A Benchmark Dataset and Instruction Fine-Tuning Methods for Metaphorical Comprehension and Explanation

Senqi Yang[1], Dongyu Zhang[1,2](✉), Wei Guo[1], Mingshuo Pan[1], Haojia Li[1], Liang Yang[3], and Hongfei Lin[3]

[1] School of Software, Dalian University of Technology, Dalian, 116620, China
{ysq1997, guowei1030, mingshuopan, 32417051}@mail.dlut.edu.cn
[2] School of Foreign Languages, Dalian University of Technology, Dalian, 116023, China
zhangdongyu@dlut.edu.cn
[3] School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
{liang, hflin}@dlut.edu.cn

Abstract. Fine-tuning large language models (LLMs) with instruction sets has become an effective method to improve the performance of LLMs. However, current metaphor datasets suffer from issues such as inconsistent annotation methods, insufficient data, limited task diversity, and a narrow genre scope, making it difficult to form a unified instruction set. Additionally, existing fine-tuning strategies have not considered the logical sequence and difficulty differences among metaphor tasks. To address these issues, we have released the "Metaphor Understanding and Generation Instruction Fine-Tuning Dataset" (MetaIFD). This dataset contains 113,384 high-quality entries from over ten genres and supports five metaphor tasks and sentiment analysis tasks. Furthermore, MetaIFD introduces two key annotations: the metaphor identification analysis process and sentiment labels. In this paper, we also propose a three-stage fine-tuning strategy, following the principles of "explicit sentiment tasks preceding implicit sentiment tasks" and "understanding tasks preceding generation tasks," along with a difficulty-based training approach that considers the logical sequence and difficulty differences between tasks. Experimental results show that the fine-tuned model significantly improves performance across all tasks. The dataset and instruction set are publicly available at: https://github.com/DUTIR-YSQ/MetaIFD.

Keywords: Text dataset · Large Language Model · Metaphor · Finetune.

1 Introduction

Metaphor is not only a linguistic technique but also a cognitive method that guides thinking and decision-making. Research shows that metaphors are very common in language, appearing on average once every three sentences [?,22].

According to Conceptual Metaphor Theory (CMT)[13], metaphors are expressed through conceptual mappings between the target and source domains. For example, "Life is a journey" maps the target domain "life" onto the source domain "journey," metaphorically presenting the process, goals, and challenges of life. This mapping is not only a linguistic phenomenon but also a cognitive process that helps people connect abstract concepts to concrete experiences, leading to a better understanding of these concepts[24].

With the development of large-scale language models (LLMs) such as GPT-4O [1], LLaMA [8], and Qwen [2], these models often confuse the target domain with the source domain when understanding metaphors, sometimes performing even worse than random baselines [24]. However, existing one-shot fine-tuning methods are unable to handle instructions of varying complexity [17]. Although phased fine-tuning methods [9] account for instruction difficulty, they overlook the inherent sequential dependencies in metaphor tasks.

Existing metaphor datasets have several issues that hinder the construction of high-quality instruction sets. First, the annotation methods are inconsistent [4,20,21,23,26,25], making it difficult to directly integrate the datasets. The lack of a metaphor identification process, the relatively small size of the current datasets, the limited range of supported task types, the lack of genre diversity, and the absence of emotional polarity [4,20,21,23,25] all contribute to these challenges.

To address the aforementioned issues, this paper collects 13 metaphor datasets, which, after uniform formatting and deduplication, result in 113,384 entries covering five metaphor understanding and generation tasks. Additionally, two new labels were added to the processed dataset. First, based on the MIP theory, the metaphor identification process was generated using GPT-4O. Second, emotional polarity labels were added to the sentences using a chain-of-thought (COT) approach, and with the aid of RAG (retrieval-augmented generation) technology, an emotional lexicon was retrieved, with emotional labels generated through GPT-4O.

Considering the logical relationships between metaphor tasks, this paper proposes a progressive fine-tuning strategy—three-phase fine-tuning, which comprehensively considers the sequential dependencies and difficulty differences between metaphor understanding and generation tasks. Specifically, the first phase focuses on explicit emotional tasks, while subsequent phases progressively address metaphor tasks.

