Howard University-AI4PC at SemEval-2025 Task 1: Using GPT-4o and CLIP-ViT to Decode Figurative Language Across Text and Images.

Saurav K. Aryal and **Abdulmujeeb Lawal**
Electrical Engineering & Computer Science
Howard University

Abstract

Correctly identifying idiomatic expressions remains a major challenge in Natural Language Processing (NLP), as these expressions often have meanings that cannot be directly inferred from their individual words. The SemEval-2025 Task 1 introduces two subtasks, A and B, designed to test models' ability to interpret idioms using multimodal data, including both text and images. This paper focuses on Subtask A, where systems were given a context sentence that contains a potentially idiomatic nominal compound, with the goal being to rank the images based on how accurately they represent the meaning of the nominal compound used in the sentences. To address this, we employed a two-stage approach. First, we used GPT-4o to analyze sentences, extracting relevant keywords and sentiments to better understand the idiomatic usage. This processed information was then passed to a CLIP-ViT model, which ranked the available images based on their relevance to the idiomatic expression. Our results showed that this approach performed significantly better than directly feeding sentences and idiomatic compounds into the models without preprocessing. Specifically, our method achieved a Top-1 accuracy of 0.67 in English, whereas performance in Portuguese was notably lower at 0.23. These findings highlight both the promise of multimodal approaches for idiom interpretation and the challenges posed by language-specific differences in model performance.

1 Introduction

Deep learning and language models have substantially improved multilingual language understanding, as well as content and sentiment classification across domains (Aryal et al., 2023b,a), at both the sentence and token levels (Prioleau and Aryal, 2023; Aryal and Prioleau, 2023). However, idiomatic words and compounds pose a huge challenge for Large Language models. While humans can easily discern figurative usage from literal usage of words, models trained predominantly on idiomatic usage of words tend to "overlook" their literal meaning and assume idiomatic usage as long as such a word or compound is spotted. The SemEval-2025 Task 1, AdMIRe, aims to address this problem by trying to examine how image and textual modalities, similar to humans, can help improve idiom interpretations.

The dataset provided consists of potentially idiomatic nominal compounds and target sentences in which the compound is used, in both English and Brazilian Portuguese. It also includes five candidate images, which are to be ranked according to how closely they capture the intended meaning, be it either literal or idiomatic (Pickard et al., 2025).

While the task comprises two Sub tasks, A and B, this paper focuses solely on Subtask-A where systems are given a context sentence that contains a potentially idiomatic nominal compound and the goal is to then rank the images based on how accurately they represent the meaning of the nominal compound used in the sentences (Pickard et al., 2025). For example, given the image in Figure 1 below and the phrase *bad apple*, if the phrase is used literally in the target sentence, then the spoiled fruit depicted on the right would be more relevant and thus should be ranked higher, and if the phrase is used idiomatically in the target sentence, the more figurative illustration on the left should be ranked higher.

Our approach aims to capture the high-level semantic nuances and keywords of each compound in the given sentences and then compare the images using a vision-language model. Also, this approach does not involve training models on any data; rather, we just focus on leveraging large language model prompts and visual-textual alignment to allow us to correctly identify idiomatic nominal compounds.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1842–1848

Figure 1: Bad-apple Illustration from task description document

2 Related Work

Recent research on idiom detection has highlighted the challenge of modeling both the compositional and non-compositional aspects of idiomatic expressions. A fairly recent study introduced a multilingual dataset and evaluation framework, demonstrating that current models struggle to capture the dual nature of idioms (Tayyar Madabushi et al., 2022). Another study proposed IBERT, a BERT-based model for cloze-style idiom comprehension (Qin et al., 2021). Although this approach differs from ours, it underscores the importance of incorporating both local (immediate surrounding words) and global context (whole sentence/document). This aligns with our prompting strategy, which encourages models to focus not just on the word but on how it's used.

Also, while the aforementioned studies have mostly focused on textual data, recent developments, especially in multimodal models, do show that visual information can offer much. Greater contextual grounding has the potential to improve model performance. This idea is also supported by the embodied cognition framework, which posits that real-world imagery aids in semantic interpretation (Lakoff and Johnson, 1980). Extending on this, our work uses a vision-language model, combined with careful prompt engineering, to enable us to align visual and textual representations for more effective detection of literal versus idiomatic usage of certain nominal compounds.

3 System Overview

We designed our system to work on the idiomatic ranking without any fine-tuning. We describe each part of our pipeline in detail.

3.1 Data Preparation

The dataset provided by the organizers required minimal processing on our end. As such, our initial step involved extracting the given nominal compound and its corresponding sentence.

3.2 Textual Analysis and Nominal Compound Interpretation

We then analyzed the nominal compounds and their sentences. Here, we employed GPT-4o to assess the context and then determine whether the compound is used idiomatically or literally in the given sentence. The model evaluates the surrounding text and then classifies the usage accordingly. We also generate some sentiment cues and keywords related to the compound based on the usage. These cues are very important and helpful as they serve as the descriptors that inform the next stage of visual matching.

