AlexUNLP-NB at SemEval-2025 Task 1: A Pipeline for Idiom Disambiguation and Visual Representation

Mohamed Badran, Youssef Nawar, and **Nagwa El-Makky**
Computer and Systems Engineering Department
Alexandria University
{es-mohamedmostafabadran, es-YoussefNawar2023, nagwamakky}@alexu.edu.eg

Abstract

This paper describes our system developed for SemEval-2025 Task 1, subtask A. This shared subtask focuses on multilingual idiom recognition and the ranking of images based on how well they represent the sense in which a nominal compound is used within a given contextual sentence. This study explores the use of a pipeline, where task-specific models are sequentially employed to address each problem step by step. The process involves three key steps: first, identifying whether idioms are in their literal or figurative form; second, transforming them if necessary; and finally, using the final form to rank the input images.

1 Introduction

Idioms are a class of multi-word expressions (MWEs) that present significant challenges for current state-of-the-art models, as their meanings are often not predictable from the individual words that compose them. This unpredictability can create ambiguity between the literal, surface meaning derived from the component words and the idiomatic meaning intended by the expression. Addressing the complexities of MWE handling is crucial for natural language processing (NLP) applications (Constant et al., 2017). SemEval-2025 Task 1, AdMIRe: Advancing Multimodal Idiomaticity Representation (Pickard et al., 2025), focuses on evaluating models that incorporate both visual and textual information to assess their ability to capture idiomatic representations in two languages: English and Brazilian Portuguese.

Previous datasets and tasks, such as SemEval-2022 Task 2 (Madabushi et al., 2022) and the MAGPIE (Haagsma et al., 2020) and FLUTE (Chakrabarty et al., 2022) datasets, have primarily focused on idiomaticity detection. However, existing idiomaticity identification benchmarks can be exploited by models that ignore the nominal compound (NC) or its contextual usage, thus failing to develop robust semantic representations of idiomatic expressions (Boisson et al., 2023). SemEval-2025 Task 1 addresses these limitations by moving beyond binary classification and introducing richer representations of meaning through visual and visual-temporal modalities. SemEval-2025 Task 1 consists of two subtasks. Subtask A presents a set of five images alongside a contextual sentence containing a potentially idiomatic NC. The objective is to rank the images based on how well they represent the sense in which the NC is used within the given context. Subtask B presents a target expression and an image sequence missing the final image, the goal is to choose the best fill from four candidates images.

To tackle Subtask A, we propose a simple yet effective pipeline comprising three stages. The first stage involves detecting whether the potentially idiomatic NC in the given context is used literally or idiomatically. If the NC is identified as idiomatic, a separate model generates its literal meaning. In the final stage, images are aligned with the appropriate interpretation: either the generated literal meaning for idiomatic expressions or the direct literal meaning for non-idiomatic cases. This approach effectively combines idiomaticity detection, literal meaning generation, and multimodal image alignment to ensure accurate ranking of images based on the intended sense of the compound.

In our analysis, we evaluate various large language models (LLMs) for the first two stages of our pipeline and experiment with different zero-shot classification models in the third stage, across both English and Portuguese datasets. Our method achieves competitive results on SemEval-2025 Task 1, Subtask A, reaching 3rd and 6th places on the English and Portuguese benchmarks respectively. The code for our approach is publicly available at `https://github.com/MBadran2000/Idiom-MultiModal-Representation.git`.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 546–550

2 Background

2.1 Related work

LLMs have gained significant popularity across academic, industrial, and public domains due to their strong performance on a variety of tasks in zero-shot or few-shot prompting setups. These tasks include question answering, common-sense reasoning, and machine translation. Previous research has demonstrated that LLMs achieve competitive results on idiomaticity detection datasets, offering general applicability without the need for type-specific fine-tuning. However, they often lag behind fine-tuned encoder-only models on specific datasets and benchmarks (Phelps et al., 2024) . Despite this, LLMs possess a notable ability to disambiguate a wide range of nominal compounds without additional fine-tuning, as they tend to overlook construction artifacts present in idiomaticity detection datasets (Boisson et al., 2023). This allows them to generalize idiomaticity detection better than fine-tuned encoders.

Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) has emerged as the dominant model and learning approach in the vision-language domain. It employs a simple yet powerful contrastive learning loss to align visual and linguistic signals within a shared feature space, leveraging large-scale image-text datasets. Unlike earlier vision encoders such as VGG and ResNet, which were trained on the limited ImageNet dataset with simple categories described by only a few words, CLIP benefits from web-scale data paired with rich, descriptive text. This alignment between vision and language enables CLIP to excel in various multimodal tasks, including image-text retrieval, establishing it as the state-of-the-art model for diverse applications in the field (Huang et al., 2024).

2.2 Task description

AdMIRe: Advancing Multimodal Idiomaticity Representation focuses on developing high-quality representations of idioms, which are essential for improving applications such as machine translation, sentiment analysis, and natural language understanding. Enhancing models' ability to interpret idiomatic expressions can significantly boost the performance of these tasks. Unlike previous datasets, which often allow models to excel at idiomaticity detection without capturing the true semantic depth of idiomatic expressions, AdMIRe emphasizes meaning representation through visual and visual-temporal modalities. This approach aims to ensure that models develop a more comprehensive understanding of idioms beyond surface-level recognition.

AdMIRe consists of two subtasks, A and B. In this subtask, participants are presented with a set of five images alongside a contextual sentence containing a potentially idiomatic NC. The objective is to rank the images based on how accurately they represent the sense in which the NC is used within the given context. A variation of Subtask A replaces the images with text captions describing their content, offering two distinct settings for the subtask. Subtask B presents a sequence of three images resembling a comic strip, where the final image has been removed. Given a target expression, the objective is to select the most suitable completion from a set of four candidate images. The NC sense being depicted (idiomatic or literal) is not provided and should also be predicted. In this study, we concentrate on the Subtask A version that uses images for ranking.

Language	Training	Dev	Test	Ext. Eval.
English	70	10	15	100
Portuguese	32	15	13	55

Table 1: Number of samples in each split for English and Portuguese datasets.

The task leverages a dataset of potentially idiomatic expressions, building on the foundation of the SemEval-2022 Task 2 dataset. Table 1 summarizes the number of samples in each split for English and Portuguese, detailing the training, development, test, and extended evaluation sets. The extended evaluation set contains overlapping compounds from training, development, and test sets but in different contexts, offering a more robust assessment of model performance and generalization.

Performance in Subtask A is evaluated using two key metrics. The first is Top Image Accuracy, which measures the model's ability to correctly identify the most representative image. The second is Rank Correlation, assessed using Spearman's rank correlation coefficient, which evaluates how closely the model's image rankings align with the ground truth.

3 System Overview

Our proposed pipeline consists of three main stages: idiom detection, literal meaning generation, and

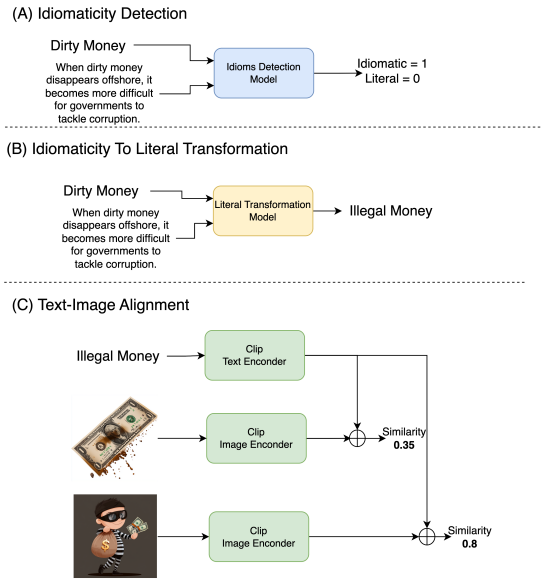

Figure 1: Idiomatic Case: First, we detect that it is idiomatic, then we obtain its literal meaning, and lastly, we align images according to the literal meaning.

