbol{O}_{idiom}), \texttt{Var}(\boldsymbol{C}_{idiom}))$ // Obtain the semantic transparency of sample $\mathcal{S}$
  return $S_{st} = \texttt{Nor}(\boldsymbol{S}_{st})$

from easy to hard. As shown in Figure 2, due to the long-tail characteristics of data distribution, using equal quantiles as thresholds results in too few samples in the T_4 subset. On the other hand, using quantiles as thresholds leads to minimal differences in transparency values between different subsets. Therefore, we use the three steps to split subsets:

- Calculating the quartiles Q_1, Q_2, and Q_3 of the data distribution, which guarantees a balanced sample size for each subset.
- Calculating the four equal points of the semantic transparency values E_1, E_2 and E_3, which ensures that the otherness of each subset are balanced.
- Taking the mean of quantile Q_i and equal points E_i as the partition threshold to balance the tradeoff between otherness and size.

Finally, we obtain the four transparency subsets

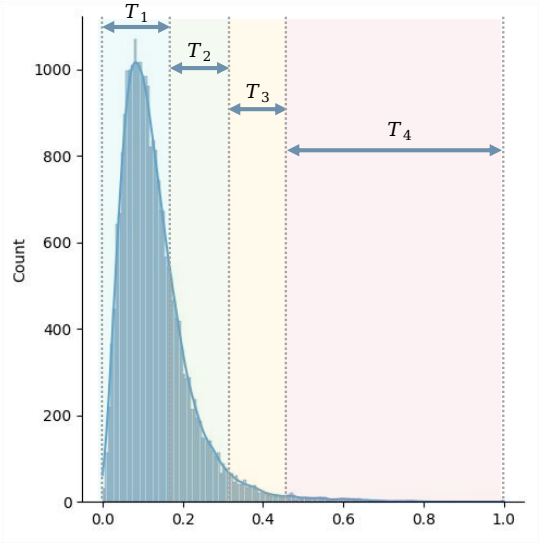


Figure 2: Sample distribution of ChID dataset by consistency rating of semantic transparency.

Dataset	Attrs	T_1	T_2	T_3	T_4
Chid	Threshold	≤ 0.160	≤ 0.305	≤ 0.460	> 0.460
	Size	18,149	5,691	808	300
	Mean	0.090	0.209	0.359	0.581
CINLID	Threshold	≤ 0.145	≤ 0.395	≤ 0.460	> 0.460
	Size	18,566	6,389	1,382	369
	Mean	0.064	0.204	0.356	0.554

Table 1: Attributes of the dataset by consistency rating of semantic transparency, including threshold, sample size and sample mean.

whose properties are shown in Table 1. In Section 5.4, we evaluate the results of the current mainstream LLMs on transparency subsets and proved the effectiveness of the CR-ST for idiom understanding.

4 Paraphrase Augmentation Strategy with Self-checking

In order to further improve the metaphorical ability of the large model to understand idioms, we construct **P**araphrase **A**ugmentation **S**trategy with **S**elf-checking (PASS) based on prompt technology to guide the LLM to fully explore its mastery of idiom paraphrase, and analyze the effect of paraphrase augmentation on different semantic transparent idioms. The overall template as shown in Prompt 1, which is divided into three steps. Please refer to Appendix A Prompt 2 for CINLID.

4.1 Paraphrase Augmentation Strategy

Metaphor competence refers to the model's ability to map from literal meanings to metaphorical meanings, which fundamentally relies on the paraphrase or usage of idioms (Petroni et al., 2019). For instance, when we understand the idiom "阳春白雪 (yáng chūn bái xuě)", we can quickly grasp it by reading its paraphrase and historical anecdote. Although LLMs possess extensive knowledge, they still lack logical reasoning abilities (Roberts et al., 2020). Therefore, we propose Paraphrase Augmentation Strategy, which explicitly guides LLMs to generate paraphrase for the current idiom based on their knowledge and then provide the correct answer in idiom comprehension tasks. The following idiom cloze as an example to explain in detail.