2 Related Work

2.1 Metaphor Instruction Fine-tuning Datase

Current metaphor datasets often lack sentiment polarity annotations and the reasoning process for metaphor identification [4,5,19], which hinders the effectiveness of LLMs in metaphor understanding and sentiment analysis tasks. Most of these datasets contain only a few thousand entries, which is inadequate for fine-tuning LLMs. Furthermore, the range of tasks supported by existing datasets is relatively limited. For instance, the VUA, TSV, LCC, GAN-KEVIN, and RE-MAnn datasets focus exclusively on metaphor identification; the MOH and Bizzoni datasets cover only two tasks; even the MUHCH dataset, which supports the most tasks, accommodates only three.

Additionally, metaphor datasets tend to lack genre diversity, with many being concentrated in specific domains or text types, thereby restricting their applicability to broader contexts. For example, the TSV and Bizzoni datasets are limited to public resources, the Katz dataset is focused solely on literary texts, and the MOH dataset is entirely reliant on WordNet. This lack of diversity significantly undermines the generalizability of these datasets, thereby limiting the performance of LLMs in a variety of real-world applications.

To address these issues, this paper collects 11 metaphor datasets, including TroFi [4], VUA [23], TSV [25], MOH [21], Bizzoni [5], LCC [16], Katz [18], GAN-kevin [20], Reimann [19], figqa [14], and MUHCH [24]. These datasets are processed into a unified format and duplicates are removed. The final dataset consists of 113,384 entries, covering five metaphor understanding and generation tasks. The data comes from a wide range of sources, including news reports, literary works, social media texts, dialogue data, etc., ensuring the diversity and representativeness of the dataset.

Additionally, two new types of labels have been added to the dataset: the analysis process for metaphor identification and sentiment polarity. After adding sentiment labels, the dataset now supports sentiment analysis tasks.

2.2 Instruction Fine-Tuning

The current mainstream instruction fine-tuning methods typically use one-off fine-tuning (One-off IFT), which involves training the entire instruction dataset in a single pass [27,6,3,11]. Although this approach emphasizes the quality and diversity of the dataset, it overlooks the complexity of the instruction set, making it difficult for the model to deeply understand and effectively execute diverse instruction tasks.

Pang et al. [9] proposed a staged fine-tuning strategy, dividing the dataset into multiple subsets based on instruction difficulty and training them from simple to complex. While this method is based on instruction difficulty grading, it does not fully consider the sequence and inherent logical relationships between tasks. For example, metaphor identification should serve as the foundation for metaphor generation, and metaphor tasks, as part of implicit sentiment computation, should follow a progressive learning strategy: starting with explicit sentiment recognition and gradually moving to implicit sentiment identification.

To address these issues, this paper proposes a three-phase fine-tuning strategy. The core idea is to first fine-tune the model for sentiment analysis tasks to master explicit sentiment recognition, and then further fine-tune the model for metaphor tasks, gradually improving its performance in metaphor understanding and generation tasks.

3 MetaIFD

Table 1. Metaphor and Literal Dataset Statistics.

Dataset	Metaphor	Literal	Total
TroFi	2,110	1,627	3,737
VUA	1,194	13,319	14,513
TSV	211	211	422
MOH	1,204	399	1,603
Bizzoni	187	0	187
LCC	18,235	18,133	36,368
Katz	598	0	598
GAN-kevin	45,164	0	45,164
Reimann	1,338	0	1,338
figqa	5,384	0	5,384
MUHCH	4,070	0	4,070
Total	**79,695**	**33,689**	**113,384**

3.1 Data Source

The MetaIFD dataset is sourced from 11 different datasets, each focusing on different aspects of metaphorical and literal instances. We collected over 500,000 English entries. The data comes from various genres, including news, fiction, academic texts, dialogues, religious websites, and more. After deduplication and data cleaning, a total of 113,384 entries were retained. Table 1 provides detailed statistical information.

Table 2. MetaIFD Data Example.

