daalft at SemEval-2025 Task 1: Multi-step Zero-shot Multimodal Idiomaticity Ranking

David Alfter
Gothenburg Research Infrastructure in Digital Humanities
University of Gothenburg
Sweden
david.alfter@gu.se

Abstract

This paper presents a multi-step zero-shot system for SemEval-2025 Task 1 on Advancing Multimodal Idiomaticity Representation (AdMIRe). The system employs two state-of-the-art multimodal language models, Claude Sonnet 3.5 and OpenAI GPT-4o, to determine idiomaticity and rank images for relevance in both subtasks. A hybrid approach combining o1-preview for idiomaticity classification and GPT-4o for visual ranking produced the best overall results. The system demonstrates competitive performance on the English extended dataset for Subtask A, but faces challenges in cross-lingual transfer to Portuguese. Comparing Image+Text and Text-Only approaches reveals interesting trends and raises questions about the role of visual information in multimodal idiomaticity detection.

1 Introduction

The SemEval-2025 Task 1 tests multimodal language models' ability to understand idioms by having them rank images based on how well they match idiomatic or literal uses of expressions in context, addressing previous datasets' limitations and exploring whether adding visual information can improve models' comprehension of figurative language; the task consists of two subtasks: ranking 5 images based on how well they match an idiomatic expression used in a sentence (Subtask A), and selecting the most appropriate final image to complete a 3-image sequence while determining if the expression is being used idiomatically or literally (Subtask B) (Pickard et al., 2025).

The data consists of a text file containing the textual data (expression, sentence, image names) and subfolders for each expression containing the images proper. The data is provided by the organizers and partitioned into Train/Dev/Test, plus an additional Extended test set. Table 1 summarizes the data for both Subtask A and B.

Data	# items	
	Subtask A	**Subtask B**
English		
Train	70	20
Dev	15	5
Test	15	5
Extended	100	30
Portuguese		
Train	32	-
Dev	10	-
Test	13	-
Extended	55	-

Table 1: Data summary

2 Related Work

Recent advancements in multimodal language models and the growing availability of datasets that integrate textual and visual information have propelled the task of multimodal idiomaticity representation and detection to the forefront of research (Filippatou, 2024; Pickard et al., 2025). However, even state-of-the-art language models, including large language models (LLMs), struggle to match human performance in comprehending idiomatic expressions (Tayyar Madabushi et al., 2021; Chakrabarty et al., 2022; Phelps et al., 2024). To bridge this gap, multimodal representation learning models, such as CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022), and generative models such as GPT-4 (OpenAI et al., 2024), have emerged as promising solutions, exhibiting strong performance in tasks that require cross-modal understanding, making them particularly well-suited for idiomaticity detection.

Cross-lingual transfer remains a challenging area in multimodal contexts, with models like mBERT (Devlin et al., 2019) and XLM-R (Conneau and

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 127–140

Lample, 2019) often experiencing performance degradation when applied to multimodal datasets. Recent studies have explored methods for improving cross-lingual transfer, such as multilingual embeddings and adversarial training (Wang et al., 2021), but consistent performance across diverse languages is yet to be achieved. Hybrid approaches that combine the strengths of multiple models are increasingly adopted for complex multimodal tasks (Guo et al., 2024). The role of visual information in idiomaticity detection remains an open question, with some studies suggesting that visual cues can enhance accuracy (Gu et al., 2023), while others argue that their contribution is context-dependent (Gupta et al., 2022). Artifacts present in existing datasets may allow models to perform well at idiomaticity detection without necessarily developing high-quality representations of the semantics of idiomatic expressions (Boisson et al., 2023). However, good representations of idioms are crucial for downstream applications such as sentiment analysis, machine translation, and natural language understanding (Tayyar Madabushi et al., 2021).

3 Methodology

3.1 System Overview

Our system for SemEval-2025 Task 1: Multimodal Idiomaticity employs two state-of-the-art multimodal language models: Claude Sonnet 3.5 and OpenAI GPT-4o.[1] Given the performance on the original test dataset, we opt to use only OpenAI for the extended dataset.[2] The system first determines whether the expression in the given context is used idiomatically or literally using a zero-shot classification approach. For Subtask A, the input is the provided sentence, while for Subtask B, the image descriptions of the first two images in the sequence are used. The model then ranks the candidate images based on their relevance to the literal or idiomatic interpretation of the expression.