3.3 Visual Matching and Image Selection

Taking the sentiment cues and the keywords generated from GPT-4o, we then focus on the next stage, which is to select and rank the most representative image. Here, we make use of CLIP-ViT to extract visual features from each candidate image. Rather than just comparing the original sentence to the 5 candidate images, our approach compares the sentiment descriptors and keywords generated by GPT-4o with the visual features of the images. This makes sure that the selected image best represents the intended literal or figurative meaning of the nominal compound.

3.4 Final Output

The final output, which consists of the expected order of the ranked candidate images and the compound name, is then formatted according to the task requirements and stored as tab-separated files. This output is then submitted as our system's final result.

4 Implementation Details

All code was written using Python with a focus on integrating pre-trained models through API interfaces as opposed to training.

4.1 Infrastructure and Model Integration

The models integrated into our pipeline were CLIP-ViT (openai/clip-vit-base-patch32) and GPT4o.

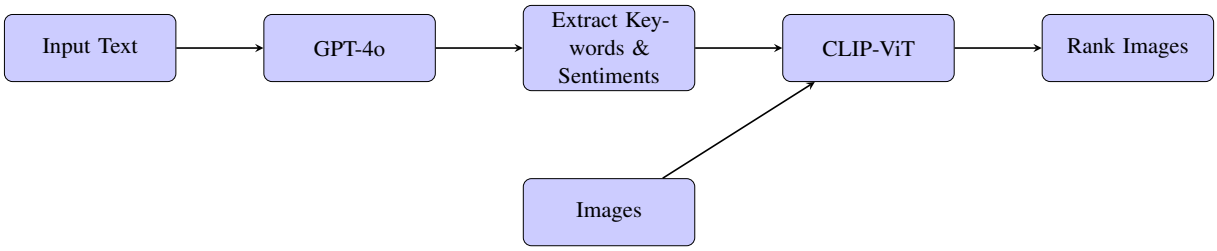

Figure 2: Diagram Illustrating the System Overview

As for the programming language, we used Python 3.10. Also, there were little to no storage concerns, and it was all stored on a MacBook Pro with 512GB storage space and 18GB RAM without CUDA support. There were also no performance concerns since all the models, aside from CLIP, were accessed through their respective APIs. The CLIP-ViT model was accessed through the Hugging Face transformers library and was downloaded locally for use.

4.2 Prompt Engineering

Our method of prompt engineering was very critical to the system's performance. We made sure to include in our prompts:

- Clear instructions to distinguish between literal and idiomatic usage of the compounds.
- Some examples of both types of usage to help guide the model's reasoning.
- Clear and explicit requests for sentiment cues and descriptive keywords based on the global context.
- Formatting requirements for the output to ensure proper parsing when getting results.

The final prompts utilized can be found in the Appendix.

4.3 Data Processing Pipeline

The Portuguese data was first translated into English to work with. After that, both the English and the now-translated Portuguese datasets had their compounds and sentences extracted. The next step was to then prompt the GPT4o model to extract the literal meaning, literal keywords, literal sentiments, as well as the idiomatic meaning, idiomatic keywords, and idiomatic sentiments of the nominal compound. The images were then loaded, opened, and prepped for comparisons. Another prompt then took in the sentence and the compound, and after reasoning, decided whether the word usage was idiomatic or literal. If the usage was literal, then the images were compared against the literal sentiments and keywords and then ranked. If the usage happened to be idiomatic, then the images were compared against the idiomatic keywords and sentiments and then ranked. Then the output data was processed in the submission format required by the task organizers.

5 Evaluation

We evaluated our system using Top-Image Accuracy and Discounted Cumulative Gain (DCG) scores, which were provided upon submission by the Coda bench platform. The Top-Image Accuracy reflects only whether the single most representative image – the one that should be ranked first – is correctly identified, as opposed to the DCG, which accounts for the entire ranking of images according to the weighting scheme detailed in the task description paper. Our approach showed a relatively strong performance with our latest submission averaging a Top Image Accuracy of 0.67 in English. Further experimentation, particularly with a higher temperature setting in the GPT4 model, did achieve a submission with up to 0.8. While we do recognize that this could be due to chance, it still does highlight the impact of varying temperatures alongside prompting.

Language	Top-1 Accuracy	DCG Score
English	0.67	3.13
Portuguese	0.23	2.64

Table 1: Final ranking scores for English and Portuguese.

To further give some context to the importance of our results, we note that the task organizers involved human annotators, who were all self-described fluent English Speakers, to work on an extended English dataset. The human annotators

achieved a Top-image accuracy score of 0.71 and a DCG of 3.22 in English, which indicates that our system, while simple, approaches human-level accuracy and ranking quality. (Pickard et al., 2025).