Label	Value
Sentence	The faculty meeting was a real war.
Metaphorical or literal	Metaphorical
Metaphor word	war
Interpretation order	1. The faculty meeting was a peaceful gathering. 2. At the faculty meeting, we talked about real war. 3. The faculty meeting was ridiculous. 4. The faculty meeting was highly aggressive.
Interpretation	Using the characteristics of "war" to describe the intensity of the meeting and the challenging nature of the situation.
Metaphor novelty rating	0.23
Metaphor_MIP	According to MIP theory, "The faculty meeting was a real war" is a metaphor because the literal meaning of "war" (actual combat) differs from its contextual meaning (an intense and confrontational meeting).
Sentiment	-1

3.2 Unified Data Format

Metaphorical or literal: Indicates whether the sentence conveys a literal or metaphorical meaning. **Metaphor word**: Highlights specific words or phrases in the sentence that carry metaphorical significance. **Interpretation order**: Rates the quality of interpretations for metaphorical sentences on a scale of 1 to 4. Higher scores indicate interpretations that more accurately capture the intended metaphorical meaning. **Interpretation**: Provides a detailed explanation of the metaphor, specifying its meaning and its relationship to the context. **Metaphor novelty rating**: Assesses the novelty of the metaphorical word on a scale from 1 to -1. Scores closer to 1 suggest higher novelty, while scores closer to -1 indicate lower novelty.

3.3 Dataset Annotation

Metaphor_MIP Metaphor_MIP identifies metaphors by analyzing and comparing the basic and contextual meanings of metaphorical words in a sentence, as shown in Table 2. MIP systematically identifies metaphors by comparing the basic and contextual meanings of words. In this study, GPT-4O is used to annotate the Metaphor_MIP labels in the dataset and generate metaphor identification analyses based on MIP theory. Metaphor_MIP plays a crucial role in fine-tuning for metaphor recognition, helping the model deeply understand the logic of metaphors and significantly improving its comprehension, going beyond superficial reliance on annotation results. The length distribution of the Metaphor_MIP field is concentrated between 0 and 50, with frequency decreasing as length increases, though some longer texts still appear frequently.

Sentiment Metaphorical language often conveys emotions [12,15,7], so performing sentiment recognition helps LLMs better understand metaphors. In this study, sentences are annotated with five levels of sentiment labels: -2 (strongly negative), -1 (negative), 0 (neutral), 1 (positive), and 2 (strongly positive). Additionally, a COT-based sentiment inference method is designed, combined with RAG technology to retrieve dictionaries and infer sentiment orientation. Statistical analysis of sentiment label distribution shows that negative labels (-2 and -1) are the most frequent.

3.4 Quality Control

We validated the sentiment labels of the dataset with the participation of three researchers: two master's students in computational linguistics and one PhD student. A random sample of 600 instances was selected, and the Kappa values for the three-class (Sentiment(3)) and five-class (Sentiment(5)) classification schemes were evaluated. For Sentiment(3), the average Kappa value was 78.13%. For Sentiment(5), the average Kappa value was 53.31%. According to the standards of Fleiss et al. [10], the three-class classification showed high consistency, while the five-class classification showed moderate consistency. Additionally, the Metaphor_MIP content was validated with an average of 81.49%, demonstrating high consistency.

4 Three Stage Finetune

Task Definition: Sentiment Analysis: Analyzing the sentiment tendency of a sentence and determining whether it is very positive, positive, neutral, negative, or very negative. **Metaphor Identification**: Determining whether a sentence contains metaphorical expressions. **Metaphor Word Identification**: After confirming that a sentence contains a metaphor, further identifying the specific metaphorical words or phrases within the sentence. **Metaphor Novelty Scoring**: Evaluating the novelty of metaphorical words in a sentence. The score ranges from -1 to 1, with higher scores indicating greater novelty. **Metaphor Explanation Rating**: Evaluating the quality of metaphor explanations, based on whether the explanation effectively conveys the core meaning of the metaphor. The score ranges from 1 to 4, with higher scores indicating better quality. **Paraphrase Generation**: Generating accurate explanations or paraphrases of metaphorical expressions, clearly articulating the deeper meaning of the metaphor.

