JNLP at SemEval-2025 Task 1: Multimodal Idiomaticity Representation with Large Language Models

Blake Matheny[1], Phuong Minh Nguyen[2,1], Minh Le Nguyen[1]

[1]Japan Advanced Institute of Science and Technology
[2]ROIS-DS Center for Juris-Informatics, NII, Tokyo, Japan
{matheny.blake,phuongnm,nguyenml}@jaist.ac.jp
Correspondence: matheny.blake@jaist.ac.jp

Abstract

Idioms and figurative language are nuanced linguistic phenomena that transport semanticity and culture via non-compositional multi-word expressions. This type of figurative language remains difficult for small and large language models to handle. For language models to become more valuable as translators and even companions, natural language must be understood and generated. A portion of this is idiomatic and figurative language. Various attempts have been made to identify idiomaticity in text. The approach presented in this paper represents an intuitive attempt to accurately address Task 1: AdMIRe Subtask A to correctly order a series of images and captions by concatenating the image captions as a sequence. The methods employ the reliability of a pre-trained vision and language model for the image-type task and a large language model with instruction fine-tuning for a causal language model approach to handle the caption portion of the task. The models chosen for development in the pipeline were based on their respective reliability in captioning and instruction fine-tuning.

1 Introduction

The idiomaticity of a multi-word expression has been traditionally difficult for many language models, due to the non-compositional nature of this type of figurative language. When interpreting meaning from a sequence that may contain an idiom, a language model must chunk accurately and identify the possible multi-word expression with phrase analysis. For general pre-trained models without training or fine-tuning for this downstream task, the contextual embeddings tend to interfere with this detection. The embeddings must change internally or externally to accommodate for the lack of "sense" the phrase would make if treated as a literal and compositional phrase. Disregarding tokenization, consider a word by word embedding of this sentence, "They became relaxed when they saw that the test was about transformers (Vaswani et al., 2017), thinking that this was their bread and butter." The transformer is not bread nor is it butter, so we can assume that the embeddings for this sentence would skew from the intended meaning of the sentence. Now, we consider mapping "*their bread and butter*" with a word or phrase of similar meaning:
"They became relaxed..., thinking this was *their bread and butter*."
"They..., thinking this was *semantically similar word or phrase*."
"They..., thinking this was *easy*."
Or
"They... this was *going to be easy*."

In order for the language model to make this leap, expressions must be analyzed, the semanticity must be determined, and the phrase must be mapped to an equivalent or similar literal phrase or word in the embedding space.

Extending this to a vision model seems like a logical step. Vision models have become more ubiquitous recently with the introduction of QWEN-VL (Wang et al., 2024), DINOv2 (Oquab et al., 2023), GIT (Wang et al., 2022), Vision Transformers (Dosovitskiy et al., 2021) and adapted distillation models (Chen et al., 2022a) (Chen et al., 2022b) , and the reliable CNNs (LeCun et al., 1998). A vision and language model provides a concise description of an image. The existing models perform this task at a SOTA level for complicated images with complex description requirements with notable results from GIT on the MS COCO dataset (Lin et al., 2014) with CIDEr evaluation score of 138.5 (Vedantam et al., 2015) and QWEN-VL on the Flickr30k dataset (Young et al., 2014)with a BLEU-4 score of 41.2 (Papineni et al., 2002). Despite the impressive evaluation scores, these models still output a caption based on the actual facts of

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1479–1484

the image rather than an inferred meaning about the image.

Combining the vision and figurative language ideas can be considered difficult in isolation, but concatenating the image captions as a sentence, could allow for the multi-word expression treatment as described above. For the task of sequencing images and captions based on their relative position that may contain idiomatic expressions in visual representation or text, we employed the functionality of the BLIP vision and language model for image captioning (Li et al., 2022), while an instruction fine-tuned (Chung et al., 2022) QWEN-2.5 handled the image captions for sequencing (Wang et al., 2024).

