CMDAG: A Chinese Metaphor Dataset with Annotated Grounds as CoT for Boosting Metaphor Generation

Yujie Shao[2*] **Xinrong Yao**[3*] **Xingwei Qu**[1,6*]
Chenghua Lin[6] **Shi Wang**[8] **Stephen W. Huang**[9] **Ge Zhang**[1,4,5,7] **Jie Fu**[1,5]
[1]HKUST [2]University of California, San Diego [3]Massachusetts Institute of Technology [4]University of Waterloo
[5]Multimodal Art Projection Research Community [6]University of Manchester [7]Stardust.ai
[8]Institute of Computing Technology, Chinese Academy of Sciences [9]harmony.ai

Abstract

Metaphor is a prominent linguistic device in human language and literature, as they add color, imagery, and emphasis to enhance effective communication. This paper introduces a large-scale high quality annotated Chinese Metaphor Corpus, which comprises around 28K sentences drawn from a diverse range of Chinese literary sources, such as poems, prose, song lyrics, etc. To ensure the accuracy and consistency of our annotations, we introduce a comprehensive set of guidelines. These guidelines address the facets of metaphor annotation, including identifying tenors, vehicles, and grounds to handling the complexities of similes, personifications, juxtapositions, and hyperboles. Breaking tradition, our approach to metaphor generation emphasizes grounds and their distinct features rather than the conventional combination of tenors and vehicles. By integrating "ground" as a CoT (Chain of Thoughts) input, we are able to generate metaphors that resonate more with real-world intuition. We test generative models such as Belle, Baichuan, and Chinese-alpaca-33B using our annotated corpus. These models are able to generate creative and fluent metaphor sentences more frequently induced by selected samples from our dataset, demonstrating the value of our corpus for Chinese metaphor research. The code is available in the https://anonymous.4open.science/r/Chinese_Metaphor_Explanation-63F2.

Keywords: chinese metaphor corpus, metaphor annotation

Figure 1: Sketch Map of the Metaphorical Language Writing Process.

1. Introduction

Metaphor is a prominent linguistic device in human language and literature, typically to draw a comparison between disparate objects or concepts with the intent to make the expression more vivid, or make abstract concepts easier to understand.

With the progression of computational linguistics, there is an increasing focus on metaphor generation through machine learning techniques, notably in chatbot applications. Zheng et al. (2020) shows how machine-generated nominal metaphors (NMs) can significantly enhance user interest during interactions with chatbots. Li et al. (2022a) finds substantial applications in shaping downstream task outputs in Natural Language Generation (NLG). Notably, a series of evaluations by Chakrabarty et al. (2020, 2021) conduct human evaluations comparing literal expressions from machine-generated stories and poems with machine-generated metaphors and find users prefer the text with metaphors.

However, metaphors are referred to as novel linguistic expressions where an object or concept is used outside of its normal conventional meaning to express another meaning under a given context. Intrinsically, metaphors do not reside within the language itself but in the way they conceptually map one mental domain onto another in application (Lakoff, 1992), as shown in Fig. 1.

With this consideration, we establish metaphor sentences centered on identifying the conceptual mappings within metaphors, specifically GROUNDS (喻意). Metaphors consist of two components: TENORS (本体), representing the actual subject, and VEHICLES (喻体), symbolizing the comparative element. Employing GROUNDS can enhance sentence fluency and creativity, achieving human-like metaphorical expression (Yang et al., 2023). This work introduces an annotated Chinese metaphor corpus (**CMDAG**) that is derived from a diverse range of Chinese literature. Every metaphorical sentence within the corpus is accompanied by its corresponding GROUNDS. The central aim of our annotation effort is to accurately annotate each metaphor with a well-defined tuple of features (TENOR, VEHICLE and GROUND), capturing its in-

LREC-COLING 2024, pages 3357–3366
20-25 May, 2024.

trinsic elements.

To evaluate the effectiveness of our annotated corpus for Chinese metaphor generation, we undertake two evaluative setups, both incorporating GROUNDS and the Chain of Thought (CoT) capability of language models. In the first setup, we prompt with TENOR and VEHICLE, and we allow the language model to deduce GROUNDS. For the second setup, we prompt TENORS and GROUNDS, and fine-tune metaphorical sentence generation by asking the language model to deduce a plausible VEHICLE. In summary, our paper outlines the following three contributions:

1. We present **CMDAG**, a unique Chinese metaphor dataset, wherein a key feature is the inclusion of GROUNDS. This dataset's thoughtful design enables the intuitive generation of metaphorical constructs, addressing a notable absence in contemporary research literature.