Furthermore, the clear performance gap between English and Portuguese in our results highlights the need for improved adaptability and better models in different languages, as some information or context could have been lost during the translation process.

6 Limitations

While our approach demonstrates some decent performance on the SemEval-2025 Task 1, there are some limitations we need to mention.

Making our system rely on out-of-the-box models like GPT4 and CLIP-ViT could limit the performance on the task. This doesn't mean the models aren't very good in their own right, but the fact that we didn't include any fine-tuning could potentially limit how well they can perform when they meet certain words for the first time

Also, there was a notable gap in performance between the English and Portuguese datasets. We believe this could have been caused by the translation process, as the Portuguese texts were automatically translated into English without any additional quality control. This could have resulted in a loss of important contextual or semantic information. Consequently, we do recognize the need for better models that understand more languages to allow them to natively handle multiple languages without relying solely on translation.

Additionally, since we didn't use a traditional classifier but relied on the model's outputs, we were unable to explicitly validate the accuracy of the classifier step due to the absence of ground truth labels. This was done to avoid introducing bias into the work. As such, while the GPT classifiers' outputs might have been qualitatively useful, we do acknowledge that this would limit our ability to report standalone metrics on that.

Lastly, given that our GPT model was accessed primarily through API's, there's the possibility that the model could be updated and hence do better or even worse on certain tasks, and this means that moving forward, we could have less control over some of the model's behaviors using the prompts.

7 Conclusion

In this paper, we presented a multimodal system that used GPT4o to extract important context information from text and then used CLIP-ViT to rank the images based on that. Our approach relied very heavily on prompting as opposed to training models on new data, and this showed that it was possible to get cues from text which could help in image idiomatic image identification. Our system also revealed a very important limitation when working without finetuning any models - a lot of context information could be lost when translating from one language to another. Looking to the future, we think it's important that future research should focus on creating more datasets for idiomatic training in other languages and also focus on refining prompting strategies to ensure the models are always guided.

Acknowledgements

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.

References

Saurav Aryal and Howard Prioleau. 2023. Howard university computer science at semeval-2023 task 12: A 2-step system design for multilingual sentiment classification with language identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2153–2159.

Saurav K Aryal, Howard Prioleau, Gloria Washington, and Legand Burge. 2023a. Evaluating ensembled transformers for multilingual code-switched sentiment analysis. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 165–173. IEEE.

Saurav K Aryal, Ujjawal Shah, Legand Burge, and Gloria Washington. 2023b. From predicting mmse scores to classifying alzheimer's disease detection & severity. *Journal of Computing Sciences in Colleges*, 39(3):317–326.

George Lakoff and Mark 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. *arXiv preprint arXiv:2503.15358*.

Howard Prioleau and Saurav K Aryal. 2023. Benchmarking current state-of-the-art transformer models on token level language identification and language

pair identification. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 193–199. IEEE.

Ruiyang Qin, Haozheng Luo, Zheheng Fan, and Ziang Ren. 2021. Ibert: Idiom cloze-style reading comprehension with attention. *arXiv preprint arXiv:2112.02994*.

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. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

A Appendix

Figure A: Prompt for Sentiments and Keywords

Analyze the given term: "{idiomatic_word}".
1. Literal Meaning:
- Provide a short and descriptive explanation of the literal meaning of "{idiomatic_word}". Include things that can help an AI model find it
- Then, generate 5 strongly associated keywords that could strongly describe the literal meaning of the word. Focus on the most relevant and descriptive words.
- Also generate 3 sentiments that could relate strongly to it based on the sentence given.
2. Idiomatic Meaning:
- Provide a short and descriptive explanation of the idiomatic meaning of "{idiomatic_word}" including "{idiomatic_word}" in the explanation. Include things that can help an AI model find it.
- Then, generate 5 descriptive keywords that strongly describe the idiomatic meaning based on the sentence, focusing on its core essence and usage.
- Also generate 3 sentiments or emotions that relate strongly to it based on the sentence given.

Return the result in the following JSON format:
"literal_meaning": "<brief explanation of the literal meaning>",
"literal_keywords": ["<word1>", "<word2>", "<word3>", ...],
"literal_sentiments" : ["<word1>", "<word2>", "<word3>", ...],
"idiomatic_meaning": "<brief explanation of the idiomatic meaning>",
"idiomatic_keywords": ["<word1>", "<word2>", "<word3>", ...]
"idiomatic_sentiments" : ["<word1>", "<word2>", "<word3>", ...],

This is important to me

Figure B: Classification Prompt

"Determine whether the usage of the word '{word}' in the following sentence is idiomatic or literal. This is important to me: \n"
"Sentence: {sentence}\n"
"Respond with 'idiomatic' or 'literal' and don't give an explanation"