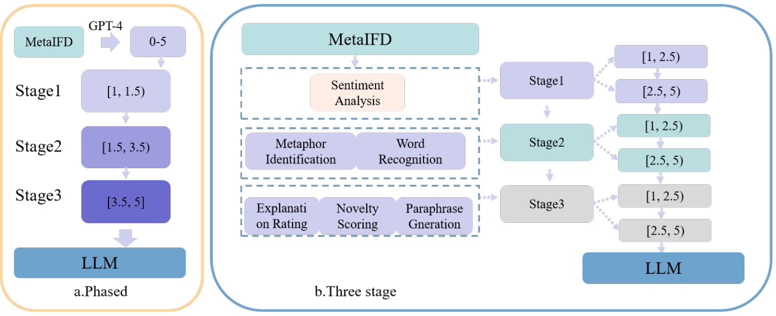

Fig. 1. Phased and Three-stage: The Phased method uses GPT-4 to score and categorize instructions by task difficulty for fine-tuning; the Three-stage method fine-tunes in three stages based on task difficulty.

Three stage Finetune: The three-stage fine-tuning process consists of three phases: The first phase is sentiment analysis, as metaphorical language often contains rich emotional information [12,15,7]. Sentiment analysis helps LLMs better understand metaphors and lays a solid foundation for subsequent tasks. Explicit emotions are easier to recognize, so performing sentiment analysis before metaphor comprehension and generation aligns with the progressive learning strategy from simpler to more complex tasks.

The second phase includes metaphor identification and metaphorical word identification, while the third phase focuses on metaphor explanation scoring.

This study argues that relying solely on the difficulty of the instruction set for classification does not fully capture the characteristics of metaphor-related tasks. Metaphor tasks have strong sequential dependencies, and there are significant difficulty differences between tasks. Therefore, the design of fine-tuning phases should be based on the inherent difficulty of the tasks themselves, rather than just instruction difficulty. Hence, this study adopts a progressive learning strategy following the principle of "explicit before implicit, simple before complex."

To improve fine-tuning effectiveness, each phase is divided into two difficulty intervals: (1, 2.5) and (2.5, 5). Statistical analysis shows that data in the [1, 1.5) interval accounts for only 2.76%, which is insufficient to support phased fine-tuning. Additionally, the Phased method is not suitable for metaphor tasks. Therefore, the range is extended to [1, 2.5), increasing the data proportion to 14.12%, significantly broadening the coverage and including simpler instances from more complex tasks. Another part of the data is trained in the [2.5, 5] interval, providing sufficient support for fine-tuning and promoting more efficient phased, difficulty-based learning.

5 Experiment

5.1 Parameter Settings

During the training process, we used a single A100-40G VRAM GPU with a batch size of 48, gradient accumulation steps of 4, a learning rate of 1e-4, model saving enabled on each node (True), and gradient checkpointing enabled. In the LoRA section, the rank was set to 8, LoRA alpha to 32, and LoRA dropout to 0.1.

5.2 Baseline

The fine-tuning models used in this paper are based on Llama 3.1-8B [8] and include three baseline methods: zero-shot, One-Off, and phased fine-tuning methods [9]. Zero-shot refers to the un-tuned Llama 3.1-8B model. The One-Off method involves inputting all data to the LLM at once. The phased fine-tuning method first uses GPT-4 to score 113,384 instructions, with scores ranging from 0 to 5, where a score closer to 0 indicates easier tasks, and a score closer to 5 indicates more difficult tasks. After scoring, the data is divided into three categories: [1, 1.5), [1.5, 3.5), and [3.5, 5]. Fine-tuning is then performed in three stages: In the first stage, instructions with scores in the range of [1, 1.5) are fine-tuned. In the second stage, the model from the first stage is loaded and fine-tuned on instructions in the range of [1.5, 3.5). In the third stage, the model from the first stage is loaded again, and fine-tuning is performed on instructions in the range of [3.5, 5).