2. We introduce a metaphor annotation pipeline by leveraging academically specialized annotators' expertise, achieving enhanced precision in pinpointing the GROUNDS of metaphors.

3. We propose the first work introducing CoT into metaphor generations. Given TENOR and VEHICLE, deriving GROUNDS using CoT, language models can generate coherent and integrative metaphor sentences. Furthermore, by combining TENORS and GROUNDS, we enhance the generation of VEHICLE, improving the quality of the generated metaphorical expressions.

2. Related Work

2.1. Chinese Metaphor Corpora

Metaphor is not only a literature of rhetoric but also a way of thinking rooted in Chinese culture (Lin, 2021). However, due to the shortage of Chinese corpora (Zhang et al., 2023), researchers are still in lack of high-quality Chinese metaphor corpora. Lyrics and Poetry corpora released by (Liu et al., 2019) provide a great source of metaphorical Chinese language, but they do not dig in and provide fine-grained annotations of existing Chinese similes and metaphors in Lyrics and Poetry corpora. CS (Zhang et al., 2021), another large Chinese rhetorical corpus, is in shortage of Fine-grained annotation as well. CMC (Li et al., 2022b) is a valuable Chinese metaphor corpus with cautious annotation of tenors and vehicles, but CMC is pretty small and without the annotation of GROUNDS (喻意). GraCe (Yang et al., 2023), an amazing contemporaneous research work, claims to provide a carefully annotated Chinese simile corpus but hasn't been released yet, and only focuses on clearly stated Chinese similes. As a sharp contrast, CMDAG is a carefully annotated large Chinese metaphor (also with simile) corpus with annotations of all TENOR, VEHICLES, and GROUNDS, which is a valuable resource for researchers interested in Chinese metaphor processing. We briefly compare the existing major Chinese metaphor/simile corpora in Tab. 1.

2.2. Boosting NLG via Chain-of-Thought

Chain-of-Thought (**CoT**) is the most important inference trick inducing Large-scale Language Models (**LLMs**) to output reasonable results (Wang et al., 2023) since proposed by Wei et al. (2022). It has been widely used in different LLM-based Natural Language Generation (**NLG**) tasks, including human moral value alignment (Liu et al., 2023, 2022), math problem solving (Yue et al., 2023), and evaluation of NLG results (Jiang et al., 2023; Chan et al., 2023).

As illustrated in Fig. 1, we believe that GROUNDS is the natural CoT connecting TENOR with VEHICLE, which has been discussed in literature research works (Black et al., 1979; End, 1986) and NLP research works (Gong, 2003; Stowe et al., 2021; Wachowiak and Gromann, 2023). Specifically, Li et al. (2023) propose to introduce explicit basic meaning modeling to boost metaphor detection. Additionally, Yang et al. (2023) reveal that simile generation could benefit from pre-specified constraints, especially explicitly stated GROUNDS. As a sharp comparison, CMDAG directly verifies how LLMs perform on Chinese metaphor generation in various settings, especially with the assistance of GROUNDS (喻意) as CoT.

3. Chinese Metaphor Dataset

In this section, we present our annotated dataset of Chinese metaphors. Subsequent subsections establish basic definitions used in our dataset, and provide detailed insights into the data collection and annotation processes.

3.1. Definition

A **Metaphor (**暗喻/隐喻**)** is a linguistic device in which a word or phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness or analogy between them. For example, the metaphor "何等动人的一页又一页篇章！这是人类思维的花朵。" compares the tenor, the pages of literature (一页页篇章), to the vehicle, bloom of human thoughts (人类思维的花朵), to convey the beautiful nature of the literature's expressions. In CMDAG, we uniformly formalize and

Table 1: Statistic characteristics and annotation information of main existing Chinese metaphor/simile datasets of metaphor and simile and CMDAG dataset. W and F separately denote the tenor/vehicle words and the corresponding feature words.

Dataset	# Nums	Tenor	Vehicle	Ground	Context	Open-source
		W/F	W/F		Above/Below	
Poetry (Liu et al., 2019)	43,051	−/−	−/−	−	✓/−	✓
Lyrics (Liu et al., 2019)	246,669	−/−	−/−	−	✓/−	✓
CS (Zhang et al., 2021)	5,490,721	−/−	−/−	−	✓/✓	✓
CMC (Li et al., 2022b)	2,787	✓/−	✓/−	−	−/−	✓
GraCe (Yang et al., 2023)	61,360	✓/✓	✓/✓	✓	✓/✓	−
CMDAG	27,989	✓/✓	✓/✓	✓	✓/✓	✓

Source Type	# Literature Works	# Likely-Metaphors	# Annotated Metaphors
Prose/Poem	3,459	28,553	5,294
Song Lyrics	102,197	109,827	21,276
Contemporary Poem	4,494	7,268	939
HipHop/Rap Lyrics	3,004	7,603	480
Total	113,154	153,251	27,989

Table 2: Statistics of CMDAG and its raw data collection literature sources.

process Chinese similes and metaphors for convenience, since similes are also sometimes referred to as direct metaphors[1].