Specifically, as Step1 in Prompt 1 shows, for each of the candidate idioms $\mathcal{S}$ in the sample, we instruct the model to generate a corresponding paraphrase for each idiom, a procedure that explicitly guides the LLM through the idiom understanding [4]. Prompt strategies for generating paraphrase can help us to: 1) intuitively evaluate the model's understanding ability for each idiom and 2) evaluate the difference in the LLM metaphor competence of paraphrase augmentation under different semantic transparency combined with the CR-ST.

Prompt 1

Step 1: Paraphrase Augmentation
Prompt: You are an expert in Chinese idioms, please generate paraphrase for the following idioms according to your knowledge.
Input: candidate idioms $\mathcal{S}$
Output: idioms paraphrase $\mathcal{P}$

Step 2: Answer Selection
Prompt: This is a cloze task where you need to understand the context and the candidate idioms, and choose the one from the candidate idioms according to the idiom paraphrase that best fits into the placeholder #$idiom$# in the context.
Input: context, candidate idioms $\mathcal{S}$, idioms paraphrase $\mathcal{P}$
Output: predictive idiom $idiom_i$

Step 3: Self-checking
Prompt: This is a cloze task where you have to determine whether the predictive idiom is correct. If it is correct, the result is printed directly. If it is not correct, you need to re-select the most suitable idiom from the candidate idioms.
Input: context, candidate idioms, idioms paraphrase $\mathcal{P}$, predictive idiom $idiom_i$
Output: predictive idiom $idiom_i$ / revised idiom $idiom_j$

4.2 Answer Selection based on Self-checking

In the Step 2 of Prompt 1, we first declare a task description that requires the LLM to understand the context and the corresponding placeholders. Secondly, the paraphrases generated by the LLM will be used as input for the Step 2, helping the model select the most appropriate idiom to fill in the context of the text content. Finally, the model predicts the idiom as the answer.

[4] In generating paraphrase, we try to avoid offensive or discriminatory words

	Training	CINLID		ChID				
		Dev	Test	Dev	Test	Ran	Sim	Out
BLSTM (Zheng et al., 2019)	✔	61.9	62.1	71.8	71.5	80.7	65.6	61.5
SAR (Zheng et al., 2019)	✔	62.1	62.0	71.8	71.5	80.0	64.9	61.7
AR (Zheng et al., 2019)	✔	64.3	65.2	72.7	72.4	82.0	66.2	62.9
Bert (Devlin et al., 2019)	✔	65.0	65.4	68.6	69.3	80.2	62.2	61.6
Roberta (Liu et al., 2019b)	✔	63.0	62.5	72.8	**73.3**	**84.2**	**66.3**	**65.7**
Llama2-7B (Hugo Touvron and et al., 2023)	✘	-	43.01	-	29.89	39.65	32.11	36.83
Chatglm2-6B (Du et al., 2022)	✘	-	48.87	-	30.07	42.62	31.06	40.39
Baichuan2-7B (Yang et al., 2023)	✘	-	54.76	-	35.62	45.90	41.86	42.45
Qwen1.5-7B (Bai et al., 2023)	✘	-	59.57	-	52.37	65.50	48.33	54.42
Internlm-7B (Cai et al., 2024)	✘	-	62.24	-	55.02	59.60	32.29	41.08
Llama3-8B (AI@Meta, 2024)	✘	-	72.83	-	41.92	52.08	31.92	38.82
Chatglm3-6B (Zeng et al., 2023)	✘	-	73.02	-	72.58	82.85	47.53	56.25
GPT-4o (OpenAI, 2024)	✘	-	**77.86**	-	66.25	72.47	32.45	49.87
Human (Zheng et al., 2019)	✘	-	89.14	-	87.1	97.6	82.2	86.2

Table 2: Evaluation results on CINLID and ChID datasets. The experimental results of the gray are from ChID (Zheng et al., 2019), and the other experimental results are from our works. The best experiment is the **bold** term, and the next best experiment is the underline term.

