ChengyuSTS: An Intrinsic Perspective on Mandarin Idiom Representation

Le Qiu[1][*], Emmanuele Chersoni[1], Aline Villavicencio[2,3,4]

[1]The Hong Kong Polytechnic University, Hong Kong SAR

[2]The University of Exeter, UK [3]The University of Sheffield, UK

[4]Federal University of Rio Grande do Norte, Brazil

Abstract

Chengyu, or four-character idioms, are ubiquitous in both spoken and written Chinese. Despite their importance, *chengyu* are often underexplored in NLP tasks, and existing evaluation frameworks remain limited in scope and depth. In this paper, we introduce an intrinsic evaluation task for Chinese idiomatic understanding: idiomatic semantic textual similarity (iSTS), which evaluates how well models can capture the semantic similarity of sentences containing idioms. To this purpose, we present a curated dataset: ChengyuSTS. Our experiments show that current pre-trained sentence Transformer models generally fail to capture the idiomaticity of *chengyu* in a zero-shot setting. We then show results of fine-tuned models using the SimCSE contrastive learning framework, which demonstrate promising results for handling idiomatic expressions. We also presented the results of DeepSeek for reference [1]

1 Introduction

Among multi-word expressions (MWEs), idiomatic expressions (IEs) are distinctive as they are often non-compositional, suggesting their meanings may not be straightforwardly derived from individual words (Baldwin and Kim, 2010; Zeng and Bhat, 2021). For example, the phrase *spill the beans* can be interpreted either literally as *tipping over the pulse*, or figuratively as *revealing a secret*. The mix of compositionality and non-compositionality of IEs has imposed significant challenges on various natural language processing (NLP) tasks, including sentiment analysis, machine translation, and more. Proper representation of IEs hence holds significant impacts on linguistics and NLP research.

	画蛇添足
Literal:	to draw a snake and add feet to it
Figurative 1:	to ruin the effects by adding something superfluous
Figurative 2:	to fabricate something out of thin air

Table 1: A prototypical example of *chengyu*: 画蛇添足. It originates from an ancient fable: In a snake-drawing contest, a man finished first but tried to improve it by adding extra legs. He ended up ruining the drawing and lost to another contestant, who kept his snake simple and unadorned. The phrase 画 (*to draw*) 蛇 (*snake*) 添 (*to add*) 足 (*foot*) compositionally outlines the story, and two extra figurative meanings have been derived on the basis.

The present study brings its attention to IEs in the Chinese language or 成语 (*chengyu*), which literally mean *set phrases* and are often referred to as *Chinese idioms* or *four-character idioms*, as they're prototypically exhibited to be collocations of four characters (*chengyu* are not always made up with four characters, nor four-character combinations are necessarily *chengyu*). While Chinese idioms can include various forms of IEs, such as slangs (俚语, *liyu*), sayings (俗语, *suyu*), proverbs (谚语, *yanyu*), and more, the term *Chinese idiom* in this study refers exclusively to *chengyu*.

In contrast to idioms in Western languages, which usually resemble regular phrasal constructions (Cacciari, 2014), Chinese idioms frequently deviate from the syntactic norms of modern Mandarin. They are typically compact and synthetic in structure, and exhibit opacity in semantics (Tsou, 2012; Liu and Su, 2021), as shown in Table 1. This characteristic adds to their expressive richness but poses challenges for interpretation without adequate socio cultural knowledge.

Researchers have made efforts to improve IE representations. For instance, Zeng and Bhat (2023) have created context-aware representations for English PIEs, by unifying BART's (Lewis, 2019) ability to generate compositional meanings with an *idiomatic adapter* that captures their non-compositional meanings (Zeng and Bhat, 2022).

[*] Correspondence: lani.qiu@connect.polyu.hk

[1] Data and code are available upon https://github.com/Laniqiu/ChengyuSTS.

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 1–12
November 8-9, 2025 ©2025 Association for Computational Linguistics

For Chinese IEs, Wu et al. (2024) have proposed to boost idiomatic representations by mitigating the inconsistency between different semantic representations using multi-semantic contrastive learning. However, evaluations of Chinese idiomatic representations so far have been carried out in more limited settings, compared to English studies. Idiomaticity representation can be accessed extrinsically, through performance on downstream tasks such as machine translation, or intrinsically, by probing the linguistic properties encoded within the representations (Zeng and Bhat, 2023; He et al., 2024a). While studies on English IEs leverage both extrinsic and intrinsic methods (e.g. Zeng and Bhat, 2021, 2022, 2023), varying in task genres, work on Chinese idioms has largely depended on cloze-style tasks (e.g. Long et al., 2020; Wu et al., 2024). While such tasks directly assess language ability and are intrinsic in nature, they emphasize specific contextual use and are thus relatively less intrinsic than evaluations based on STS or embedding analysis, which focus more on semantic structure and are therefore reflect a deeper level of intrinsic evaluation. For simplicity, we refer to the former as *relatively extrinsic*, without implying the standard extrinsic evaluation based on downstream tasks.

