CTYUN-AI at SemEval-2025 Task 1: Learning to Rank for Idiomatic Expressions

Yuming Fan and **Dongming Yang** * and **Zefeng Cai** and **Binghuai Lin**
fanym@chinatelecom.cn, yangdongming@pku.edu.cn,
caizf@chinatelecon.cn, linbinghuai@gmail.com
China Telecom Cloud Technology Co., Ltd

Abstract

This paper presents our solution for SemEval-2025 Task 1: Learning to Rank Idiomatic Expressions, which addresses the challenge of ranking visual representations for figurative language understanding. We propose a multimodal approach that combines textual context with image caption analysis through systematic data augmentation and model fine-tuning. Our method includes three main components: (1) an option-shuffling strategy to eliminate positional bias in ranking tasks, (2) lexical perturbation through synonym replacement and back-translation to enhance linguistic diversity, and (3) parameter-efficient fine-tuning of large language models optimized for cross-modal ranking. The system achieved first place in Portuguese (Top-1 Acc: 0.92, DCG: 3.43) and second place in English (Top-1 Acc: 0.87, DCG: 3.51) on the CodaBench leaderboard. Through extensive experimentation with models ranging from 7B to 72B parameters, we demonstrate that mid-sized 32B models achieve optimal performance by balancing capacity and trainability. Our analysis reveals that while larger models (72B) suffer from overfitting and optimization challenges, traditional knowledge distillation approaches using GPT-4 prove ineffective for this task. The results highlight the importance of controlled data augmentation and parameter scaling for idiomatic representation learning, providing valuable insights for future work in multimodal figurative language processing.

1 Introduction

Idiomatic expressions are a fundamental component of natural language and often pose challenges to human interpreters and computational models. Unlike literal expressions, idioms convey meanings that are not directly inferred from the individual words, but are instead shaped by cultural and contextual usage. These expressions are essential for natural language understanding, influencing tasks such as sentiment analysis, machine translation, and automated summarization. However, despite significant advances in large-scale language models (LLMs), understanding and accurately interpreting idioms remains a key challenge in NLP.

The AdMIRe (Aesthetic Multi-modal Idiomatic Representation) task(Pickard et al., 2025) was introduced to address these challenges by combining textual and visual information to better represent idiomatic expressions. This multimodal approach aims to move beyond traditional text-only models, which often struggle with the figurative meanings of idioms. Through the use of images alongside context sentences, AdMIRe seeks to improve model comprehension by providing a richer, more nuanced understanding of idiomatic expressions.

In this paper, we present our approach to Subtask A - Static Images, where we were tasked with ranking a set of images based on their ability to represent the meaning of a given idiomatic expression in a specific context. We participated in the competition in both English and Portuguese, achieving notable results: first place in Portuguese with a score of 0.93 and second place in English with a score of 0.86. Our approach leverages state-of-the-art language models that integrate textual cues, offering an improved representation of idiomatic expressions.

This paper outlines our methodology for tackling the task, discusses the challenges we encountered, and provides insight into how the integration of visual information can significantly enhance the performance of language models in understanding figurative language.

2 Related Work

Idiomatic expressions are a key component of natural language, posing significant challenges for both human interpreters and computational mod-

[*] Corresponding Author.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 16–19
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

els. Early research highlighted the cognitive difficulty of processing idioms, with Lakoff and Johnson(Lakoff and Johnson, 1980) emphasizing that idioms often carry meanings beyond their literal interpretations.

Although previous tasks have explored how language models represent idioms, Boisson (Boisson et al., 2023) argue that artifacts in these datasets may enable models to perform well on idiomaticity detection without producing high-quality semantic representations.

Traditional NLP models struggled with idiomaticity due to their reliance on literal word meanings, but recent advancements in deep learning have improved idiom detection. Models like BERT(Devlin et al., 2019) and GPT-3 have shown progress in leveraging large-scale contextual embeddings. Currently, generative models in the realm of NLP, exemplified by the GPT series(Brown et al., 2020; Bai et al., 2023; Yang et al., 2023; Wang et al., 2023; Y et al., 2024c,b,a), have shown remarkable abilities in interpreting and producing natural language.

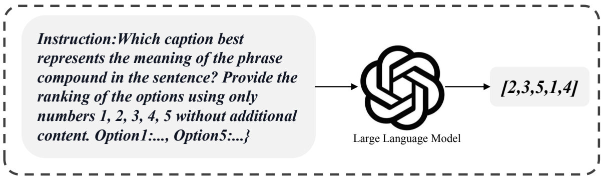
More recently, multimodal approaches have gained attention, integrating visual information to enhance understanding of idioms. AdMIRe demonstrated that combining text and images can significantly improve idiomatic representation, suggesting that multimodal models may offer a promising direction for future research.