5.3 Comparative Experiment

Table 3 presents the experimental results of three fine-tuning methods across six tasks. The evaluation metrics are as follows:

Table 3. Fine-tuning results. The performance of four methods across six tasks. The best results are highlighted in bold.

Task	**Zero-shot**	**One-Off**	**Phased**	**Three Stages**
Metaphor (F1)	60.80	0.732	0.752	**0.830**
Metaphor Words (F1)	36.04	42.11	**0.651**	0.634
Explanation Rating (MSE)	1.67	1.505	1.464	**0.402**
Novelty Scoring (MSE)	0.055	0.0497	0.054	**0.044**
Paraphrase Generation (GPT-4O)	0.7370	0.7490	0.7960	**1.3700**
Sentiment Analysis (F1)	0.409	0.4600	0.472	**0.618**

For metaphor recognition, sentiment analysis, and metaphor word identification tasks, the F1 score is used as the evaluation metric. In the metaphor word identification task, singular/plural forms and tense differences are ignored; as long as the identified word matches the target word, it is considered correct. Explanation quality and novelty are evaluated using mean squared error (MSE), which measures the deviation in output quality. For the paraphrase generation task, GPT-4O is used for evaluation. GPT-4O compares the paraphrase with the original text and assigns scores based on the following criteria: 3 points: almost identical, with the same core meaning and expression; 2 points: core meaning is consistent, but with slight differences in expression or details; 1 point: partially similar, but with significant deviation in core information; 0 points: very different, with core information completely changed.

The experimental results show that, compared to the zero-shot performance, the fine-tuned models achieved varying degrees of improvement across all tasks, validating the effectiveness of MetaIFD. The three-stage fine-tuning method significantly outperformed the phased fine-tuning method and also surpassed the one-off fine-tuning method. This indicates that adopting a staged training strategy based on the difficulty of metaphor tasks and the logical relationships between tasks, followed by further refinement of task instructions according to their difficulty, is an effective approach to improving model performance.

Error Analysis: In the metaphor word identification task, the three-stage fine-tuning method slightly underperformed compared to the phased fine-tuning. This may be because the difficulty of the metaphor word identification task is concentrated in the 2.5-5 range, while the data in the 1-2.5 interval is relatively sparse, leading to insufficient training in the initial stages and thus affecting overall performance. In the paraphrase generation task, the three-stage fine-tuning method significantly outperformed the other methods. Other fine-tuning methods often generated paraphrases that were too lengthy and deviated significantly from the original, resulting in lower scores.

5.4 Ablation Experiment

To validate the effectiveness of the "explicit-to-implicit progressive learning" approach and to demonstrate the contribution of sentiment analysis to metaphor

Table 4. Ablation experiment results.

Task	**W/o sentiment**	**With sentiment**
Metaphor (F1)	0.758	**0.830**
Metaphor Words (F1)	0.595	**0.634**
Explanation Rating (MSE)	0.487	**0.402**
Novelty Scoring (MSE)	0.0807	**0.044**
Paraphrase Generation (MT-Bench)	1.2820	**1.3700**
Sentiment Analysis (F1)	0.412	**0.618**

recognition tasks, ablation experiments were conducted. In these experiments, "W/o sentiment" refers to the method that excludes the first-stage sentiment analysis training. The results, as shown in Table 4, indicate that sentiment analysis significantly enhances the performance of metaphor understanding and generation tasks, helping LLMs gain a more comprehensive mastery of metaphors, which is consistent with previous findings.

6 Conclusion

To address issues such as inconsistent annotations, data scarcity, and limited task types in existing metaphor understanding and generation instruction sets, this study constructed the MetaIFD fine-tuning dataset, which contains 113,384 records. Additionally, a three-stage fine-tuning strategy based on task difficulty was proposed. Experimental results demonstrated that this approach significantly improved model performance on metaphor-related tasks and sentiment analysis tasks, thereby validating the effectiveness of both MetaIFD and the fine-tuning strategy. Future optimization of task performance could be achieved through alignment techniques for LLMs.