The observed limitations arise from multiple aspects, including a lack of diverse evaluation methods and datasets. In light of this, the study takes the initiative to explore the intrinsic evaluation of idiomatic representation in Mandarin. Inspired by Tayyar Madabushi et al. (2022), we adopt the iSTS task (see section 3) as an intrinsic method and present an STS (semantic textual similarity) benchmark on Chinese idioms: **ChengyuSTS**. The dataset could provide a novel alternative for assessing Chinese idiom representations from an intrinsic perspective. Additionally, we investigated the performance of several Transformer models and DeepSeek (DeepSeek-AI, 2025) on the dataset. To our knowledge, this is the first evaluation of this new LLM (large language model) on Chinese idiomatic representations. Plus, we introduce several baseline models trained with contrastive learning, with the best ones yielding results comparable to DeepSeek.

The following sections are organized as follows: Section 2 introduces some related work and section 3 details the task setup and dataset creation. Section 4 presents the evaluation setup and results. Finally, section 5 concludes the current study and suggests directions for future work.

2 Related Work

2.1 MWE Representation

MWEs, such as IEs, have been a longstanding hardnut for NLP (Sag et al., 2002; Liu et al., 2017; Shwartz and Dagan, 2019; Biddle et al., 2020; Zeng and Bhat, 2022). Studies have suggested that PLMs (pre-trained language models) such as BERT (in its basic version) cannot model idiomaticity properly (Shwartz and Dagan, 2019; Wu et al., 2024). Even ChatGPT models (Brown, 2020), as powerful as they can be, have exhibited limitations in interpreting IEs (Raunak et al., 2023).

Early attempts to model MWEs, including IEs, generally employed either a *distributional* or a *compositional* approach. The distributional approach treats the entire phrase as a single, inseparable unit, learning its embedding from the context, much like the process for individual word embeddings (Mikolov, 2013; Yin and Schütze, 2014, 2016). While effective, this approach often requires large-scale corpora and thus is prone to data sparsity issues. On the other hand, the compositional approach constructs phrase embeddings by aggregating the embeddings of constituent words (Mitchell and Lapata, 2010; Yu and Dredze, 2015), but it struggles to represent non-compositional phrases such as IEs. Given these limitations, hybrid approaches have been developed to jointly learn phrase embeddings (Hashimoto and Tsuruoka, 2016; Li et al., 2018a,b).

Recent work has sought PLMs for IE representation using adaptive and contrastive learning techniques. Zeng and Bhat (2022) pioneered this direction by introducing GIEA, a contextualized embedding model that captures non-compositional semantics with an *idiomatic adapter* added to BART (Lewis, 2019). Building on this, they further developed PIER (Zeng and Bhat, 2023), a refined model that has been shown to effectively balance compositional and non-compositional representations of IEs. Additionally, He et al. (2024a) and Wu et al. (2024) integrated contrastive loss into their methods, respectively, enabling the models to better distinguish between different semantic interpretations. Studies have also found that incorporating external linguistic knowledge, such as hypernymy, synonyms or definitions can enhance model performance (Long et al., 2020; Wang et al., 2020; Sha et al., 2023).

2.2 Representation Evaluation

Word embeddings, by design, represent words as vectors, such that the proximity between vectors reflects semantic similarity relationships between the corresponding words (Schnabel et al., 2015; Bakarov, 2018). Their evaluation can be categorized into extrinsic and intrinsic methods.

Extrinsic methods evaluate representations based on their contribution to downstream tasks, such as machine translation, sentiment analysis or natural language inference, using the task-specific metrics as indicators of quality (Chiu et al., 2016; Zhou et al., 2024). However, the final performance can be influenced by several confounding factors, in addition to evaluating idiomatic accuracy. In contrast, intrinsic evaluation directly examines embeddings, often assessing their alignment with human judgments regarding *similarity* or *relatedness* between words (Schnabel et al., 2015; Chiu et al., 2016; Tsvetkov et al., 2016). By approximating these relevant tasks (e.g., to examine semantic similarity), intrinsic methods evaluate the general properties of word embeddings, without the need to perform each task of the sort (Tsvetkov et al., 2016) and indeed they represent a popular choice for analyzing the linguistic knowledge encoded in embeddings (Lenci et al., 2023; A et al., 2024; Ascari et al., 2024). Besides, they could provide insights into the traits that influence model performance in downstream tasks, and thus guide targeted improvements. As an example of intrinsic evaluation targeting compositionality, Senaldi et al. (2016) built a dataset of Italian verbal idioms and compositional expressions, and then built lexical variants of their items by replacing some of their constituents with semantically-related words. Their results showed that idioms have a lower similarity to their lexical variants in distributional embedding spaces, compared to compositional verb phrases.