3 Method

3.1 Preprocessing

During the data pre-processing stage, we first processed each input record by extracting the idiomatic expressions, contextual sentences, and descriptions and names of five images, constructing input-output pairs from the image description data. For each record, we extractedmpound words, sentences, image captions, and image names formulated an English prompt. The prompt asks: *Which caption best represents the meaning of the phrase compound in the sentence? Provide the ranking of the options using only numbers 1, 2, 3, 4, 5 without additional content. Option1:... Option5*, as shown in fig. 1. Using this data, we trained a large language model (LLM) to perform the ranking task.

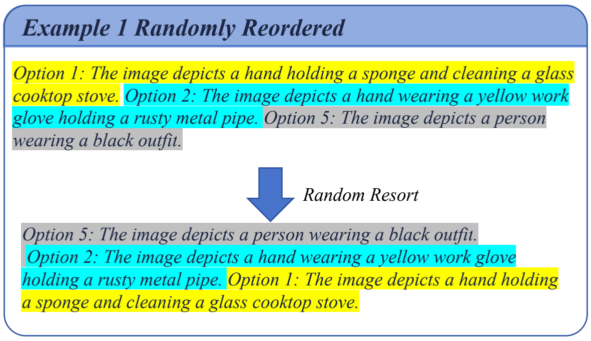
In the testing phase, we applied the trained model to the test set for inference, prompting the LLM to generate a context-based ranking of the five image captions. The resulting ranking, represented by numbers from 1 to 5, was then mapped to the corresponding image names and saved as the final ordered output.



Instruction:Which caption best represents the meaning of the phrase compound in the sentence? Provide the ranking of the options using only numbers 1, 2, 3, 4, 5 without additional content. Option1:..., Option5:...}

Figure 1: Prompt Construction.

However, relying solely on the original data may lead to model overfitting to specific linguistic expressions, limiting its generalization capability. To mitigate this issue, we introduced a series of data augmentation strategies to enhance model robustness and adaptability.



Example 1 Randomly Reordered

Option 1: The image depicts a hand holding a sponge and cleaning a glass cooktop stove. Option 2: The image depicts a hand wearing a yellow work glove holding a rusty metal pipe. Option 5: The image depicts a person wearing a black outfit.

Option 5: The image depicts a person wearing a black outfit. Option 2: The image depicts a hand wearing a yellow work glove holding a rusty metal pipe. Option 1: The image depicts a hand holding a sponge and cleaning a glass cooktop stove.

Figure 2: Randomly reordered method.

3.2 Enhancing Ranking Diversity

To prevent the model from developing a dependency on fixed option positions and improve its generalization in ranking, we applied an option-shuffling strategy to augment the dataset. Specifically, we randomly reordered Options 1-5 while simultaneously adjusting the expected_order field to reflect the new arrangement, as shown in Fig. 2. This process reduces the model's reliance on positional biases and encourages it to focus on the actual content of the options rather than learning patterns from their fixed order.

3.3 Data Self-Augment

Furthermore, to enhance model robustness and enrich data diversity, we perform lexical perturbations in the option texts. We randomly selected words from each input-output pair and replaced them with synonyms, introducing minor variations in the image captions while preserving their core semantics. Additionally, we employed back-translation,

Text Only - Portuguese				
CodaBench Username	Top 1 Acc.	DCG Score	Top 1 Acc. (Extended)	DCG Score (Extended)
CTYUN-AI	0.92	3.43	0.56	2.97
artrsousa	0.85	3.27	0.44	2.78
GPT4	0.6	3.06	-	-
Text Only - English				
CodaBench Username	Top 1 Acc.	DCG Score	Top 1 Acc. (Extended)	DCG Score (Extended)
dd101bb	0.93	3.52	0.83	3.43
CTYUN-AI	0.87	3.51	0.64	3.10
GPT4	0.7	3.17	-	-
phuongnm	0.67	3.04	0.51	2.86
dadonapo97	0.67	3.07	0.59	3.04
artrsousa	0.53	2.82	0.51	2.86
wiepet	0.47	2.82	0.54	3.04
gladysflacks	0.40	2.61	0.39	2.69
arash3908	0.27	2.41	0.20	2.38

Table 1: CodaBench Evaluation Results for Portuguese and English

where captions were translated into other languages (e.g., Chinese) and then translated back into English. This approach introduces linguistic variations, allowing the model to better adapt to different paraphrases and reducing the risk of overfitting to specific expressions.