2 Related Work

This general approach for idiomaticity has been refined using a variety of novel methods, including contrastive learning in the form of adaptive triplets (He et al., 2024), BERT-based binary classification (Devlin et al., 2019; Wang et al., 2021), single token approaches (Yin and Sch"utze, 2015; Li et al., 2018; Cordeiro et al., 2019; Phelps, 2022), analyses of pre-trained language model performance using standard evaluation techniques (Tayyar Madabushi et al., 2021), such as STS (Agirre et al., 2012) and cosine similarity (Salton and McGill, 1983), and recent studies concerned with recent LLM performance on various datasets (Phelps et al., 2024) . A number of datasets have also been created for evaluation and training in model development, including the Semeval 2022 task B dataset (Agirre et al., 2012), for multilingual idiomaticity detection and sentence embedding evaluation across languages; EPIE (Saxena and Paul, 2020), English possible idiomatic expressions, designed for binary idiomaticity classification tasks; MAGPIE (Haagsma et al., 2020), for idiom interpretation, paraphrasing, and contextual understanding; and Llidioms, multilingual linked idiom dataset. Vision models including the aforementioned QWEN-VL (Wang et al., 2024), text generation and image understanding; DINOv2 (Oquab et al., 2023), self-supervised image representations, GIT (Wang et al., 2022), a generative image-to-text transformer; Vision Transformers, replacing convolutional layers with an attention mechanism; distillation models, transfer learning; and CNNs (LeCun et al., 1998), extracting hierarchical structures from images which have primarily been pre-trained for generalizability for SOTA performance on various tasks or fine-tuned to perform exceedingly well on specific tasks, such as VQA (visual question answering) or image captioning. The performance of these general models is leveraged as the basis on which we attempted to build a fine-tuned and idiom-robust vision and language model.

3 System Overview

The goal of this task was to rank images based on how accurately they reflect the meaning of a nominal compound (potential idiom phrase) as used in a given context sentence. Formally, given a short text nominal compound (x), its corresponding context sentence (S), and a set of five images or image captions ($\mathcal{I}$), a machine learning system must determine the ranking of these five images ($y = [\text{image}_i]_{1 \leq i \leq 5}$).

For ranking the images, this approach was based on the instruction fine-tuning a Vision Language model and prompting technique to develop two approaches: *end-to-end* and *comparing operator* -based methods. In an *end-to-end* system, depicted in Figure 1, all information provided for this task is shown, including (1) instruction message, (2) nominal compound, (3) sentence context, (4) picture information and train a model to generate the ranking of all images together. In the approach based on *comparing operator*, the focus was on training a model to learn the operator of the comparison between two images, which image is closer to the meaning of the potential idiom phrase in sentence context, and then used these results to implement the insertion sort algorithms to achieve the final ranking of all images. Formally, the two approaches can be represented in the following formulas:

$$\textit{end-to-end}: \quad \mathbb{P}(y \mid x, S, \{\mathcal{I}_i\}_{1 \leq i \leq 5}) \tag{1}$$

$$\textit{comparing operator}: \quad \mathbb{P}(\{\text{Y, N}\} \mid x, S, \mathcal{I}_i, \mathcal{I}_{j|i \neq j}) \tag{2}$$

In the first approach (*end-to-end*), the Vision LMs were expected to be able to understand both the meaning of the focal phrase (nominal compound) expressed in the context sentence and the content of the provided images for the ranking process. This approach can utilize the full advantage of VLMs, which require the system to process whole images as input, thereby needing additional computing resources as the number of images increases. To avoid this problem, a second approach was developed, *comparing operator*, which only

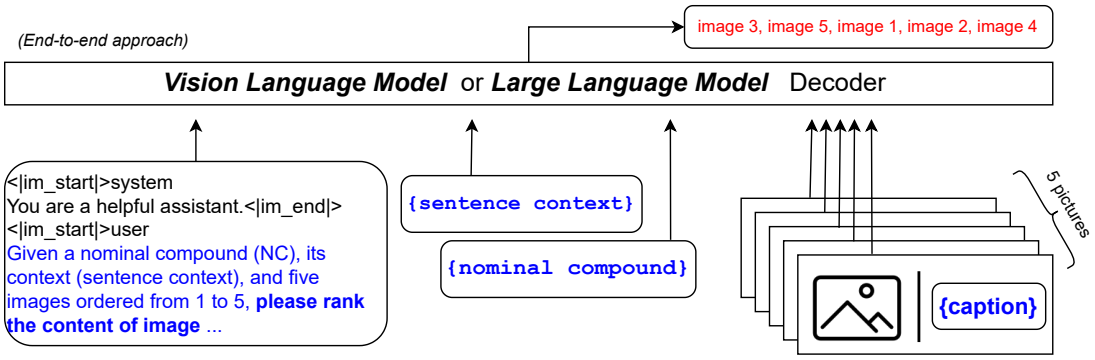

Figure 1: Overview of our approach for Subtask A.