On IE representations specifically, extant work adopts both extrinsic (e.g. Škvorc et al., 2022; Chakrabarty et al., 2022) and intrinsic methods (e.g. Dankers et al., 2022; He et al., 2024a). However, datasets that could be leveraged for intrinsic evaluations are mostly available in Western languages, including English, Portuguese and Galician, such as MAGPIE (Haagsma et al., 2020), AStitchInLanguageModels (Tayyar Madabushi et al., 2021), NCTTI Garcia et al. (2021a), and data from SemEval2022 (Tayyar Madabushi et al., 2022). For instance, Zeng and Bhat (2022, 2023) discussed their intrinsic evaluation tasks — *embedding clustering* and *embedding differentiation* — using the MAGPIE dataset. The clustering task involves grouping IE embeddings into clusters and examining the homogeneity within each cluster to determine if the model produces high-quality embeddings for PIEs that share similar meanings. The differentiation task, on the other hand, assesses a model's ability to distinguish between the literal and idiomatic meanings of the same PIEs within the given context. In a separate stream, He et al. (2024a), focusing on noun compounds (NC) in MWEs, presented their dataset NCIMP for probing the idiomaticity of NCs in English and Portuguese. Besides, they extend the probes from previous work such as Garcia et al. (2021b) and Klubička et al. (2023) and propose a set of measures: *Affinity* and *Scaled Similarity*. Both are relevant to similarity measuring. Affinity examines if an NC is more similar to its synonyms than to other semantically related targets and distractors, while Scaled Similarity magnifies the similarities in a specific vector space by introducing a rescaling factor (e.g., a random item as the lower bound).

In contrast, Chinese idiom datasets such as ChID (Zheng et al., 2019), PETCI (Tang, 2022), CCT (Jiang et al., 2018), CIP (Qiang et al., 2023), and IDIOMKB (Li et al., 2024), are primarily constructed to examine idiom behaviors in applied tasks, including cloze tests, translation, and paraphrasing. CHENGYU-BENCH (Fu et al., 2025), a most recent benchmark, integrates Chengyu-Bench, encompasses three distinct tasks. However, its core focus remains on evaluating surface-level contextual competence, rather than deeper semantic understanding. These resources often lack fine-grained annotations on idiom usage, semantic similarity, or contextual flexibility, making them less suitable for deeper intrinsic exploration. We argue that this limitation contributes to the scarcity of idiom-focused intrinsic evaluation frameworks in current research.

2.3 Idiomatic Semantic Textual Similarity

STS has long been a prominent area in the linguistics and NLP communities. Notable studies revolve around the general domain, such as OCNLI (Hu et al., 2020), and STS benchmarks (Agirre et al., 2016, etc.), yet few are dedicated to the idiomatic STS (iSTS) in the Mandarin language.

The iSTS task, initially introduced by Tayyar Madabushi et al. (2021), is rooted in such a theory: if a model accurately encodes an MWE, the embedding of a sentence containing the MWE should be semantically similar to the embedding of the same sentence where the MWE is substituted with another phrase. Two cases arise depending on the substitution:

1. The MWE has been paraphrased appropriately. In this case, the sentence pair $P(S, A_c)$, where S is the original sentence, A_c is derived from S by correctly paraphrasing the target MWE, should have a semantic similarity approximating 1.
2. The MWE has been replaced inappropriately. In this case, two sentence pairs $P(S, A_i)$ and $P(A_c, A_i)$, where A_i is derived from incorrect paraphrasing, should have roughly the same similarities (see Table 2 for demonstration).

This iSTS setting can be formulated as in Equation 1. It assesses if a model genuinely captures the meaning of an MWE by evaluating its ability to identify semantically similar/ dissimilar sentence pairs. A major strength is that it doesn't require human-annotated scores for incorrect sentence pairs, making it highly efficient for data annotation. The method was adopted by SemEval 2022 (Tayyar Madabushi et al., 2022; He et al., 2024a,b) as an intrinsic alternative for MWE evaluation on non-Chinese datasets.

$$\forall_{i \in I,\, c \in C} \begin{array}{l} Sim(S, A_c) \approx 1; \\ Sim(S, A_i) \approx Sim(A_c, A_i) \end{array} \tag{1}$$

The current study has the goal of the intrinsic evaluation of the representation of Chinese idioms, to provide a benchmark that is independent of downstream task performance and to diversify evaluation in idiom research.

3 Chinese Idiomatic STS

3.1 Task Statement

We adapted the iSTS framework (see section 2.3) for the intrinsic evaluation of Chinese idioms and created a new dataset: ChengyuSTS, inspired by the work of Tayyar Madabushi et al. (2021, 2022). ChengyuSTS requires constructing sentence pairs via paraphrasing or replacements. The replacements can take varied forms:

For a correct replacement, a Chinese idiom can be substituted with a synonymous or a near-synonymous unit, which could be a single word, an idiom, or a phrase. Alternatively, it can be paraphrased using an explanatory expression that conveys its meaning or a literal interpretation that allows the figurative meaning of the idiom to be inferred. Conversely, incorrect replacements may involve words, idioms, or phrases that are antonymous, contextually inappropriate, or distort the sentence's original meaning. In this sense, the replacements could also be idiomatic. Examples can be found in Table 3.

The subsequent sections detail the dataset creation pipeline.