In order to improve the accuracy, we introduced a Self-checking strategy to guide the LLM to reflect on whether the predicted idioms were correct in the Step 3. In this process, the prediction idiom and sample information are used together as inputs, and the LLM will eventually output what it thinks is the final answer.

5 Experimental Setup and Result

5.1 Dataset

ChID is a large-scale Chinese idiom dataset for cloze testing (Zheng et al., 2019), which contains 581K paragraphs and 729K blanks. In ChID, idioms in paragraphs are replaced with blank symbols (#*idiom*#). For each blank, provide a list of 7 candidate idioms, including golden idioms, as a choice. In addition to having a common **Test** set, the ChID is also designed **Out** set for out-of-domain test to assess the generalization ability of models. **Ran** and **Sim** set have the same paragraph as Test, but the design of the candidate idioms is different. In Ran, all candidates are drawn from idioms that do not resemble the golden idiom. Instead, in Sim, all candidates are drawn from the 10 most similar idioms.

CINLID (Chinese Idioms Natural Language Inference Dataset) comes from Baidu LUGE database[5], which collected 106,832 idioms (training set 80,124, test set 26,708), including a few short texts such as proverb and allegory. In this paper, we split the original training set into training set 64,099 and validation set 16,024. Based on the four basic semantic categories of same relation, including relation, overlapping relation and separation relation, the idiom pairs are artificially labeled as entailment, contradiction and neutral. **Entailment** indicates a similar meaning ("拾陈蹈故 (shí chén dǎo gù)" and "因循守旧 (yīn xún shǒu jiù)"[6]), **Neutral** means semantically neutral ("沉滓泛起 (chén zǐ fàn qǐ)"[7] and "凤泊鸾飘 (fèng bó luán piāo)"[8]), and **Contradiction** means two words with opposite meanings ("稀奇古怪 (xī qí gǔ guài)"[9]and"平淡无奇 (píng dàn wú qí)"[10]).

5.2 Baseline and Setting

The model we evaluated consisted of three classical models: **BLSTM** (Hochreiter and Schmidhuber, 1997), **AR** and **SAR** (Zheng et al., 2019) , two PLMs **Bert** (Devlin et al., 2019) and **Roberta** (Liu et al., 2019b), and seven LLMs: **Llama2-7B** (Hugo Touvron and et al., 2023), **Chatglm2-6B** (Du et al., 2022), **Baichuan2-7B** (Yang et al., 2023), **Qwen1.5-7B** (Bai et al., 2023), **Internlm-7B** (Cai et al., 2024), **Llama3-8B** (AI@Meta, 2024), **Chatglm3-6B** (Zeng et al., 2023) and **GPT-4o** (OpenAI, 2024).

In this work, all experimental metrics were **ACC**,

[5]https://www.luge.ai/#/luge/dataDetail?id=39

[6]"拾陈蹈故" and "因循守旧": Stick to the old ways, lack of innovation.

[7]"沉滓泛起": Things that have disappeared reappear.

[8]"凤泊鸾飘": Talented people don't succeed.

[9]"稀奇古怪": Out of the ordinary.

[10]"平淡无奇": Plain and ordinary.