We selected Claude and OpenAI models for their state-of-the-art multimodal reasoning capabilities, strong zero-shot performance, and complementary strengths in handling both textual and visual inputs. Both models exhibit efficient and tightly integrated vision-language processing, which is especially valuable in multimodal tasks, and robust multilingual understanding. Both models are widely regarded for their reliability, accessibility through stable APIs, and support for intermediate reasoning chains, making them well-suited for a hybrid system with an intermediate interpretation step.While alternative models like Gemini, LLaVA, or open-source LLMs (e.g., LLaMA or Mistral-based variants) were considered, they either lacked comparable multimodal maturity, cross-lingual robustness, or were not readily deployable at the time of experimentation. The selected models provided a pragmatic balance of performance, versatility, and ease of integration.

3.2 Idiomaticity Classification

To determine whether the expression is being used idiomatically or literally, we employ a zero-shot classification approach using the pre-trained language models. For Subtask A, the input sentence is directly fed to the model, while for Subtask B, the concatenated image descriptions of the first two images in the sequence are used. The model predicts the idiomaticity label based on its understanding of the expression in context, without any additional fine-tuning or examples provided during the task.

3.3 Image Ranking

Once the idiomaticity of the expression has been determined, the model is tasked with ranking the candidate images based on their relevance to the literal or idiomatic interpretation. For both Subtask A and B, the target expression and the predicted idiomaticity label are used to construct the prompt. The model then scores each candidate image using its knowledge of the expression's meaning and the visual content, producing a ranked list.

3.4 Improvement for Portuguese

Upon observing subpar performance on the Portuguese subset of the data, we experiment with translating some of the prompts to Portuguese before feeding them to the model. This allows the model to better understand the nuances of the expressions in their original language context. The translations are performed using GPT-4o.

3.5 Explanation-based Ranking

As an additional experiment for Subtask A, we introduce an intermediate explanation step to improve the model's understanding of the expression in context. After classifying the idiomaticity, the model is prompted to provide a brief explanation

[1] Parameters and prompts can be found in Appendix A

[2] We implement a fallback to Claude in case the model responds with "I apologize..." or "I'm unable to..."

of the literal or idiomatic meaning of the expression as used in the sentence. This explanation is then incorporated into the prompt for ranking the images, providing additional context to guide the model's selection.

3.6 Hybrid Approach with o1-preview and GPT-4o

In an effort to further improve the system's performance, we investigated a hybrid approach leveraging the complementary strengths of OpenAI o1-preview and GPT-4o. o1-previewexhibits strong performance on natural language understanding and generation tasks. We employ o1-preview for the idiomaticity classification and explanation steps, capitalizing on its robust language understanding capabilities. However, as o1-preview does not have the capability to directly process and reason about images, we continue to use GPT-4o for the visual ranking component. This hybrid strategy allows us to benefit from o1-preview's language understanding while still incorporating the visual reasoning capabilities necessary for the task. Interestingly, we found that this combination of models produced the best overall results on the SemEval-2025 Task 1 datasets, suggesting that the strengths of the two models are indeed complementary and can be effectively combined to tackle multimodal idiomaticity challenges.

3.7 Output Parsing and Post-processing

A key challenge in using large language models like GPT-4o for this task is that their generated outputs do not always strictly adhere to the specified prompt format, necessitating robust parsing and post-processing steps. For instance, when prompted to provide a ranking of the candidate images, the model's response may not be a well-formed array or list, requiring additional effort to extract the intended ranking. Additionally, we observed that the model occasionally produces rankings that are offset by one position, likely due to confusion about whether to use zero-based or one-based indexing. To mitigate these issues, we implement a flexible parsing system that can handle a variety of potential output formats. This includes using regular expressions to identify and extract ranked lists or arrays, as well as heuristics to detect and correct off-by-one errors in the rankings. By applying these post-processing techniques, we ensure that the final output of our system is consistent and aligns with the expected format for evaluation, even if the raw model outputs are somewhat noisy or inconsistent.