3.2 Sample Selection

The idiom vocabulary and raw samples are collected from existing resources, including ChID (Zheng et al., 2019), CIP (Qiang et al., 2023) and CCT (Jiang et al., 2018). ChID holds a large collection of four-character Chinese idioms and serves as a gold benchmark for the task of Chinese idiom cloze reading comprehension. The CIP dataset is converted from a machine translation corpus – WMT18 (Bojar et al., 2018) and the CCT dataset contains idioms and sentences crawled online. Given that the texts in ChID are primarily long paragraphs extracted from formal sources (novels, essays and news articles), the inclusion of CIP and CCT could contribute to the diversity of the ChengyuSTS dataset.

The idiom vocabulary was primarily constructed using ChID, as its idioms have been specially collected and filtered based on frequency. We also ensured that the selected idioms are included in Xinhua Dictionary[2], an authorized Chinese-language dictionary that contains over 45,000 idioms with rich meta information. For each idiom, about 3 contexts were sampled across all corpora. We excluded those sentences where the target idiom appears more than once or is only mentioned or referred to. For example, in 他在演讲中用到了“画蛇添足”这个词。(*He used the word of "hua she tian zu" in his speech.*), the idiom 画蛇添足 is simply referred to, instead of being used in a real context. We try to exclude such instances. Also, to minimize the noise from excessively long texts, the text length is restricted to [20, 70], with ChID samples truncated to fit within this range.

[2] https://github.com/pwxcoo/chinese-xinhua.git

S	A_c	A_i	Expectation
It's a **blood bath**.	It's a **massacre**.	It's a **sanguine fluid bath**.	$Sim(S, A_c) = 1$; $Sim(S, A_i) = Sim(A_c, A_i)$
I feared that taking it would make me a **guinea pig**.	I feared that taking it would make me a **test subject**.	I feared that taking it would make me a **pig**.	

Table 2: Examples of the iSTS data from Tayyar Madabushi et al. (2022). MWEs and their replacements are in **bold**. For brevity, sentences are shortened.

	Original sentence	Correct replacement
E1	巴士的乘客被撞死，火车中的旅客却{安然无恙}。 *Passengers on the bus were killed in the crash, while the passengers on the train were {safe and sound}.*	巴士的乘客被撞死，火车中的旅客却{毫发无损}。 *Passengers on the bus were killed in the crash, while the passengers on the train were {completely unscathed}.*
E2	这些规定{朝三暮四}，叫人无所适从。 *These regulations are {full of chop and change}, leaving people at loss.*	这些规定{经常变动}，叫人无所适从。 *These regulations are {consistently changing}, leaving people at loss.*
E3	你会让这个{稳如泰山}的星座苦恼不安。 *You will disturb this zodiac that is {unshakable as Mount Thai}.*	你会让这个{像泰山一样稳固}的星座苦恼不安。 *You will disturb this zodiac that is {as solid as Mount Thai}.*

Table 3: Examples of paraphrasing, with target idioms and their replacements enclosed in *{}*. Only samples of correct replacement are presented: E1 illustrates the synonym rewording, E2 provides an explanatory substitution, and E3 presents a literal interpretation. Note that these categories are not strictly distinct, as a synonymous phrase may also be an explanation, etc.

3.3 Paraphrasing

For each raw sentence from section 3.2, we construct its homogeneous and adversarial samples via correct and incorrect replacements, as outlined in section 3.1.

In practice, the construction is a collaborative effort between humans and AI. We aim to ensure that the paraphrasing process is dynamic and context-dependent, instead of mechanically substituting an idiom with a set, pre-determined word/ phrase. The use of AI, besides reducing the annotation workload, brings more diversity to the process, thanks to its stochasticity.

Initially, ChatGPT (OpenAI, 2024) is used to paraphrase the target idiom in each sentence using a given candidate, while retaining grammaticality or coherence (an example prompt can be found in Table 4). These candidate replacements, including definitions, synonyms, and antonyms, are extracted from Xinhua Dictionary. Two human experts – both native Mandarin speakers with a Master's degree in linguistics – then compare and review the AI-generated results. Sentences with inappropriate paraphrasing are either revised or discarded. In case of an incorrect replacement, the generated sentence might be slightly incorrect in grammar, which we deem acceptable considering that this is a result of manipulation.

This context-aware and flexible paraphrasing process can produce sentence instances that integrate both the static meaning of an idiom (its dictionary definition, literal or figurative) and its dynamic interpretation (its meaning as shaped by context). Sometimes, the replacements may be morphological variants of the original idioms, through reorganization, abbreviation, or modification, such as 总而言之 -> 言而总之 (*in brief*); 必恭必敬 -> 恭敬 (*being respectful*); 故步自封-> 固步自封 (*being conservative*). By analyzing these sentence pairs, we can better examine whether a model has truly learned an idiom as a holistic unit and grasped the full range of its meaning, rather than merely memorizing its components or capturing a context-insensitive meaning.