compared a pair of images. The image chosen was closer in meaning to the focal phrase in the context sentence. Although this approach was able to deal with scaling the number of images for ranking, it may have ignored some features regarding the relationships among various images, allowing VLMs to avoid having an overall view of images.

The following methods represent work pertaining to Task1: AdMIRe subtask A, including *images and text* and *text only*. Overall, this method was shared between both these settings. For the images and text settings, the content of images was used to rank the similarity with the context sentence supported by pre-trained VLMs. For the text-only model, captions of images were utilized to represent the content and train LLMs. Moreover, extensive experiments were also conducted with a new caption generated by the BLIP (Li et al., 2022) language model to compare the effectiveness of the image's caption to the overall system.

Building on the strong vision-language understanding abilities of pre-trained models (Wang et al., 2024), instruction prompting was utilized to help the model interpret the task requirements. This approach aligned with instruction fine-tuning as described by Chung et al. (2022), employing a *causal language modeling* objective to train the LLM in generating the ranking of images or determining which image was closer to the context sentence. LoRA (Hu et al., 2022) was used to enhance efficiency, as it functions as a lightweight training method that minimizes the number of trainable parameters. The fine-tuned LLM was trained to model and generate the expected output based on the given input information.

$$s = \text{prompting}(y, x, S, \{\mathcal{I}_i\}_{1 \leq i \leq 5}) \quad (3)$$

$$\mathbb{P}(s) = \Pi_{z=1}^{|s|} \mathbb{P}(s_z | s_0, s_1, \ldots, s_{z-1}) \quad (4)$$

where s and x denote token sequences, and z represents the token index within the prompting input.

4 Experimental Setup

Dataset. For evaluating our methods, the original emotional dataset provided by the SemEval Task 1 (Pickard et al., 2025) organization was used. The dataset is divided into three subsets: training, development, test, and extended test sets, covering two phases of the competition: development and test. The held-out portion of the data is used for hyperparameter tuning, ensuring that the optimized checkpoints are chosen based on this internal development set.

Evaluation Metric. According to the competition guidelines, the evaluation metrics used to assess the quality of our model ranking are Top-1 Accuracy and the DCG score.

Settings. Experiments were conducted under the following settings, with results presented in Table 1:

1. Instruction fine-tuning of an LLM (`Qwen/Qwen2.5-72B-Instruct`) using the *end-to-end* strategy with original caption data for image information.

2. Similar to the first experiment, but using the *comparing operator* strategy.

3. Instruction fine-tuning of a Vision-Language Model (`Qwen/Qwen2.5-VL-72B-Instruct`) using the *end-to-end* strategy on raw image data (without caption text).

4. Similar to the first experiment, but using synthesized captions generated by the BLIP (Li et al., 2022) pre-trained model.

Task	Strategy	Caption Data	Model type	Top1 Acc.	DCG Score	Top1 Acc. (XE)	DCG Score (XE)
(Results obtained during the competition)							
Text only	*end-to-end*	*O*	LLM	**0.67**	3.04	0.51	2.857
Text only	*comparing operator*	*O*	LLM	0.20	2.30	0.44	2.769
Images and Text	*end-to-end*	G	LLM	0.53	**3.14**	0.55	**3.126**
Images and Text	*end-to-end*	*O+G*	LLM	0.53	2.91	0.58	3.037
(Results obtained in the post-evaluation phase)							
Images and Text	*end-to-end*	-	VLM	0.60	3.05	**0.62**	3.053

Table 1: Results on the test set. The values *O* and *G* in the *Caption Data* column represent the provided **o**riginal image caption and the synthesized caption **g**enerated by the BLIP LLM model, respectively.