3.8 Evaluation

The system's performance is evaluated using the official metrics for each subtask. For Subtask A, we calculate the average ranking score across all test instances. For Subtask B, we measure both the ranking score and the idiomaticity classification accuracy. The submitted rankings and labels are compared against the gold standards provided by the task organizers. We report results on both the original and extended English datasets, as well as the Portuguese subset, to assess the effectiveness of our proposed improvements.

4 Results and Discussion

Tables 2 and 3 show the results for Subtask A and Subtask B, respectively. Additional plots can be found in Appendix B. Claude models are prefixed with *C-*, while OpenAI models are prefixed with *O-*. *DR* stands for "Detect [idiomaticity] and Rank", *DER* stands for "Detect, Explain, Rank". *DER2* models use o1-preview as reasoning LLM and GPT-4o as ranking LLM. Note that the *DER* and *DER2* models were only used in Subtask A Image+Text. For Portuguese, models suffixed with *-P* use prompts translated into Portuguese.

4.1 Subtask A: Image and Text

Our system achieves competitive performance on the English extended dataset for Subtask A, which involves ranking images based on their relevance to an idiomatic or literal expression in a given sentence. The best-performing model, O-DER2, attains an overall accuracy of 0.81, only slightly behind the top score of 0.83 reported by other participants. This result demonstrates the effectiveness of our hybrid approach combining o1-preview for idiomaticity classification and GPT-4o for image ranking. Binary classification scores (literal/idiomatic) are quite high, with accuracies of 0.93 on English, 0.97 on English Extended, 0.85 on Portuguese and 0.75 on Portuguese Extended.

Interestingly, the model exhibits a higher accuracy on literal expressions (0.94) compared to idiomatic ones (0.65), suggesting that identifying and ranking images for literal language use is an easier task. The Discounted Cumulative Gain (DCG) metric, which assesses the quality of the ranked image lists, shows a similar trend, with a higher

Model	Acc all	Acc lit	Acc id	Corr all	Corr lit	Corr id	DCG all	DCG lit	DCG id
Image and Text									
English									
C-DR	0.66	**0.86**	0.50	0.15	0.01	0.26	3.17	**3.37**	3.00
O-DR	0.80	**0.86**	0.75	0.17	0.10	0.22	3.30	3.35	3.26
O-DER	0.80	**0.86**	0.75	0.17	0.10	0.22	3.30	3.35	3.26
O-DER2	**0.87**	**0.86**	**0.88**	**0.52**	**0.29**	**0.73**	**3.43**	3.35	**3.49**
English Extended									
O-DER	0.78	0.79	**0.76**	0.40	0.45	**0.34**	3.30	3.33	**3.25**
O-DER2	**0.81**	**0.94**	0.65	**0.43**	**0.56**	0.28	**3.35**	**3.54**	3.13
Portuguese									
C-DR	0.46	0.29	0.67	0.11	0.23	-0.03	2.74	2.58	3.03
O-DR	0.46	0.29	0.67	0.21	0.20	0.22	2.80	2.51	3.10
O-DER	0.62	0.43	0.83	0.12	0.14	0.08	3.01	2.71	3.35
O-DER-P	0.69	0.42	**1.0**	0.29	**0.27**	0.32	3.11	2.72	3.56
O-DER2-P	**0.77**	**0.57**	**1.0**	**0.41**	0.21	**0.63**	**3.31**	**3.04**	**3.63**
Portuguese Extended									
O-DER-P	0.51	0.33	0.64	**0.26**	**0.27**	**0.25**	2.90	2.58	3.15
O-DER2-P	**0.56**	**0.42**	**0.68**	0.23	0.20	0.24	**2.95**	**2.66**	**3.17**
Text Only									
English									
C-DR	0.60	0.43	**0.75**	**0.35**	**0.27**	**0.41**	3.04	2.85	**3.21**
O-DR	**0.66**	**0.57**	**0.75**	0.21	0.07	0.34	**3.07**	**3.10**	3.04
English Extended									
O-DR	0.33	0.48	0.15	0.09	0.18	-0.01	2.61	2.90	2.28

Table 2: Results for Subtask A. Best scores per column and test set in bold. Bold omitted for last row.

score for literal expressions (3.54) than idiomatic ones (3.13).