3.4 Final Corpus

A fine-tuned model may *cheat* in evaluation by simply assigning a perfect 1 to any sentence pair (Tayyar Madabushi et al., 2021). To prevent this, we deliberately spiced the final corpus with standard STS data. Note that the paraphrased sentences from section 3.3 may exhibit high lexical and structural similarity to their original, while regular STS data may not. For this reason, we chose not to use common datasets such as the Chinese STS-B (Cer et al., 2017), as they don't meet our criteria. Instead, we retrieved about 6,000 Mandarin sentence pairs from PAWS-X (Yang et al., 2019a) and LCQMC

请对以下句子中的成语部分进行改写，尽量保持句子原意和结构不变，并保持改写后的句子语法正确、自然。

句子：他总是对牛弹琴，没人听得懂他的专业术语。

Idiom：对牛弹琴

Please paraphrase only the idiom in the following Chinese sentence. The paraphrased version should preserve the original meaning and structure as much as possible, and be grammatically correct and natural.

Sentence: *He is always cating pearls before swine — no one understands his technical jargon.*

Idiom: 对牛弹琴 (*to cast pearls before swine*)

Table 4: An example prompt for paraphrasing using ChatGPT.

(Liu et al., 2018), prioritizing those sharing similar structure and wording. Detailed statistics of the final corpus are provided in Table 5 and Table 6 displays some data examples.

	Idiom	Homo.	Advrl.	Total
Train	3,452	9,129 (1,911)	8,898 (2,223)	18,027 (4,134)
Dev	1,219	1,968 (491)	1,895 (515)	3,863 (1,006)
Test	1,153	1,853 (460)	2,010 (549)	3,863 (1,009)
All	3,452	12,950 (2,862)	12,803 (3,287)	25,753 (6,149)

Table 5: Statistical details of the ChengyuSTS dataset. Numbers enclosed in *()* correspond specifically to the counts from the standard STS data. *Homo.* represents a homogeneous sentence pair, e.g., (S, A_c), and *Avrl.* denotes an adversarial sentence pair, i.e., (S, A_i) or (A_c, A_i).

4 Experiments

To establish baseline performance for the ChengyuSTS dataset, we use PLMs to generate sentence embeddings for each sentence pair and then evaluate their alignment with annotations by measuring their pairwise cosine similarity.

We first evaluated the performance of current models on the ChengyuSTS dataset in a *zero-shot* setting in section 4.1, where a model was tested without continuing pre-training or fine-tuning on the ChengyuSTS data. Subsequently, we presented fine-tuned models under the contrastive learning framework in section 4.2. The performance was assessed with the Spearman correlation coefficient, following Tayyar Madabushi et al. (2022); He et al. (2024a,b). [3]

4.1 Zero-shot Evaluation

Importantly, the sentence representations produced by most PLMs, such as BERT (Devlin et al., 2018), are not inherently suitable for the iSTS task due to *anisotropy*(Gao et al., 2019; Li et al., 2020), meaning that the embeddings may not be distributed uniformly across the entire space but rather concentrate within a hypercone, leading to a lack of semantic isometry of the embedding space and introducing biases in cosine similarity measurements (Gao et al., 2019). As evidenced by Reimers and Gurevych (2019), sentence embeddings generated by simply averaging word embeddings or using the CLS-token output from a Transformer demonstrate poor performance in STS tasks. Given this, we sought PLMs that are specifically tailored for sentence representations and are available in Mandarin[4]:

- **Chinese-SBERT**$_{general}$: a Mandarin-adapted sentence Transformer trained for Mandarin STS task.
- **Chinese-SBERT** (Reimers and Gurevych, 2019; Zhao et al., 2019, 2023): a Mandarin-adapted sentence Transformer, trained for Mandarin NLI (natural language inference) task.
- **XLM-SBERT** a multilingual Sentence Transformer for clustering or semantic search tasks.
- **XLM-SBERT-MPNet**: a multilingual Sentence Transformer for clustering or semantic search tasks.

[3] The experimental details, including prompts and parameters, can be found in the GitHub repo.

[4] Model keys on Hugging Face (sorted in chronological order). sbert-base-chinese-nli, sbert-chinese-general-v2, distiluse-base-multilingual-cased-v1, paraphrase-multilingual-mpnet-base-v2, simcse-roberta-large-zh and promcse-bert-large-zh.

	S	A_c	A_i	sim
E1	有的网友{付之一笑}，同时也有15名网友积极响应... *Some netizens {brushed it off with a smile}, while 15 others responded actively...*	有的网友{一笑了之}，同时也有15名网友积极响应... *Some netizens {laughed it off}, while 15 others responded actively...*	有的网友{大笑不止}，同时也有15名网友积极响应... *Some netizens {couldn't stop laughing}, while 15 others responded actively...*	
E2	这个丑闻有可能使原本大有前途的政治生涯{戛然而止}。 *This scandal could {bring an abrupt end to} what was once a highly promising political career.*	这个丑闻有可能使原本大有前途的政治生涯{突然中断}。 *This scandal could {abruptly interrupt} what was once a highly promising political career.*		1.0
E3	然而，大多数白马皮肤呈粉红色，有些则有蓝眼睛。 *However, most white horses have pink skin, and some have blue eyes.*		然而，大多数粉红马有白色皮肤和一些蓝眼睛。 *However, most pink horses have white skin, and some have blue eyes.*	0.0

Table 6: Examples from the ChengyuSTS dataset. Target idioms and their replacement in iSTS data are denoted by *{}*. The original sentence S and its correct paraphrase A_c share the same translation, and the translation for its incorrect paraphrase A_i might be partially omitted. For instances with an explicit sim label (e.g., E2 and E3), we expect the model to produce a score approximating the label. In other cases (e.g., E1), the model is expected to assign similar scores to (S, A_i) and (A_c, A_i).