5. Similar to the first experiment, but using a concatenation of the synthesized captions with the original captions.

5 Results and Analysis

Performance differences were observed across model types, input modalities, and task strategies, based on both competition and post-evaluation results. When instruction fine-tuning was performed on a language-only model (Qwen/Qwen2.5-72B-Instruct) using the end-to-end strategy with original caption data (O), the highest Top-1 accuracy (0.67) and strong Discounted Cumulative Gain (DCG) score (3.04) were obtained among all competition-phase models. Alternatively, when the same model was trained using the comparing operator strategy, performance dropped substantially (0.20 Top-1 accuracy, 2.30 DCG), indicating that comparison-style supervision was less compatible with the model's inference behavior.

Introducing multimodal input—either through synthesized captions (G), concatenated caption sources (O+G), or raw image data—led to important differences. Models fine-tuned with the end-to-end strategy using BLIP-generated captions (G) or concatenated original and generated captions (O+G) achieved moderately strong Top-1 accuracies (0.53 for both), though only slight improvements in DCG scores were observed. These models also demonstrated higher Top-1 accuracy when measured using a cross-entropy variant (0.55 and 0.58, respectively), suggesting better ranking reliability under probabilistic evaluation.

Post-evaluation results for the vision-language model (Qwen/Qwen2.5-VL-72B-Instruct), trained on raw image data using the end-to-end strategy, revealed the most balanced profile. While its Top-1 accuracy (0.60) was marginally lower than the best text-only system, the DCG score (3.05) and cross-entropy-based metrics (Top-1 accuracy of 0.62, DCG of 3.053) surpassed those of all other systems. These findings indicate that deeper integration of vision and language components—rather than only relying on intermediate captions—provides stronger generalization across both metrics.

Across all settings, models trained with the end-to-end strategy consistently outperformed those trained with comparative strategy, reinforcing the conclusion that generative instruction tuning better aligns with the strengths of both LLMs and VLMs. Further, multimodal enrichment via raw images or multiple caption sources proved beneficial, particularly under subjective and rank-sensitive evaluation.

6 Conclusion

This analysis reinforces the impact of task framing, modality integration, and caption strategy on the performance of instruction-tuned models for multimodal reasoning. Generative approaches consistently outperformed comparative ones, and native vision-language models showed strong post-evaluation performance, particularly in the DCG metric. Supplementing or replacing original captions with synthesized alternatives also yielded benefits, especially when used in combination. These findings support the continued development of instruction-tuned, generative pipelines with integrated multimodal architectures, and suggest that future systems should explore richer forms of input representation to maximize both top-1 and ranking-based performance across evaluation phases. Developments of this type will reach their optimal purpose by achieving the natural language understand-

ing and generation of figurative language sought after to help AI systems bridge gaps in communication.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

H. Chen, Y. Wang, and Z. Zhang. 2022a. Dearkd: Data-efficient early knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19206–19215.

H. Chen, Y. Wang, and Z. Zhang. 2022b. Vitkd: Feature-based knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19316–19325.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Silvio Ricardo Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. A computational model of nominal compound compositionality. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2762–2773. Association for Computational Linguistics.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*, arXiv:2010.11929.

Hessel Haagsma, Shiva Taslimipoor, and Siva Reddy. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024. Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss. In *Findings of the ACL: ACL 2024*, pages 12473–12485, Bangkok, Thailand. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Preprint*, arXiv:2201.12086.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, H. Touvron, H. Jégou, and I. Laptev. 2023. Dinov2: Learning robust visual features without supervision.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Dylan Phelps. 2022. drsphelps at SemEval-2022 task 2: Learning idiom representations using BERTRAM. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 158–164, Seattle, United States. Association for Computational Linguistics.

Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

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)*, Proceedings of Machine Learning Research. Association for Computational Linguistics.

Gerard Salton and Michael J. McGill. 1983. Introduction to modern information retrieval. In *McGraw-Hill Book Company*. McGraw-Hill.

P. Saxena and S. Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575. IEEE.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

S. Wang, L. Thompson, and M. Iyyer. 2021. Phrase-bert: Improved phrase embeddings from bert with an application to corpus exploration.

Wenpeng Yin and Hinrich Sch"utze. 2015. Convolutional neural network for paraphrase identification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 901–911. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 67–78.