4 Experiment Results

We conducted an evaluation of Portuguese and English text only data on the CodaBench platform, as presented in Table 1. The primary evaluation metrics were Top-1 accuracy and DCG score, with additional extended criteria also considered. For the Portuguese dataset, the CTYUN-AI system achieved the highest performance, achieving a Top-1 accuracy of 0.92 and a DCG score of 3.43 in the base test set. In the English setting, CTYUN-AI ranked second, with a Top-1 accuracy of 0.87 and a DCG score of 3.51, showcasing its strong competitive edge. Moreover, under the extended evaluation criteria, CTYUN-AI scored 0.56/2.97 for Portuguese and 0.64/3.10 for English, further reinforcing its robustness and stability. These results underscore the significant advantages of our approach in text-processing tasks. Furthermore, we performed a ranking using GPT-4 on the task data, with scores of 0.6 and 0.7 for English and Portuguese, respectively, which were lower than those achieved by our proposed method.

We employed the Qwen2.5(Bai et al., 2023) model series as the backbone and trained our models using the dataset constructed in the Method section. Specifically, we conducted training and inference using four Ascend-910B nodes, each equipped with eight GPUs. The learning rate was set to 5e-6, the gradient accumulation steps were configured as 8, and the models were trained for a total of five epochs. We experimented with models of different parameter scales, as summarized in Table 2.

Model Size	Top-1 Acc. (PT)	DCG Score (PT)
7B	0.70	3.06
14B	0.67	3.14
32B	**0.92**	**3.61**
72B	0.87	3.42

Table 2: Performance of Different Model Sizes on Portuguese Data

4.1 Unsuccessful Attempts

Larger Models and Parameter Scaling: We experimented with models of different parameter sizes, including the Qwen2.5(Bai et al., 2023) model series, ranging from 7B to 72B parameters. While the 72B model had a significantly larger capacity, it did not outperform the 32B model. We hypothesize that this is due to an optimal balance between parameter size and dataset scale, allowing the model to learn complex patterns effectively while avoiding excessive optimization challenges. In contrast, the 7B and 14B models likely lacked sufficient parameters to fully capture the intricate relationships in the input data, thereby limiting their performance. Meanwhile, although the 72B model featured a larger parameter size, it did not outperform the 32B model. We attribute this to two potential factors: first, larger models tend to overfit when trained on a limited dataset, resulting in reduced generalization ability. Second, the computational overhead of training and inference with the 72B model was significantly higher, which may have constrained the batch size and negatively impacted the stability of the gradient.

Leveraging GPT-4 for Data Augmentation and Knowledge Distillation: We initially intended to leverage GPT-4 to augment our dataset and distill its capabilities for improved performance. However, GPT-4's performance in this context was suboptimal, likely due to its inherent limitations when applied to this specific task. This was particularly disappointing given the recent surge in interest around knowledge distillation techniques (e.g., DS-R1(DeepSeek-AI and et al., 2025)) for transferring model knowledge. Despite these efforts, GPT-4 did not provide the anticipated improvements, and we decided to focus on optimizing the core model instead.

These explorations underscore the challenges of scaling up the model parameters and using external models such as GPT-4 for distillation, which, although promising in some contexts, did not yield the expected benefits for this particular task.

5 Conclusion

In this paper, we present our approach to SemEval-2025 Task 1, focusing on ranking idiomatic expressions using a multimodal framework. By integrating textual and visual information, along with data augmentation and fine-tuning, we achieved strong results, securing first place in Portuguese and second place in English on the CodaBench platform. Our approach demonstrated improved understanding of idiomatic expressions and better generalization. Although experiments with larger models and GPT-4 for knowledge distillation were less effective, they provided valuable information. This work highlights the potential of multimodal models in enhancing figurative language processing, and we plan to refine these methods further in future work.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, and et al. 2023. Qwen technical report. *arXiv:2309.16609*.

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6581–6590, Singapore. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and et al. 2020. Language models are few-shot learners.

DeepSeek-AI and Daya Guo et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

G. Lakoff and M. Johnson. 1980. The metaphorical structure of the human conceptual system. *Cognitive Science*, 4(2):195–208.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, Carolina Scarton, and Marco Idiart. 2025. Semeval-2025 task 1: Admire - advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*, Vienna, Austria. IEEE, Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators.

Fan Y, Yang D, Zhang J, et al. 2024a. Fake-gpt: Detecting fake image via large language model. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 122–136, Singapore. Springer Nature Singapore.

Fan Y, Yang D, and Cao L. 2024b. Ctyun-ai@ smm4h-2024: Knowledge extension makes expert models. In *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 5–9.

Fan Y, Yang D, and He X. 2024c. Ctyun-ai at semeval-2024 task 7: Boosting numerical understanding with limited data through effective data alignment. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 47–52.

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