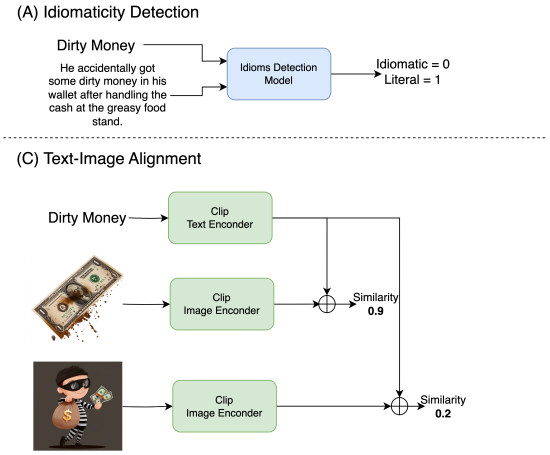

Figure 2: Literal Case: First, we detect that it is literal, then we align images according to the literal input.

multimodal image alignment. This pipeline ensures that images are aligned with the intended meaning of a given NC based on its contextual usage. Figures 1 and 2 illustrate the two possible workflows, depending on the outcome of the idiom detection stage.

3.1 Idiomatic vs. Literal Classification

The first step in our pipeline involves determining whether a NC in a sentence is used idiomatically or literally, using a zero-shot approach. For this task, we utilize an instruction-tuned large language model (LLM), which is prompted with the contextual sentence and predicts whether the NC conveys an idiomatic or literal meaning.

3.2 Literal Meaning Generation

If the NC is identified as idiomatic, we use a separate generative model in a zero-shot setting to generate its literal meaning. This stage also relies on an instruction-tuned large language model (LLM), ensuring that the generated interpretation remains contextually relevant. If the NC is classified as literal, this step is skipped, and the original NC is directly used for image retrieval.

3.3 Multimodal Image Alignment

Once the literal meaning of the NC is determined, we retrieve images that best correspond to the intended interpretation. We leverage a zero-shot multimodal image retrieval model based on CLIP to rank images based on their alignment with either the original NC (if used literally) or its generated literal meaning (if used idiomatically).

3.4 Portuguese Data

We also evaluated our approach on non-English data by testing the pipeline on the Portuguese dataset. Two strategies were explored. In the first, the translation approach, the NC is first classified as idiomatic or literal. After this classification and the generation of its literal meaning (if necessary), the Portuguese text is translated into English before proceeding to multimodal image alignment. This allows the full pipeline to operate in English, using translation to ensure cross-lingual compatibility. In the second strategy, the direct multilingual approach, we bypass translation by using multilingual multimodal image alignment models capable of processing multiple languages. This enables direct alignment of images with the Portuguese text, leveraging the models' ability to handle both textual and visual representations across languages.

4 Experiments

4.1 Experimental Setup

Our experiments were conducted in a zero-shot setting using a NVIDIA V100 GPU. For stages 1 and 2, we evaluated four instruction-tuned LLMs:

Stage 1 & 2 Model	Stage 3 Model	Train Top-1 Acc.	Train Rank Corr.	Dev Top-1 Acc.	Dev Rank Corr.
Llama-3.3 70B	CLIP ViT-H/14	0.5333	0.4617	0.7333	0.2133
Qwen2.5 72B	CLIP ViT-H/14	0.6167	0.5217	0.8000	0.4867
calme-3.2 78b	CLIP ViT-H/14	0.5833	0.4800	0.8000	0.3533
shuttle-3	CLIP ViT-H/14	0.5833	0.4817	0.8000	0.4933
Llama-3.3 70B	SigLIP SoViT-400m patch14	0.5167	0.4633	0.8000	0.4867
Llama-3.3 70B	BLIP ViT-large	0.6333	0.4917	0.5333	0.0750
Llama-3.3 70B	ALIGN	0.4833	0.4450	0.5333	0.2067
Llama-3.3 70B	MetaCLIP h14	0.5667	0.5100	0.6667	0.3200
Llama-3.3 70B	LLM2CLIP EVA02	0.5333	0.4183	0.7333	0.4200
Llama-3.3 70B	LLM2CLIP Openai	0.5500	0.4367	0.7333	0.1600
Ensemble of models combinations using CLIP ViT-H/14				0.8667	0.5067
Ensemble of models combinations using Llama-3.3 70B				0.8000	0.3200
Ensemble of all tested models combinations				0.8667	0.5200

Table 2: Comparison of different models on the English training and development sets