	S	A_c
E1	指挥官还制订在全球多个地点发动{先发制人}或者报复性袭击的计划。 *The commander has also formulated plans to launch {preemptive} or retaliatory strikes at multiple locations around the world.*	指挥官还制订在全球多个地点发动{先声夺人}或者报复性袭击的计划。 *The commander has also formulated plans to launch {show-of-force} or retaliatory strikes at multiple locations around the world.*
E2	许巍的歌声给一代人留下了{不可磨灭}的回忆。 *Xu Wei's singing has left an {inedible} memory for a generation.*	许巍的歌声给一代人留下了{不会随时间消逝}的回忆。 *Xu Wei's singing has left a generation with a memory that won't fade with time.*

Table 7: Hard examples: Only homogeneous sentence pairs (i.e., pairs with an expected similarity of 1) are included. These examples receive calculated similarity scores lower than 0.6 from both DeepSeek and Chinese-SBERT. It can be observed that the model struggles to identify similarity when the target idiom is replaced with either a near-synonym (e.g., E1) or a literal interpretation (e.g., E2), suggesting the models' weakness in iSTS.

	Overall	iSTS only
Chinese-SBERT	30.13	47.38
Chinese-SBERT$_{general}$	37.90	42.32
XLM-SBERT	10.59	32.21
XLM-SBERT-MPNet	21.84	35.98
SimCSE-RoBERTa	40.44	58.75
PromCSE-BERT	42.37	56.81
DeepSeek (R1)	**72.67**	**89.89**

Table 8: Spearman correlation scores the *zero-shot* evaluation results on the test set (reported in %). *Overall* indicates scores on the entire test dataset (3,863 instances), while *iSTS* shows the scores solely on the iSTS data (2,834 instances).

- **SimCSE-RoBERTa**: a RoBERTa trained using the SimCSE framework (Gao et al., 2021) for sentence embeddings.

- **PromCSE-BERT**: a BERT trained using the PromCSE framework (Jiang et al., 2022) for sentence embeddings.

Besides, we also presented the evaluation results with DeepSeek (DeepSeek-AI, 2025), a more recent and advanced LLM that has outperformed other LLMs, including GPT-4o (OpenAI, 2024) and Claude-3.5 Sonnet (Anthropic, 2024), in multiple Chinese benchmarks. In practice, we prompted the model to assign a similarity score to each sentence pair within [0, 1], where 0 suggests complete dissimilarity and 1 perfect similarity.[5]

The results are provided in Table 8. DeepSeek achieves a strong performance, while others show their weakness in discriminating similar/ dissimilar sentences when Chinese idioms are replaced, even though they have been (continuing) pre-trained on large-scale corpora for NLI or STS tasks. This suggests that such models fall short of modeling the underlying meanings of idioms (see Table 7).

While the performance of the DeepSeek model can be seen as an upper bound, the adoption of specialized approaches may be necessary to improve the understanding of less powerful models like PLMs. In the next section, we propose an additional experiment with contrastive learning.

4.2 Contrastive Learning to Mandarin Idiomatic Representation

Contrastive learning, which brings similar embeddings closer and pushes dissimilar ones apart, has been suggested to enhance the semantic isometry of embedding spaces. As further evidenced in Table 8, the SimCSE framework exhibits notable advantages in capturing semantic similarity. Motivated by these findings, we fine-tuned models directly on Chengyu-STS using the supervised version of SimCSE (Gao et al., 2021). The training objective of SimCSE within a mini-batch N is defined in Equation 2, where h_i denotes the representation of sample x_i, $sim(h_1, h_2)$ the cosine similarity between h_1 and h_2, and τ is a temperature hyperparameter.

$$-\log \frac{e^{sim(h_i,h_i^+)}/\tau}{\sum_{j=1}^{N}(e^{sim(h_i,h_j^+)/\tau} + e^{sim(h_i,h_j^-)/\tau})} \quad (2)$$

The key aspect of this approach is to construct a triplet (x_i, x_i^+, x_i^-) for instance x_i, where x_i^+ is semantically similar to x_i and x_i^- is dissimilar. In our experiments, we used the correctly paraphrased instances of x_i as positive samples x_i^+, and their incorrect paraphrases as negative samples x_i^-. If an x_i^+ is unavailable, we duplicate x_i as its positive sample and in the case of a missing x_i^-, we sample a random instance from the dataset as a substitute.