To further explain other annotated elements, **Tenor (**本体**)** is the literal object or idea being described, and **Vehicle (**喻体**)** is the object or idea carrying the weight of comparison. The **Ground (**喻意**)** of a metaphor/simile is the concept or concepts the tenor and vehicle share, enabling the metaphor to align with common sense.

3.2. Data Collection

In constructing our corpus, we first collect a raw set of ∼153K probable metaphoric sentences from various Chinese literary sources online, with a focus on genres such as prose [2], poems [3], and song [4] and rap/hip-hop [5] lyrics, which are often renowned for their rich usage of literary techniques and devices. Statistics of our raw and annotated metaphor datasets, separated by source types, are shown in Tables 2 and 3. We applied the following set of heuristic rules to detect sentences which are likely to be of metaphoric usage, as opposed to literal ones, if either:

- The sentence contains Chinese simile comparators ("像", "好似", "如同", etc.), or
- We identify metaphors by applying a similar method as in Su et al. (2017), where the sentence is classified as metaphoric if its subject and object, identified through dependency parsing, are not highly related and do not have a hyponym/hypernym relationship. We query whether the subject is a hyponym or a hypernym of the object in WordNet, and determine the relatedness between the subject and object by computing their cosine similarity score (a low score indicates the subject and object are less related, and hence there exists little shared information between them).

  Suppose the subject and object are represented by n-dimensional vectors **w** and **v** respectively, then their cosine similarity score is computed as:

$$\cos(\mathbf{w}, \mathbf{v}) = \frac{\sum_{i=1}^{n} w_i v_i}{\sqrt{\sum_{i=1}^{n} w_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}$$

  where sentences with a score below a set threshold of $\cos(\mathbf{w}, \mathbf{v}) \leq 0.575$ (which from the results by Su et al. (2017) gives the best performance and accuracy) are considered likely-metaphors and are kept for annotation.

3.3. Data Annotation

Our data annotation goal is to accurately mark each metaphor with a well-defined tuple of features: (TENOR, VEHICLE, GROUND).

[1] Relation to metaphor from BNC Baby specifications.
[2] https://www.ppzuowen.com/book/sanwen/
[3] https://github.com/yuxqiu/modern-poetry
[4] https://github.com/dengxiuqi/ChineseLyrics
[5] https://github.com/djwackey/chinese-hiphop-lyrics

Source Type	Average Context Length (Tokens)
Prose/Poem	101
Song Lyrics	49
Contemporary Poem	51
HipHop/Rap Lyrics	52
Overall	59

Table 3: Statistics of CMDAG and its raw data collection literature sources.

Consider the metaphor "天上的云像奔腾的骏马" (translates to: "clouds in the sky are like galloping horses"), annotated as: (云(clouds), 奔腾的骏马(galloping horses), 相似的形态(similar forms)). A compilation of examples from our annotated dataset is presented in Table 4.

Our annotation process consists of two main stages, focusing on both coarse-level and fine-level annotations:

Preliminary Annotation: Here, we engage a 20-people team of Chinese college students to identify genuine metaphors from the initial dataset, and highlight potential TENOR's and VEHICLE's. Each sample is annotated by two annotators at this stage.

Refined Annotation: Leveraging the groundwork from the initial round, a second cohort of annotators, primarily Chinese native speakers with at least undergraduate credentials in Chinese Literature, refines the annotations. Their specialized background enables them to further pinpoint the GROUND of the metaphors with higher precision. We provide comprehensive guidelines to ensure consistent annotation quality, which mandates that each data piece is assessed by at least three annotators, improving label consistency and accuracy. Our labeling strategy emphasizes sophisticated composition of our GROUND labels, ensuring a structure combining an *Adjective* and a *Noun* (形容词 + 名词). The noun part delineates the shared characteristic between the TENOR and VEHICLE, while the adjective highlights the dimension underscoring their connection. Fig. 2 showcases the diverse adjectives and noun elements of our annotated grounds via word clouds.