	PASS	CINLID					ChID				
		Test	T_1	T_2	T_3	T_4	Test	T_1	T_2	T_3	T_4
Llama2-7B	✘	17.90	18.92	17.90	16.07	15.89	35.47	46.07	38.13	38.08	34.15
	✔	43.01	43.27	49.96	41.34	32.77	20.89	27.29	27.80	26.49	31.00
Chatglm2-6B	✘	27.64	27.51	28.00	28.59	26.33	35.57	34.10	38.39	36.32	31.71
	✔	48.87	48.64	49.99	47.92	44.69	30.07	31.27	30.79	33.04	27.00
Baichuan2-7B	✘	35.81	35.88	36.00	33.42	34.67	38.48	38.48	30.46	30.90	41.67
	✔	54.76	56.07	51.71	50.96	55.80	35.62	36.89	35.42	34.97	32.14
Qwen1.5-7B	✘	48.63	48.47	48.34	48.27	46.00	51.51	48.47	45.44	40.01	40.11
	✔	59.57	61.20	57.13	52.76	45.50	52.37	52.37	52.87	49.38	50.67
Internlm-7B	✘	42.20	42.34	42.21	39.98	39.33	54.65	57.52	49.52	43.63	40.11
	✔	62.24	64.70	57.39	53.34	53.09	55.02	55.40	52.21	51.58	48.67
Llama3-8B	✘	43.08	43.31	42.82	42.33	36.33	58.05	60.88	52.59	46.74	52.57
	✔	72.83	74.40	70.59	65.56	59.62	41.92	42.22	41.15	41.96	38.33
Chatglm3-6B	✘	53.87	53.87	53.05	52.72	51.00	61.81	64.40	57.38	50.51	50.14
	✔	73.02	74.58	69.79	67.15	**66.09**	**72.58**	62.90	62.87	62.02	62.33
GPT-4o	✘	74.52	75.33	75.29	65.80	51.76	66.13	66.49	65.99	65.47	65.46
	✔	**77.86**	**79.85**	**78.06**	**69.73**	57.30	66.25	**66.83**	**66.02**	**65.33**	**65.27**

Table 3: Evaluation results of CINLID and ChID datasets divided based on consistency rating of semantic transparency. The best experiment is the **bold** term. The gray of experimental results indicates results that violate expectations.

and the evaluation was performed on a single NVIDIA A40 GPU. In order to ensure the fairness of the experiment, we set the temperature of the LLM to 0. Please refer to Appendix B for the selection of temperature, and Refer to Appendix A for the prompt template used in our work. In addition to the seven LLMs, we adopted the training method in the original paper for other models. All experiments had 5 epochals, $Adam\ optimizer$ was used, warmup_ratio of 0.1, and learning rate of 5e-5. In the CINLID dataset, max_length is 32 and batch_size is 64. In the ChID dataset, max_length is 128 and batch_size is 32.

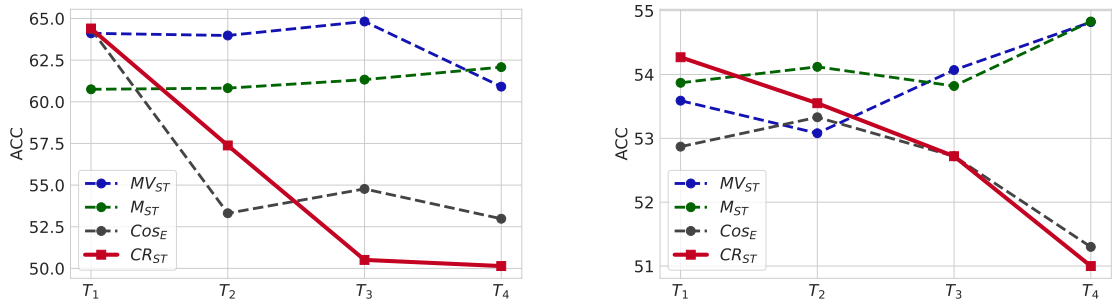
5.3 Main result

This section analyzes the basic evaluation results of LLMs, classical models, and PLMs in idiom understanding, as shown in Table 2. On CINLID dataset, GPT-4o and Chatglm3-6B perform well, on ChID dataset, Roberta performs best and Chatglm3-6B is second.