In zero-shot evaluation, only a limited range of models are available. However, for fine-tuning, we explored a broader variety of base models, including Mandarin-adapted BERT, RoBERTa, and XLNet (Yang et al., 2019b; Cui et al., 2019, 2020)[6]. The results, obtained in 4 training epochs, are reported in Table 9.

	Overall	iSTS only
BERT-Chinese	60.03	82.81
Chinese-BERT-WWM	**63.65**	83.95
Chinese-RoBERTa-WWM	63.28	**84.40**
Chinese-XLNet	44.50	79.87

Table 9: Spearman correlation scores of the *fine-tuning* experiments results on the test set (reported in %). *WWM* denotes whole-word-masking, a pre-training strategy (Cui et al., 2019).

The *fine-tuning* results present a great improvement over the zero-shot evaluation. The contrast between them suggests that current sentence Transformers struggle to model idiomaticity, without specific adaptation. The fact that Chinese-RoBERTa-WWM (Cui et al., 2020, 2019) receives the best fine-tuning performance on iSTS data may stem from its WWM strategy during the pre-training phase. By masking entire words rather than compositional units, the model is encouraged to learn the representations of complete semantic units, which aligns well with the fixed structures of idiomatic expressions. However, the performance drop on the overall test set may arise from the inherent discrepancy between standard STS tasks and iSTS tasks (requiring figurative semantic reasoning), indicating that the PLM's generalization ability is still limited.

While the SimCSE-fine-tuned models still fall short compared to DeepSeek, the improvements they achieved remain meaningful, considering the gaps between them in model scale (e.g., DeepSeek has about 7B parameters while Chinese-RoBERTa-WWM only has about 100M[7]) and data volume. The performance gap could be further bridged in the future with other lightweight strategies.

5 Conclusion and Future Work

Chengyu idioms are an essential component of Chinese linguistics and cultural heritage, yet their computational processing remains underexplored in NLP, due to limited task-specific datasets and evaluation frameworks.

In this paper, we have introduced iSTS (Tayyar Madabushi et al., 2021), the idiom-aware se-

[5] We didn't compare other LLMs, as DeepSeek has already demonstrated superior performance on multiple Chinese benchmarks, and that evaluation with LLMs is not the major goal of the study.

[6] Model keys on the Hugging Face Hub (sorted in chronological order): bert-base-chinese, chinese-bert-wwm-ext, chinese-roberta-wwm-ext, and chinese-xlnet-base.

[7] Information obtained by using the *Transformers* library.

mantic textual similarity task, into Chinese, accompanied by a curated dataset ChengyuSTS. Our experiments reveal that pre-trained sentence Transformer models fail to capture idiomaticity in Mandarin under the zero-shot setting, and we presented fine-tuned models using SimCSE (Gao et al., 2021), which significantly improved the performance.

Finally, our work is the first presenting evaluation results for the DeepSeek model on Chinese idiom representation. DeepSeek achieved by far the strongest performance, and it will likely represent the reference for future task approaches.

Limitations

We identify the following limitations in the study:

Data. The paraphrased sentences could be rigid in syntax or inappropriate in grammar, especially in the incorrect replacement scenario. Also, due to time and resource constraints, we could only source limited amounts of standard STS data from existing datasets rather than constructing one. The differences between these datasets and our iSTS objectives may cause performance discrepancies (as seen in Table 8 and 9).

Experiments. The limited variety of Sentence Transformers and CSE models on Mandarin restricts our exploration under the *zero-shot* setting. And our *fine-tuning* experiments only employ the SimCSE framework. Future work should incorporate more diverse model families and contrastive learning paradigms.

Acknowledgements

EC was supported by a GRF grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 15612222). AV was partly supported by MRC-FAPESP AIM-Health project.

Ethical Considerations

The datasets and codes used in the study are publicly available, and we strictly followed the terms of use specified by their original providers. The annotators were provided with informed consent and were allowed to withdraw freely. AI tools were used in compliance with API terms, with no sensitive data exposure.

References

Pranav A, Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu, Alessandro Lenci, et al. 2024. Comparing static and contextual distributional semantic models on intrinsic tasks: An evaluation on mandarin chinese datasets. In *Proceedings of LREC-COLING*.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation.

Anthropic. 2024. Claude 3.5 sonnet model card addendum.

Roberto Ascari, Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, and Mario Mezzanzanica. 2024. A fistful of vectors: a tool for intrinsic evaluation of word embeddings. *Cognitive Computation*, 16(3):949–963.

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Rhys Biddle, Aditya Joshi, Shaowu Liu, Cecile Paris, and Guandong Xu. 2020. Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter. In *Proceedings of the web conference 2020*, pages 1217–1227.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cristina Cacciari. 2014. Processing multiword idiomatic strings: Many words in one? *The Mental Lexicon*, 9(2):267–293.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP*, pages 1–6.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

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*.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. Chengyu-bench: Benchmarking large language models for chinese idiom understanding and use.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*.

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

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. 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 2730–2741, Online. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics*, pages 3551–3564. Association for Computational Linguistics (ACL).

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *12th Language Resources and Evaluation Conference: LREC 2020*, pages 279–287. European Language Resources Association (ELRA).