Acknowledgments. This work is supported by National Language Commission Key Program (No.ZDI145-80); Fundamental Research Funds for the Central Universities (No.DUT24RW213); Research Project on Economic and Social Development of the Liaoning Provincial Federation of Social Sciences (No.2025lslybkt-064); and NSFC Programs (No.62076051).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv:2303.08774 (2023)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv:2309.16609 (2023)
3. Bin, Y., Shi, W., Ding, Y., et al.: Gallerygpt: Analyzing paintings with large multimodal models. ACM MM 2024 pp. 7734–7743 (2024)

4. Birke, J., Sarkar, A.: A clustering approach for nearly unsupervised recognition of nonliteral language. In: EACL 2006. pp. 329–336 (2006)
5. Bizzoni, Y., Lappin, S.: Predicting human metaphor paraphrase judgments with deep neural networks. In: FigLang:. pp. 45–55 (2018)
6. Chen, L., Li, S., Yan, J., et al.: Alpagasus: Training a better alpaca with fewer data. arXiv:2307.08701 (2023)
7. Citron, F., Goldberg, A.: Metaphorical sentences are more emotionally engaging than their literal counterparts. Journal of Cognitive Neuroscience **26**(11), 2585–2595 (2014). `https://doi.org/10.1162/jocn_a_00654`
8. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv:2407.21783 (2024)
9. Dubois, Y., Li, C., Taori, R., et al.: Alpacafarm: A simulation framework for methods that learn from human feedback. NeurIPS **36** (2024)
10. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5), 378–382 (1971)
11. Hayashi, K., Sakai, Y., Kamigaito, H., et al.: Towards artwork explanation in large-scale vision language models. ACL 2024 pp. 705–729 (2024)
12. Kövecses, Z.: Metaphor and emotion: Language, culture, and body in human feeling (2003)
13. Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago Press (1980)
14. Liu, E., Cui, C., Zheng, K., Neubig, G.: Testing the ability of language models to interpret figurative language. arXiv:2204.12632 (2022)
15. Mohammad, S., Shutova, E., Turney, P.: Metaphor as a medium for emotion: An empirical study. SemEval 2024 pp. 23–33 (2016)
16. Mohler, M., Brunson, M., Rink, B., Tomlinson, M.: Introducing the lcc metaphor datasets. In: LREC 2016. pp. 4221–4227 (2016)
17. Pang, W., Zhou, C., Zhou, X.H., Wang, X.: Phased instruction fine-tuning for large language models. Findings of the Association for Computational Linguistics: ACL 2024 pp. 5735–5748 (2024)
18. Prystawski, B., Thibodeau, P., Potts, C., Goodman, N.D.: Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. arXiv:2209.08141 (2022)
19. Reimann, S., Scheffler, T.: Metaphors in online religious communication: A detailed dataset and cross-genre metaphor detection. In: LREC-COLING 2024. pp. 11236–11246 (2024)
20. Shou, X., Huang, X., Xi, W.: Conceptual metaphor theory guides gans for generating metaphors and interpretations. IEEE Access (2024)
21. Shutova, E., Kiela, D., Maillard, J.: Black holes and white rabbits: Metaphor identification with visual features. In: NAACL 2016. pp. 160–170 (2016)
22. Shutova, E.V.: Computational approaches to figurative language. Tech. rep., University of Cambridge, Computer Laboratory (2011)
23. Steen, G.J., Dorst, A.G., Herrmann, J.B., Kaal, A.A., Krennmayr, T.: Metaphor in usage (2010)
24. Tong, X., Choenni, R., Lewis, M., Shutova, E.: Metaphor understanding challenge dataset for llms. arXiv:2403.11810 (2024)
25. Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., Dyer, C.: Metaphor detection with cross-lingual model transfer. In: ACL 2014. pp. 248–258 (2014)
26. Zhang, D., Zhang, M., Zhang, H., Yang, L., Lin, H.: Multimet: A multimodal dataset for metaphor understanding. In: AC. pp. 3214–3225 (2021)
27. Zhou, C., et al.: Lima: Less is more for alignment. NeurIPS **36** (2024)