On the Portuguese subset, our best model, O-DER2-P, achieves an overall accuracy of 0.77, with perfect performance on idiomatic expressions (1.0) but lower accuracy on literal ones (0.57). The DCG scores follow a similar pattern, with idiomatic expressions (3.63) outperforming literal ones (3.04). These results highlight the challenges of cross-lingual transfer and the need for further improvement in handling Portuguese idioms.

4.2 Subtask A: Text Only

In the text-only setting for Subtask A, our system demonstrates mixed performance. On the English dataset, the O-DR model achieves an overall accuracy of 0.66, with higher accuracy on idiomatic expressions (0.75) compared to literal ones (0.57). The DCG scores are relatively balanced, with 3.10 for literal expressions and 3.04 for idiomatic ones.

However, on the English extended dataset, the performance drops significantly, with an overall accuracy of 0.33 and a notable decrease in performance on idiomatic expressions (0.15) compared to literal ones (0.48). This suggests that the extended dataset introduces more challenging and diverse examples that require further improvements in our text-based idiomaticity classification approach.

Comparing the Text-Only results to the Image+Text setting, we observe that the inclusion of visual information generally improves performance, particularly on the English extended dataset. This highlights the importance of leveraging multimodal information for idiomaticity detection, especially in more complex and diverse scenarios.

4.3 Subtask B: Image and Text

In the image+text setting for Subtask B, our system achieves mixed performance on the English dataset. The O-DR model obtains an overall item accuracy of 0.60, with perfect accuracy on idiomatic expressions (1.0) but zero accuracy on literal ones (0.0).

Model	Item all	Item lit	Item id	Sent all	Send lit	Sent id
Image and Text						
English						
C-DR	0.20	0.0	0.33	0.80	**1.0**	0.67
O-DR	**0.60**	0.0	**1.0**	**1.0**	**1.0**	**1.0**
English Extended						
O-DR	0.23	0.17	0.33	0.77	0.94	0.50
Text Only						
English						
C-DR	0.60	0.50	0.07	0.8	**1.0**	0.67
O-DR	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
English Extended						
O-DR	0.60	0.78	0.33	0.77	0.94	0.50

Table 3: Results for Subtask B. Best scores per column and test set in bold. Bold omitted for English Extended.

However, the model achieves perfect sentence accuracy (1.0) for both literal and idiomatic expressions.

On the English extended dataset, the O-DR model's performance drops, with an overall item accuracy of 0.23 and sentence accuracy of 0.77. The model performs better on idiomatic expressions (0.33 item accuracy, 0.50 sentence accuracy) compared to literal ones (0.17 item accuracy, 0.94 sentence accuracy). This suggests that the extended dataset presents more challenging cases for image selection and idiomaticity classification, requiring further improvements in our multimodal approach.

4.4 Subtask B: Text Only

For Subtask B, which involves selecting the most appropriate final image to complete a 3-image sequence while determining the idiomaticity of the expression, our system demonstrates strong performance using only textual information. On the English dataset, the O-DR model achieves perfect scores across all metrics, correctly identifying the idiomaticity and selecting the appropriate final image for both literal and idiomatic expressions.

However, on the English extended dataset, the performance drops significantly, with an overall accuracy of 0.6 and lower scores on idiomatic expressions (0.33) compared to literal ones (0.78). This suggests that the extended dataset introduces more challenging and diverse examples that require further improvements in our text-based idiomaticity classification and image selection approach.

4.5 Comparison between Image+Text and Text Only

Comparing the results of Subtask A (Image+Text) and Subtask B (Text Only) reveals an interesting trend. While the inclusion of visual information in Subtask A generally improves performance, particularly on the English extended dataset, the text-only approach in Subtask B surprisingly outperforms the Image+Text approach on the English dataset. This suggests that the textual context alone can be sufficient for identifying idiomaticity and selecting appropriate images in some cases, and that the integration of visual information may introduce additional complexity or noise. However, it is important to note that the English extended dataset results for Subtask B show a significant drop in performance compared to the English dataset, indicating that the text-only approach may not generalize well to more diverse and challenging examples. Further investigation is needed to understand the factors contributing to this performance gap and to develop more robust multimodal approaches that can effectively leverage both textual and visual information.