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2016. Adaptive joint learning of compositional and non-compositional phrase embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 205–215, Berlin, Germany. Association for Computational Linguistics.

Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024a. Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss. *arXiv preprint arXiv:2406.15175*.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2024b. Investigating idiomaticity in word representations. *Computational Linguistics*, pages 1–48.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. 2020. Ocnli: Original chinese natural language inference. *arXiv preprint arXiv:2010.05444*.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2022. Improved universal sentence embeddings with prompt-based contrastive learning and energy-based learning. *arXiv preprint arXiv:2203.06875*.

Zhiying Jiang, Boliang Zhang, Lifu Huang, and Heng Ji. 2018. Chengyu cloze test. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana. Association for Computational Linguistics.

Filip Klubička, Vasudevan Nedumpozhimana, and John D Kelleher. 2023. Idioms, probing and dangerous things: Towards structural probing for idiomaticity in vector space. *arXiv preprint arXiv:2304.14333*.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2023. A comparative evaluation and analysis of three generations of distributional semantic models. *Language resources and evaluation*, 56(4):1269–1313.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Bing Li, Xiaochun Yang, Bin Wang, Wei Wang, Wei Cui, and Xianchao Zhang. 2018a. An adaptive hierarchical compositional model for phrase embedding. In *IJCAI*, pages 4144–4151.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Minglei Li, Qin Lu, Dan Xiong, and Yunfei Long. 2018b. Phrase embedding learning based on external and internal context with compositionality constraint. *Knowledge-Based Systems*, 152:107–116.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

Pengfei Liu, Kaiyu Qian, Xipeng Qiu, and Xuan-Jing Huang. 2017. Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1204–1213.

Te-hsin Liu and Lily I-Wen Su. 2021. Chinese idioms as constructions: Frequency, semantic transparency and their processing. *Language and Linguistics*, 22(4):558–592.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. LCQMC:a large-scale Chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1952–1962, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Siyu Long, Ran Wang, Kun Tao, Jiali Zeng, and Xinyu Dai. 2020. Synonym knowledge enhanced reader for chinese idiom reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3684–3695.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

OpenAI. 2024. Gpt-4o system card.

Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11:740–754.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 298–307.

Marco Silvio Giuseppe Senaldi, Gianluca E Lebani, and Alessandro Lenci. 2016. Lexical variability and compositionality: Investigating idiomaticity with distributional semantic models. In *Proceedings of the ACL Workshop on Multiword Expressions*.

Ying Sha, Mingmin Wu, Zhi Zeng, Xing Ge, Zhongqiang Huang, and Huan Wang. 2023. A prompt-based representation individual enhancement method for chinese idiom reading comprehension. In *International Conference on Database Systems for Advanced Applications*, pages 682–698. Springer.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Tadej Škvorc, Polona Gantar, and Marko Robnik-Šikonja. 2022. Mice: mining idioms with contextual embeddings. *Knowledge-Based Systems*, 235:107606.

Kenan Tang. 2022. Petci: A parallel english translation dataset of chinese idioms. *arXiv e-prints*, pages arXiv–2202.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benjamin K Tsou. 2012. Idiomaticity and classical traditions in some east asian languages. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 39–55.

Yulia Tsvetkov, Manaal Faruqui, and Chris Dyer. 2016. Correlation-based intrinsic evaluation of word vector representations. *arXiv preprint arXiv:1606.06710*.

Xinyu Wang, Hongsheng Zhao, Tan Yang, and Hongbo Wang. 2020. Correcting the misuse: A method for the chinese idiom cloze test. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10.

Mingmin Wu, Yuxue Hu, Yongcheng Zhang, Zeng Zhi, Guixin Su, and Ying Sha. 2024. Mitigating idiom inconsistency: A multi-semantic contrastive learning method for chinese idiom reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19243–19251.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019a. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Wenpeng Yin and Hinrich Schütze. 2014. An exploration of embeddings for generalized phrases. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 41–47.

Wenpeng Yin and Hinrich Schütze. 2016. Discriminative phrase embedding for paraphrase identification. *arXiv preprint arXiv:1604.00503*.

Mo Yu and Mark Dredze. 2015. Learning composition models for phrase embeddings. *Transactions of the Association for Computational Linguistics*, 3:227–242.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

Ziheng Zeng and Suma Bhat. 2022. Getting bart to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

Ziheng Zeng and Suma Bhat. 2023. Unified representation for non-compositional and compositional expressions. *arXiv preprint arXiv:2310.19127*.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

Zhe Zhao, Yudong Li, Cheng Hou, Jing Zhao, et al. 2023. Tencentpretrain: A scalable and flexible toolkit for pre-training models of different modalities. *ACL 2023*, page 217.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. Chid: A large-scale chinese idiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787.

He Zhou, Yu Yin Hsu, and Emmanuele Chersoni. 2024. Evaluating Chinese Noun Compound Interpretation in Natural Language Inference. In *Proceedings of the Chinese Lexical Semantics Workshop (CLSW 2024)*.