A cornerstone of our labeling strategy is the sophisticated composition of our GROUND labels. We ensure that they consistently adopt a structure melding both an *Adjective* and a *Noun* (形容词 + 名词). Specifically, the noun portion delineates the shared characteristic linking the TENOR and VEHICLE. Meanwhile, the accompanying adjective furnishes the dimension or aspect underscoring their connection. Fig. 2 showcases the diverse adjectives and noun elements of our annotated grounds via word clouds.

Guidelines for Refined Annotation in Chinese: Our rigorous annotation approach is demonstrated through strict guidelines for the second annotation round, focusing on intricate annotation of metaphorical components in Chinese text.

1) Annotation and Quality Inspection Rules: Given Chinese rhetoric's complexity, it's essential to label all rhetorical devices like metaphor, metonymy, simile, personification, etc., in a unified standard. Annotators reference prior annotations, remaining cautious against possible inaccuracies, especially regarding previous GROUND labels, as we standardize the formatting requirements in the second round. A large proportion of statements contain multiple possible tuples of (TENOR, VEHICLE, GROUND). Annotators separate different tenors, vehicles, and grounds by three predetermined quotation marks when labeling each statement. Correctness verified, multiple tuples of one statement are automatically retrieved by string matching, forming the current open-sourced corpora.

2) Selection of Nouns for Grounds: To uphold annotation authenticity and precision, a curated list of nouns is provided. The emphasis is on opting for more descriptive nouns, avoiding generic terms like 样子(appearance), 特征(feature), 特点(characteristic), 感受(feeling), 感觉(sensation), and other similar broad terms. This curated list is crucial for ensuring that the grounds aptly reflect the nuanced connections between the TENOR and VEHICLE.

4. Methodology

We consider two common scenarios that people often encounter with metaphor usage in writing. By utilizing our metaphor dataset with annotated grounds and applying a Chain-of-Thought (CoT) prompting technique with generated knowledge, we examine the importance of GROUND labels in metaphor generation tasks, including

Task 1: Ground Identification The first task that we consider is a situation where given a potential pair of TENOR and VEHICLE as the subject and object which we would like to connect and compare, the model is to generate a corresponding metaphor.

Task 2: Vehicle Identification Our second task requires the model to produce a metaphor when provided with a TENOR as the topic and a potential GROUND that signifies the features of the TENOR we aim to emphasize.

4.1. Prompt Engineering

With multiple-prompt prompting, we motivate in-context learning (ICL) by first of all providing the model with several examples (as in few-shot learn-

Source Type	Sentence	Tenor	Vehicle	Ground
Prose/Poem	雨，像银灰色黏湿的蛛丝，织成一片轻柔的网，网住了整个秋的世界。 The rain is like silver-gray sticky spider silk, weaving into a soft net that captures the entire realm of autumn.	雨 rain	蛛丝 spider silk	细长的形状 elongated shape
	佛法就好像手中的一块玉，如果没有握过许多泛泛的石头，就不能了解手中的玉是多么珍贵了。 Buddhism is like a piece of jade in your hand, if you have not held many ordinary stones, you cannot understand how precious the jade is.	佛法 Buddhism	玉 jade	珍贵的属性 preciousness
Song Lyrics	我以为旅人将我热情都燃尽 你却像一张情书感觉很初级 I thought travelers would burn out all my passion, but you are like a love letter, which feels so elementary	你 you	情书 love letter	稚嫩的感情 immature emotion
	爱像一阵风 吹完它就走 Love is like a gust of wind, it blows away and then goes away	爱 love	风 wind	短暂的经过 a brief passage
Contemporary Poem	花香仿佛消散的钟声 The fragrance of flowers is like the dissipation of bells	花香 fragrance of flowers	消散的钟声 dissipated bells	浅淡的感觉 light feeling
	我的颊像溶了的雪，我的心像热了的酒 My cheeks are like melted snow, and my heart is like warm wine	我的脸颊;我的心 my cheeks;my heart	溶了的雪; 热了的酒 melted snow;warm wine	温暖的感觉;炙热的感觉 warm feeling;hot feeling
HipHop/Rap Lyrics	我的努力依旧还不够 如同大海里的虾米 My efforts are still not enough, like shrimps in the sea	我 Myself	大海里的虾米 shrimps in the sea	渺小的状态 insignificance
	纵然现实残酷 惟有向着上天祷告 生存就仿似生活响一个战场 Even though the reality is cruel, we can only pray to God. Survival is like living a battlefield.	生活 life	战场 battlefield	激烈的斗争 fierce struggle

Table 4: Examples of annotated metaphors in CMDAG, separated by source types.