In LLMs, **Chatglm3-6B** demonstrates strong performance, achieving results that either surpass or approach PLMs. However, other LLMs show significant disparities in performance. Post-training, Bert achieves an accuracy of 65.4% on the Test of CINLID. Apart from **GPT-4o**, **Chatglm3-6B** and **Llama3-8B**, the performance of other LLMs falls below that of Bert. On ChID, aside from **Chatglm3-6B**, the performance of other models shows substantial gaps compared to the optimal performance of **Roberta**. For instance, Llama3-8B performs 31.38% lower than **Roberta** on the Test. This highlights considerable variation in the comprehension of idioms among LLMs, reflecting differences in performance across various stages of LLM development.

For different tasks, idioms cloze testing poses greater difficulty, whereas idiom matching tasks are less challenging, with LLMs performing better. For instance, **Llama3-8B** achieves 72.83% on the Test of CINLID, whereas it scores 41.92% on ChID Test, indicating a significant disparity. In contrast, PLMs demonstrate comparable comprehension abilities across CINLID and ChID. On the other hand, overly similar candidate idioms in cloze tests also present a challenge. For example, in the ChID, **Chatglm3-6B** experiences substantial performance drops on Sim and Out dataset, declining by 18.77% and 9.45% respectively compared to **Roberta**.

There remains a gap between the ability of LLMs and PLMs to understand idioms compared to humans. Drawing on human evaluation methods from ChID, we obtained benchmark results on the CINLID dataset, where **Human** accuracy rates were observed to be 89.14% and 87.1% on ChID and CINLID respectively, surpassing those of Chatglm3-6B



(a) CINLID (b) ChID

Figure 3: Comparative experiment of different semantic transparency methods.

Strategy	Para.	Self-check	Exam.	CINLID					ChID				
				Test	T_1	T_2	T_3	T_4	Test	T_1	T_2	T_3	T_4
PASS	✔	✔	✘	**73.02**	**74.58**	**69.79**	**67.15**	**66.09**	**62.38**	**62.90**	**62.87**	**62.02**	62.33
Var_1	✔	✘	✘	58.74	60.70	58.33	58.30	57.96	62.21	62.31	60.76	59.45	**62.47**
Var_2	✘	✔	✘	64.13	64.43	63.10	63.98	61.54	58.36	58.40	58.32	58.10	57.69
Var_3	✘	✔	✔	54.89	55.10	54.03	54.88	52.74	58.04	58.33	58.01	57.05	57.11

Table 4: Ablation study over PASS, which contains three variants Var_1, Var_2 and Var_3.

by 15.74% and 14.52%. Refer to the Appendix C for details of human evaluations.

5.4 Evaluation results of semantic transparency subsets

We present the experimental results of LLMs on transparency subsets T_1-T_4, as shown in Table 3. The top half of each column represents the results without the PASS module, and the bottom half represents the results with the PASS module. Note that the prompt template without PASS contains only the first step in the Prompt 1 and Prompt 2.

First, we analyze the results of the no PASS version. The result trend of most LLM on T_1-T_4 is declining, which indicates that CR-ST effectively evaluated the difficulty of idiom understanding, i.e., the higher the CR-ST, the lower the consistency of idiom cognition and the greater the difficulty of understanding. we observed abnormal experimental results, shown in gray, which deviated from expectations. Specifically, **Chatglm2-6B** and **Baichuan2-7b** showed no obvious trend in the experimental results of T_1-T_4. Moreover, **Qwen1.5-7B** and **Llama3-8B** exhibited higher performance on T_4 than on T_3 in ChID, which could be attributed to the the long-tail distribution of the dataset.

On the other hand, we focus on the performance improvement with the addition of PASS across transparency subsets. On the CINLID, PASS significantly improved evaluation results. For example, after adding PASS, **Chatglm3-6B** showed improvements of 19.15%, 20.71%, 16.74%, 14.43%, and 15.09% on the Test and T_1-T_4 datasets, respectively. In contrast, on the ChID dataset, the effect of PASS was less pronounced and even counterproductive in some cases. For instance, while PASS improved performance for **Chatglm3-6B**, **Internlm-7B**, and **Qwen1.5-7B**, it led to performance declines for other LLMs. This may be because the addition of paraphrase introduced too much noise, increasing the difficulty for LLMs to understand the candidate idioms, thereby reducing performance. In ChID, we observed that the addition of PASS reduced the performance gap between T_1-T_4, indicating that the inclusion of paraphrase has a more significant impact on idioms with lower transparency.