Stage 1 & 2 Model	Stage 3 Model	Approach	Train Top-1 Acc.	Train Rank Corr.	Dev Top-1 Acc.	Dev Rank Corr.
Qwen2.5 72B	CLIP ViT-H/14	Translation	0.500	0.416	0.300	0.210
Llama-3.3 70B	CLIP ViT-H/14	Translation	0.531	0.375	0.500	0.180
Llama-3.3 70B	SigLIP SoViT-400m patch14	Translation	0.469	0.419	0.300	0.080
Llama-3.3 70B	BLIP ViT-large	Translation	0.438	0.341	0.200	0.200
Llama-3.3 70B	ALIGN	Translation	0.438	0.316	0.600	0.240
Llama-3.3 70B	MetaCLIP h14	Translation	0.500	0.331	0.400	0.080
Llama-3.3 70B	LLM2CLIP EVA02	Multilingual	0.563	0.406	0.400	0.240
Llama-3.3 70B	LLM2CLIP Openai	Multilingual	0.594	0.403	0.800	0.320
Qwen2.5 72B	LLM2CLIP Openai	Multilingual	0.500	0.375	0.500	0.220
calme-3.2 78B	LLM2CLIP Openai	Multilingual	0.563	0.381	0.700	0.420
shuttle-3	LLM2CLIP Openai	Multilingual	0.531	0.431	0.600	0.360

Table 3: Comparison of different models and approaches on the Portuguese training and development sets

Llama-3.3 70B (Dubey et al., 2024), Qwen2.5 72B (Yang et al., 2024), calme-3.2 78B, and shuttle-3. In stage 3, we tested CLIP ViT-L trained on the LAION-2B English dataset (Schuhmann et al., 2022), ALIGN (Jia et al., 2021), BLIP (Li et al., 2022) ViT-large trained on the Flickr30k dataset, SigLIP SoViT-400m patch14 (Zhai et al., 2023), and LLM2CLIP (Huang et al., 2024) with base models EVA02-L/14 (Fang et al., 2024) and OpenAI ViT-L/14 CLIP. We explored all possible combinations of these models, resulting in 28 different configurations. Additionally, we tested ensemble approaches combining various model selections to further enhance performance.

4.2 Results

The results of our English experiments, including combinations of CLIP ViT-H/14 and Llama-3.3 70B, are presented in Table 2. Additionally, we report results for an ensemble of these models, as well as for all 28 tested model configurations. Our analysis revealed that individual models tend to make different types of errors, with failures varying across cases. As a result, model ensembling achieved the best performance by mitigating individual model errors through averaging, leading to overall improved results.

The results of both the translation approach and the direct multilingual approach on the Portuguese dataset are presented in Table 3. For multilingual experiments, we exclusively used LLM2CLIP models, as they rely on multilingual text encoders, which are essential for effectively aligning text and images across languages. Our findings indicate that the multilingual approach outperformed the translation-based method.

For Portuguese, we created an ensemble combining all four Phase 1&2 models with the two LLM2CLIP variants, which served as our final system. For English, our best-performing model was an ensemble of all tested models. The results of these top-performing models on both English and Portuguese are presented in Table 4. Our approach achieved competitive performance, ranking in the 3rd and 6th places on the English and Portuguese benchmarks respectively.

Language	Test Top-1 Acc.	Test Rank Corr.	Ext. Eval. Acc.	Ext. Eval. Rank Corr.
English	0.9300	0.4733	0.7200	0.3440
Portuguese	0.6154	0.3462	0.6153	0.3461

Table 4: Performance of the best-performing approaches on English and Portuguese test and extended evaluation datasets.

5 Conclusion

In this work, we proposed a three-stage pipeline for multimodal idiomaticity representation, consisting of idiom detection, literal meaning generation, and multimodal image alignment. We used this pipeline in our submission to subtask A of SemEval 2025 Task 1. Our approach leverages instruction-tuned LLMs for idiomaticity classification and literal meaning generation, followed by a zero-shot multimodal retrieval model for aligning images with the intended meaning of a nominal compound. We evaluated our system on both English and Portuguese datasets, exploring translation-based and direct multilingual approaches. Our results demonstrated that ensemble models improve performance by mitigating individual model errors, achieving competitive results in both languages.

References

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. *arXiv preprint arXiv:2311.00790*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations. *arXiv preprint arXiv:2205.12404*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2024. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171.

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

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, et al. 2024. Llm2clip: Powerful language model unlock richer visual representation. *arXiv preprint arXiv:2411.04997*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. Semeval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. *arXiv preprint arXiv:2204.10050*.

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. *arXiv preprint arXiv:2405.09279*.

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 Evaluation (SemEval-2025). Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.