Figure 2: Word Clouds of tenors, vehicles, and adjective and noun components of grounds; the corresponding English word clouds are in the lower row.

ing), utilizing the labels from our annotated dataset. For instance, based off our annotation of the metaphor “他的温柔像海洋” ("his gentleness is like the ocean"), an example prompted for task 1 would be: *when constructing a metaphor with "his gentleness" as the tenor and "ocean" as the vehicle, the ground could be "the ability to accommodate"* (*“包容的能力”*); and similarly an example can be prompted for task 2, with inferring the vehicle from the tenor and the ground.

Chain-of-Thought (CoT) Prompting Using the model’s response from the first prompt, we then ask the model to generate a metaphor from the TENOR-VEHICLE or TENOR-GROUND pairs, based on the GROUND or VEHICLE of its previous CoT generated response.

Clustering To ensure a diverse set of examples, we strategically selected them from distinct clusters generated through various clustering algorithms. The initial clustering methodology leverages the embeddings of the `[CLS]` tokens to produce distinct clusters. Two primary clustering techniques were employed in our study:

1. *Sentence-level Embeddings Clustering:* This method utilizes the embeddings from the `[CLS]` token of each input, and the K-means clustering algorithm is then applied to these embeddings to generate distinct clusters.

2. *Word-level Embeddings Clustering:* Rather than using sentence-level embeddings, this technique takes advantage of word-level embeddings for each token in the input. These embeddings are then subjected to K-means clustering to produce the desired clusters.

5. Experiments

As described in previous sections, we apply the unannotated and annotated versions of CMDAG.

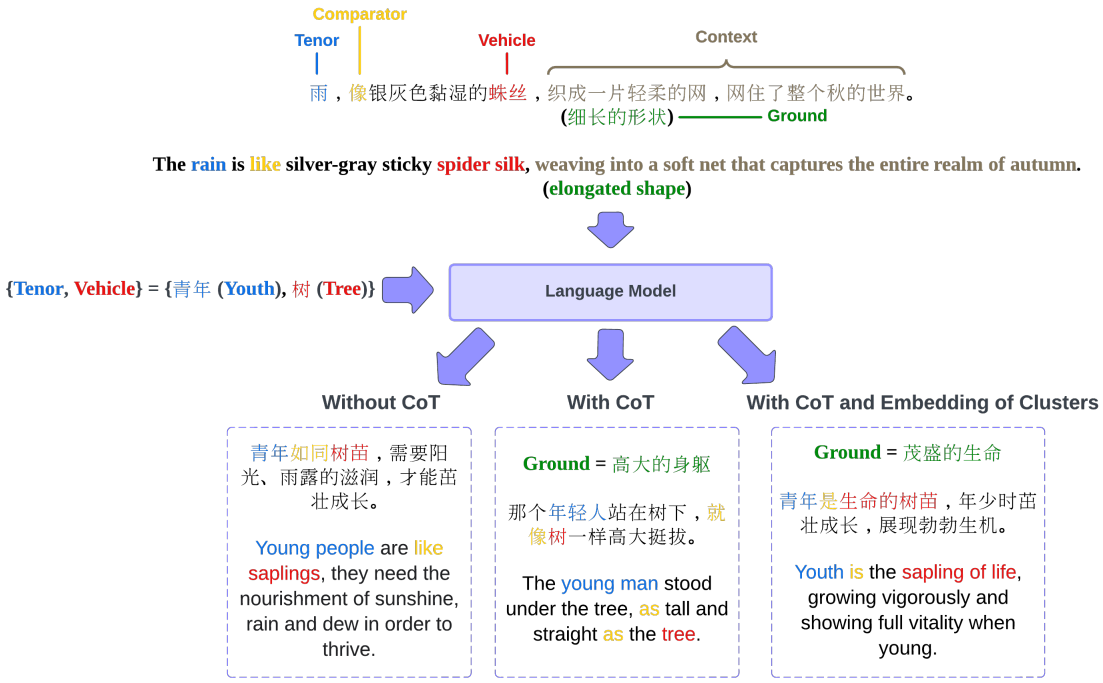

Figure 3: A flowchart that illustrates our experiment with an example of task 1.

5.1. Experimental Setting

The evaluations were consistently carried out across six standardized settings to maintain a uniform benchmark, three for each of our two metaphor generation tasks, across our selected models. For each language model:

- In Setting 0 of Task 1, we prompt the model with TENOR-VEHICLE pairs and for each pair we ask it to generate a corresponding metaphor.
- In Setting 1 of Task 1, we prompt the model with TENOR-VEHICLE pairs, as well as annotated examples selected based on our first clustering method, and for each pair we ask it to generate a corresponding GROUND. We then prompt the model again with the same TENOR-VEHICLE pairs and annotated examples, as well as the inferred GROUND, and for each pair we ask it to generate a corresponding metaphor.
- In Setting 2 of Task 1, we conduct a similar process as in Setting 1, except we select the annotated examples based on our second clustering method.
- In Task 2, we apply a similar procedure of settings as in Task 1, but instead of prompting with TENOR-VEHICLE pairs, we prompt the model and provide annotated examples with TENOR-GROUND pairs, and ask it to infer the corresponding VEHICLE for each pair in Settings 1 and 2.