5.5 Comparative experiment

In this section, we show the rationality of the CR-ST through experiments, so we choose three methods to conduct comparative experiments: MV_{ST} indicates that the semantic transparency of the sample adopts the $(Mean + Variance)$ method, M_{ST} indicates that the semantic transparency of the sample is measured by the $Mean$, and Cos_E is the semantic similarity of the traditional cosine similarity calculation idiom, and CR_{ST} is our method.

From the trend in Figure 3, CR_{ST} shows a downward trend on both datasets, which is consistent with our theory that idioms with lower semantic transparency are more difficult to understand and less accurate. Specific analysis, in the two data sets, the other three methods show different degrees of fluctuation. For example, Figure 3 (a), the M_{ST} method on CINLID shows an upward trend, which is completely contrary to our perception. Another case where T_4 accuracy is too high, it appears in the M_{ST} and MV_{ST} methods on the ChID dataset (Figure 3 (b)), which may be due to the long tail of the data distribution, resulting in too few samples. Therefore, we can argue that compared with traditional other methods, CR-ST to measure semantic transparency consistency is more in line with the model's understanding process of idioms.

5.6 Prompt template study

In this section, we propose three variations of the Paraphrase Augmentation Strategy with Self-checking (PASS), to investigate the necessity for different modules in PASS. The base model is **Chatglm3-6B**. Var_1 indicates the deletion of the Self-check from the PASS, and Var_2 indicates the deletion of Paragraph augmentation from the PASS. Notably, we propose Var_3, which uses example augmentation to replace paraphrase augmentation (Appendix A Prompt 3), which stems from the idea in cognitive linguistics that idiom paraphrase and usage help humans better understand idioms .

The results of the ablation study are shown in Table 4. We can see that in the CINLID dataset, the performance of PASS and the three variants improved compared with Table 2. Among them, the improvement effect of PASS is obvious, the ACC improvement of Test is 18.13% over Var_3, and the sub dataset is also significantly improved. The results show that the PASS is effective for the understanding of idiom pairs. On the ChID dataset, compared with the results in Table 2, PASS is improved by 4.34%, Var_1 is also improved. However, the performance of Var_2 and Var_3 is decreased, which indicates that the paraphrase and example augmentation are not enough due to the existence of multiple options in the idiom cloze task, and it may be necessary to further guide LLM to learn the internal logical thinking of idiom understanding.

6 Conclusion

This work evaluates the metaphor competence of LLM on the idioms understanding task. We propose CR-ST strategy, which evaluates the performance of LLMs at different rating levels representing their varying degrees of metaphorical understanding. Furthermore, we designs a prompt template PASS, that combines paraphrase augmentation and self-check mechanisms at different rating levels, assisting LLMs in enhancing metaphor competence when performing idiom understanding tasks. We comprehensively evaluate the performance of LLM on ChID and CINLID datasets, and prove the validity of CR-ST and PASS methods.

Acknowledgements

This work is supported in part by the Natural Science Foundation of China (grant No.62276188), TJU-Wenge joint laboratory funding.

Limitations

Firstly, the types of idiom comprehension tasks are relatively limited, particularly lacking in the area of text generation. This indicates a significant gap between understanding and applying idioms, suggesting that the next step in idiom comprehension should move towards generation tasks. Secondly, carefully designed evaluation metrics for idiom comprehension are crucial for understanding and applying idioms, differing from traditional semantic similarity metrics. Without such evaluations, it is impossible to further improve idiom understanding and usage in generative contexts. Lastly, enhancing LLMs' ability to comprehend idioms remains challenging. Exploring research feasibility in areas such as Agents and Retrieval-Augmented Generation (RAG) could be beneficial.