5.1.1. Models

Chinese metaphor generation is a novel task, we select three general generative models, a Chinese nominal metaphor generation method, and a Chinese metaphor generation model as baselines.

GPT-3.5 GPT-3.5 (OpenAI, 2023b) is a version of OpenAI's Generative Pretrained Transformer series. It is capable of handling a variety of language-processing tasks.

GPT-4.0 GPT-4 (OpenAI, 2023a) is a large-scale, multimodal model capable of accepting both image and text inputs to produce text outputs. It showcases human-level performance on various professional and academic benchmarks.

Belle Belle (Yunjie et al., 2023) is a Chinese LLM (Large Language Model) trained specifically on Chinese data and thus is able to generate precise Chinese metaphoric information.

Baichuan Baichuan (Baichuan, 2023) is a robust 13-billion parameter Chinese AI language model that is open-source and freely available for business and research purposes.

Chinese-alpaca-33B : Chinese-alpaca-33B (Cui et al., 2023) is a state-of-the-art language model that holds a massive 33 billion parameters, specifically designed for Chinese language tasks.

ERNIE : Baidu ERNIE (Research, 2023) is an innovative language model developed by Baidu Research, focusing on understanding and generating text in a more human-like manner.

Model Name	Setting	Clarity	Creativity	Authentic Expression	Final Score
Baichuan	⊙	2.94	2.06	2.36	2.4
Baichuan	◇	2.98	2.09	2.29	2.49
Baichuan	⋆	2.98	2.07	2.20	2.32
Belle	⊙	2.61	1.71	2.18	2.07
Belle	◇	2.83	1.9	2.37	2.33
Belle	⋆	2.97	1.69	2.23	2.17
GPT-4	⊙	2.92	1.64	2.16	2.25
GPT-4	◇	2.96	1.6	2.11	2.21
GPT-4	⋆	2.98	1.66	2.24	2.36
GPT-3.5	⊙	2.99	1.78	2.23	2.21
GPT-3.5	◇	2.99	1.75	2.16	2.25
GPT-3.5	⋆	2.98	1.45	1.94	2.03
Chinese-alpaca-33B	⊙	2.99	1.83	2.14	2.28
Chinese-alpaca-33B	◇	2.97	1.68	2.14	2.11
Chinese-alpaca-33B	⋆	2.99	1.86	2.29	2.20
ERNIE	⊙	2.87	1.86	2.30	2.27
ERNIE	◇	2.97	1.56	2.16	2.27
ERNIE	⋆	2.90	1.73	2.02	2.17

Table 5: The human evaluation results for each model under three settings for taks1. According to the section 5.1, ⊙ is the symbol of Setting 0, ◇ is the symbol of Setting 1 and ⋆ represents the Setting 2.

Model Name	Setting	Clarity	Creativity	Authentic Expression	Final Score
Baichuan	⊙	2.74	1.91	2.30	2.27
Baichuan	◇	2.41	1.89	1.98	2.04
Baichuan	⋆	2.53	1.98	2.00	2.08
Belle	⊙	2.47	1.93	2.29	2.14
Belle	◇	2.54	1.97	2.15	2.22
Belle	⋆	2.56	1.87	2.02	2.09
GPT-4	⊙	2.57	1.83	2.17	2.13
GPT-4	◇	2.48	1.66	2.26	2.12
GPT-4	⋆	2.58	1.62	2.21	2.07
GPT-3.5	⊙	2.60	1.94	2.22	2.18
GPT-3.5	◇	2.36	1.77	2.12	2.02
GPT-3.5	⋆	2.49	1.70	2.05	2.02
ERNIE	⊙	2.23	1.85	2.21	1.98
ERNIE	◇	2.31	1.47	2.14	1.92
ERNIE	⋆	2.28	1.45	2.15	1.87

Table 6: The human evaluation results for each model under three settings for taks2. The symbols for the settings are consistent with those in Table 5.