Ethics Statement

The research conducted in this paper adheres to the ethical principles of ACL. The process of generating interpretations and context of idioms using large language models may involve the presence of potentially offensive or discriminatory words due to the emotional polarity of idioms. Despite our extensive efforts to mitigate their impact, it may require more meticulous manual inspection. We have provided warnings to readers in the relevant sections of the paper.

References

AI@Meta. 2024. Llama 3 model card.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and Binyuan Hui et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, and Xin Chen et al. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Yanai Elazar, Shauli Ravfogel, and Yoav Goldberg. 2021. Measuring and improving consistency in pre-trained language models. *arXiv preprint arXiv:2104.05810*.

Nguyen Thiphuong Wang Zhenliang Fang Yuanyuan, Xie Ruibo and Wu Xinchun. 2023. The processing mechanism and development stages of chinese idiom comprehension. *Chinese Journal of Applied Psychology*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Zhuanglin Hu. 2023. *Metaphor and cognition*, volume 8. Springer Nature.

Louis Martin Hugo Touvron and Kevin Stone et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jesus Insuasti, Felipe Roa, and Carlos Mario Zapata-Jaramillo. 2023. Computers' interpretations of knowledge representation using pre-conceptual schemas: An approach based on the bert and llama 2-chat models. *Big Data and Cognitive Computing*, 7(4):182.

Nina Julich-Warpakowski and Paula Pérez Sobrino. 2023. Introduction: Current challenges in metaphor research.

Gal Kuhn and Sebastian Farquhar. 2023. Semantic consistency for assuring reliability of large language models. *arXiv preprint arXiv:2308.09138*.

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

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

Qiang Liu, Peng Li, and Shuming Shi. 2019a. Neural network models for idiom identification in natural language processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3510–3520.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. Gpt understands, too. *AI Open*.

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

OpenAI. 2024. Gpt-4 technical report. Technical report.

Ollomurodov Arjunbek Orifjonovich. 2023. The main features of conceptual metaphors in modern linguistics. *American Journal of Language, Literacy and Learning in STEM Education (2993-2769)*, 1(9):365–371.

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

Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, and Shilei Wen. 2024. Diffusiongpt: Llm-driven text-to-image generation system. *arXiv preprint arXiv:2401.10061*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Tal Schuster, Tal Linzen, Leon Bergen, and Massimo Poesio. 2020. Towards a better understanding of figurative language in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1044–1053.

Xu Shenghuan. 2022. Quantum thinking and language study—an exploration of analogical metaphor from the perspecfive of the non-locality principle. *Foreign Language Teaching and Research (bimonthly)*, 54(2):189–200.

Yucheng Shi, Hehuan Ma, Wenliang Zhong, Qiaoyu Tan, Gengchen Mai, Xiang Li, Tianming Liu, and Junzhou Huang. 2023. Chatgraph: Interpretable text

classification by converting chatgpt knowledge to graphs. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 515–520. IEEE.

Vered Shwartz, Evangelos Antonakos, Shoichi Sekine, Mark Johnson, and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on idiom handling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3374–3380.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, pages 348–367. Springer.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, arXiv:2309.10305.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Glm-130b: An open bilingual pre-trained model. *Preprint*, arXiv:2210.02414.

Shufan Zhang, Minghui Li, and Kai Sun. 2023. Can large language models understand uncommon meanings of common words? *arXiv preprint arXiv:2305.05741*.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. Chid: A large-scale chinese idiom dataset for cloze test. *arXiv preprint arXiv:1906.01265*.

A Prompt Template

Prompt 2 is the PASS prompt template for CINLID, which is applied in Section 5.3 and 5.4. Prompt 3 is a prompt template for Var_3 models in Section 5.6.

Prompt 2

Step 1: Paraphrase Augmentation
Prompt: You are an expert in Chinese idioms, please generate paraphrase for the following idioms according to your knowledge.
Input: idiom pair ($idioms_1$, $idioms_2$)
Output: idioms paraphrase (p_1, p_2)

Step 2: Answer Selection
Prompt: This is an idiom understanding task. For a given pair of idioms, by learning their paraphrase, output them as Entailment, Contradiction or Neutral without any other description.
Input: idiom pair ($idioms_1$, $idioms_2$), idioms paraphrase (p_1, p_2)
Output: predictive semantic relation

Step 3: Self-checking
Prompt: Please determine whether the relation of the given idiom is correct based on the paraphrase and your understanding. If yes, answer correctly, if not, give a new judgment, output your judgment from Entailment, Contradiction or Neutral.
Input: idiom pair ($idioms_1$, $idioms_2$), idioms paraphrase (p_1, p_2), predictive semantic relation
Output: predictive semantic relation

B Parameter Analysis

Temperature is a key parameter affecting LLM performance. We selected ChatGLM3 and Qwen1.5 to analyze their evaluation performance and stability at different temperatures, and the temperature range is $tem = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$.

As shown in Table 5, when $tem = 0.0$, i.e., $do\_sample = False$, the performance on both datasets is the best and the variance is small. Although ChatGLM3 performs better on the ChID dataset when $tem = 0.2$, it also has a larger variance, so in our work we set all tem to 1. The above experimental results show that for the idiom understanding task of text matching, it is more suitable to set the LLM temperature at a lower level to reduce the flexibility of text generation.

Dataset	Model	0	0.2	0.4	0.6	0.8	1.0
CINLID	ChatGLM3	**73.02±0.00**	72.98±0.04	72.86±0.26	72.98±0.41	73.00±0.38	71.02±0.48
	Qwen1.5	**59.57±0.00**	52.37±0.28	52.87±0.28	49.38±0.28	50.67±0.28	50.67±0.00
ChID	ChatGLM3	62.38±**0.00**	**62.40**±0.28	61.06±0.44	59.18±0.86	56.78±1.98	54.03±2.34
	Qwen1.5	**52.37±0.00**	**52.37**±0.28	52.87±0.28	49.38±0.28	50.67±0.28	50.67±0.28

Table 5: Evaluation results of ChatGLM3 and Qwen1.5 at different temperatures.

Human	ACC	κ
User_1	91.11	-
User_2	89.25	-
User_3	87.06	-
Average	89.14	0.763

Table 6: Caption

Prompt 3

Step 1: Example Augmentation
Prompt: You are an expert in Chinese idioms, please make sentences for the following idioms according to your understanding.
Input: candidate idioms $\mathcal{S}$
Output: idioms example $\mathcal{E}$

Step 2: Answer Selection
Prompt: This is a cloze task where you need to understand the context and the candidate idioms, and choose the one from the candidate idioms according to the idiom example that best fits into the placeholder #$idiom$# in the context.
Input: context, candidate idioms $\mathcal{S}$, idioms example $\mathcal{E}$
Output: predictive idiom $idiom_i$

Step 3: Self-checking
Prompt: This is a cloze task where you have to determine whether the predictive idiom is correct. If it is correct, the result is printed directly. If it is not correct, you need to re-select the most suitable idiom from the candidate idioms.
Input: context, candidate idioms, idioms example $\mathcal{E}$, predictive idiom $idiom_i$
Output: predictive idiom $idiom_i$ / revised idiom $idiom_j$

C Human Evaluation

To explore the ceiling of model performance, we also conducted Human Evaluation. We sampled 200 samples from the test sets, and then hired three annotators to complete the tests. These three annotators are university students and all have very good command of Chinese language. The average accuracy of the annotators and the corresponding Fleiss' kappa are reported as the final performance. We report the experimental results in Table 6.