5.1.2. Evaluation Metrics

Evaluating models' performance on metaphor sentences is extremely challenging because determining the vividness of a metaphor is often intuitive. Many of these tasks cannot be measured by automatic metrics or even be judged by normal crowd workers. To get a more faithful evaluation, we hire expert annotators to judge model predictions. All the annotators conducting the human evaluation have a Master's or Doctor's degree in Chinese Literature, Philology, or Literature. Due to cost, each sample is only analyzed by one annotator. To illustrate the annotators' responsibility, they are allowed to join the project only if their trial annotation results are verified by the authors of CMDAG.

The annotators are asked to rate the output based on whether it accurately and vividly generates the metaphors. We implemented a four aspects rating system for categorizing the quality

Model Name	Setting	Task1	Task2
Belle	⊙	0.112	0.236
Belle	◇	0.12	0.268
Belle	⋆	0.14	0.216
GPT-4	⊙	0.38	0.484
GPT-4	◇	0.448	0.548
GPT-4	⋆	0.448	0.548
GPT-3.5	⊙	0.372	0.384
GPT-3.5	◇	0.392	0.416
GPT-3.5	⋆	0.32	0.368

Table 7: Percentage of model-generated sentences that are reasonable Chinese metaphors. The symbols for the settings are consistent with those in Table 5.

of the models' outputs: **Clarity**, **Creativity**, **Authentic Expression** and **Final Score**. For every criterion, scores range from 1 to 3 points, with 1 being the minimum and 3 being the maximum score.
Clarity: Refers to the degree to which a statement is expressed without ambiguity, ensuring its comprehensibility.

Example: 「眼睛是人心灵的窗户」This phrase, meaning "The eyes are the window to the soul", is unambiguous and clearly expresses the idea that one's eyes can reveal their innermost thoughts and feelings.
Creativity: Indicates the originality of the given statement, differentiating between novel concepts and clichéd ideas.

Example: 「小朋友的脸仿佛是红苹果」This statement, which translates to "The faces of children are like red apples", is straightforward and lacks novelty.
Authentic Expression: Represents the degree to which a statement aligns with expressions that are considered authentic or native-like by the evaluators.

Example: 「心如止水」This idiom, meaning "Heart like still water", is an authentic and native-like expression conveying a sense of inner peace and tranquility.

5.2. Discussion

We propose an analysis of how grounds-based CoT assists LLMs in metaphor generation in Tab. 7. Additionally, we provide expert-level human evaluation results on how different LLMs perform on Task 1 and Task 2 in Tab. 5 and Tab. 6. As supplementary material, we also reveal different human evaluation criteria' relationships in Tb. 8. The experiments of Tab. 7 are conducted on a selected 250-sample test set selected from CMDAG. Only and all reasonable metaphorical sentences of various models and settings are manually evaluated and analyzed in Tab. 5 and Tab. 6.

5.2.1. Grounds-based CoT's Influence

Tab. 7 reveals that Grounds-based CoT can improve the percentage of model-generated sentences that are reasonable Chinese metaphors. Given Tab. 5 and Tab. 6, we notice that LLMs with Grounds-based CoT achieve comparable performance on Task 1 and Task 2, compared with LLMs without Grounds-based CoT. **Nota bene** that LLMs without grounds-based CoT often generate fewer reasonable metaphorical sentences, so their experiment results might slightly benefit from it. We also propose that Grounds-based CoT leads to a slight performance decline, especially in the Creativity and Authentic Expression criteria. An assumption of the observation is that Grounds-based CoT limits LLMs' tendency to explore novel **Vehicle (**喻体**)** and **Ground (**喻意**)**, which is a promising future research direction.

5.2.2. Various LLMs' Performance

Based on Tab. 7, Tab. 5, and Tab. 6, we have two major observations. **First**, since Baichuan performs similarly or even surpasses GPT-4 and GPT-3.5 in Task 1 and Task 2, we point out that LLMs with more Chinese corpora in their pretraining procedure might perform better on Chinese metaphor generation. **Second**, Belle generates much fewer reasonable metaphorical sentences compared to GPT-4 and GPT-3.5. Additionally, GPT-4 and GPT-3.5 cannot always generate reasonable Chinese metaphorical sentences as well. The observations reveal that the Chinese metaphor generation is still an under-explored task, and a larger model size and training corpus can lead to a noticeable performance gain on the task.

5.2.3. Criteria's Relationships

	Clarity	Creativity	Authentic Expression
Task 1	0.28	0.71	0.68
Task 2	0.85	0.72	0.41

Table 8: Pearson Correlation between final score and evaluation criteria.

Based on Tab. 8, we point out that expert-level annotators attach importance to creativity when conducting an evaluation on Chinese metaphor generation. Additionally, compared to the conventional Task 1, expert-level annotators pay more attention to clarity instead of authentic expression. We propose an assumption that human annotators

hold an implicit belief of GROUND in the conventional Task 1 setting which decreases their reliability on clarity. However, writers practically only have TENOR and the features of TENOR, in other words GROUND, in their mind, when they want to write a metaphorical expression. As a result, we point out that future metaphor generation models and benchmarks should pay more attention to the clarity of generated metaphorical sentences.

6. Conclusion

In this paper, we present an annotated Chinese Metaphor Dataset, encompassing approximately 28,000 sentences sourced from a wide array of Chinese literary forms, including poems, prose, and song lyrics. To ensure the precision and uniformity of our annotations, we have developed a thorough set of guidelines. These guidelines are instrumental in aiding annotators in the identification of tenors, vehicles, and grounds. Further more, we design a evaluation method for metaphor sentence generation that leverages a Chain of Thoughts (CoT) framework. Our experimental setup employs open-source multilingual Large Language Models (LLMs), which are tested to underscore the corpus's capability to facilitate the generation of creative and linguistically metaphors. This underscores the significant potential of our dataset to fuel advancements in the understanding and creation of Chinese metaphors.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Max Black et al. 1979. More about metaphor. *Metaphor and thought*, 2:19–41.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. Mermaid: Metaphor generation with symbolism and discriminative decoding.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Laure J End. 1986. Grounds for metaphor comprehension. In *Advances in psychology*, volume 39, pages 327–345. Elsevier.

Shu-Ping Gong. 2003. A corpus-based study on mapping principles of metaphors in politics. In *Proceedings of the ROCLING 2003 Student Workshop*, pages 287–294.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *arXiv preprint arXiv:2310.00752*.

George Lakoff. 1992. The contemporary theory of metaphor. In Andrew Ortony, editor, *Metaphor and Thought (2nd edition)*, chapter 11, pages 202–251. Cambridge University Press, Cambridge.

Yucheng Li, Chenghua Lin, and Frank Geurin. 2022a. Nominal metaphor generation with multitask learning.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022b. Nominal metaphor generation with multitask learning. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 225–235, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Yucheng Li, Shun Wang, Chenghua Lin, and Guerin Frank. 2023. Metaphor detection via explicit basic meanings modelling. *arXiv preprint arXiv:2305.17268*.

Su Lin. 2021. Metaphor and metonymy: Differences in chinese language and culture. *Open Journal of Modern Linguistics*, 11(2):135–139.

Ruibo Liu, Chenyan Jia, Ge Zhang, Ziyu Zhuang, Tony Liu, and Soroush Vosoughi. 2022. Second thoughts are best: Learning to re-align with human values from text edits. *Advances in Neural Information Processing Systems*, 35:181–196.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*.

Zhiqiang Liu, Zuohui Fu, Jie Cao, Gerard de Melo, Yik-Cheung Tam, Cheng Niu, and Jie Zhou. 2019. Rhetorically controlled encoder-decoder for Modern Chinese poetry generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1992–2001, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2023a. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2023b. How do davinci and text davinci-003 differ? Accessed: YYYY-MM-DD.

Baidu Research. 2023. Wenxin: Baidu's advanced language model. Accessed: YYYY-MM-DD.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. 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 6724–6736, Online. Association for Computational Linguistics.

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

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

Zekun Wang, Ge Zhang, Kexin Yang, Ning Shi, Wangchunshu Zhou, Shaochun Hao, Guangzheng Xiong, Yizhi Li, Mong Yuan Sim, Xiuying Chen, et al. 2023. Interactive natural language processing. *arXiv preprint arXiv:2305.13246*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Xiangpeng Wei, Zhengyuan Liu, and Jun Xie. 2023. Fantastic expressions and where to find them: Chinese simile generation with multiple constraints. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 468–486, Toronto, Canada. Association for Computational Linguistics.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Ji Yunjie, Deng Yong, Gong Yan, Peng Yiping, Niu Qiang, Zhang Lei, Ma Baochang, and Li Xiangang. 2023. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. *arXiv preprint arXiv:2303.14742*.

Ge Zhang, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, et al. 2023. Chinese open instruction generalist: A preliminary release. *arXiv preprint arXiv:2304.07987*.

Jiayi Zhang, Zhi Cui, Xiaoqiang Xia, Yalong Guo, Yanran Li, Chen Wei, and Jianwei Cui. 2021. Writing polishment with simile: Task, dataset and a neural approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14383–14392.

Danning Zheng, Ruihua Song, Tianran Hu, Hao Fu, and Jin Zhou. 2020. "love is as complex as math": Metaphor generation system for social chatbot.