4.6 Portuguese Performance

The results on the Portuguese subset for Subtask A highlight the challenges of cross-lingual transfer in multimodal idiomaticity detection. Despite the improvements achieved by translating the prompts to Portuguese and incorporating explanations, the overall performance remains lower compared to the English datasets. This suggests that there may be linguistic and cultural differences in idiomatic

Test set	Rank (O)	Rank (E)
Subtask A		
English (T+I)	5	2
English (TO)	3	5
Portuguese (T+I)	4	4
Subtask B		
English (T+I)	1	2
English (TO)	1	2

expressions that require further adaptation and fine-tuning of the models.

4.7 Overall Performance

In comparison to other submissions, according to the official leaderboard, our best models rank as follows: for Subtask A (Text+Image), we rank fifth on the original and second on the extended test set, for Subtask A (Text Only), we rank third and fifth, for Subtask A (Text+Image) Portuguese, we rank fourth on both test sets. For Subtask B, we rank first and second in both modalities.

5 Conclusion

In this paper, we present a multi-step zero-shot system for the SemEval-2025 Task 1 on Advancing Multimodal Idiomaticity Representation (AdMIRe). Our approach leverages state-of-the-art multimodal language models, including Claude Sonnet 3.5, OpenAI GPT-4o, and o1-preview, to address the challenges of idiomaticity detection and image ranking in both literal and idiomatic contexts.

The system demonstrates competitive performance on the English extended dataset for Subtask A, achieving an overall accuracy of 0.81 using a hybrid approach that combines o1-preview for idiomaticity classification and GPT-4o for visual ranking. However, cross-lingual transfer to Portuguese remains a challenge, highlighting the need for further research in adapting multimodal idiomaticity detection systems to different languages and cultural contexts.

Our analysis of the Image+Text and Text-Only approaches reveals interesting trends, with the Text-Only approach surprisingly outperforming the Image+Text approach on the English dataset for Subtask B. This raises questions about the role and effectiveness of visual information in multimodal idiomaticity detection, and calls for further investigation into the factors contributing to the performance differences across datasets and subtasks.

Future work should investigate the factors contributing to the performance differences between Image+Text and Text-Only approaches across datasets and subtasks to develop more effective multimodal idiomaticity detection.

Limitations

While our system demonstrates competitive performance on the SemEval-2025 Task 1 datasets, there are several limitations that should be acknowledged:

1. Our system relies on zero-shot classification for idiomaticity detection, which may not capture the full complexity and nuance of idiomatic expressions across different contexts and languages. Fine-tuning the models on task-specific data could potentially improve performance and generalization.

2. Although we experimented with translating prompts to Portuguese, our cross-lingual evaluation is limited to a single language. To assess the true effectiveness of our approach for multilingual idiomaticity detection, it would be necessary to evaluate on a wider range of languages and idioms. Reliance on pre-trained models: Our system heavily relies on the capabilities of pre-trained multimodal language models, such as GPT-4o and o1-preview. While these models have demonstrated strong performance on various tasks, they may have inherent biases or limitations that could impact the system's performance on specific idioms or cultural contexts.

3. The use of large pre-trained models in our system makes it challenging to interpret the decision-making process behind the idiomaticity classifications and image rankings. Developing more interpretable and explainable models could provide insights into the system's behavior and potential areas for improvement.

4. The use of large pre-trained models like GPT-4o and o1-preview requires significant computational resources, which may limit the accessibility and scalability of our approach for researchers and practitioners with limited resources.

Acknowledgments

The author would like to acknowledge Bill Noble and Mattias Appelgren for the constructive discussions on the topic.

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021).

Ethical Considerations

We acknowledge that the use of large language models for natural language processing tasks can be computationally intensive and consume significant energy and resources. For this reason, no prompt engineering or extensive fine-tuning of the LLM was conducted. The total computational costs incurred in this study are at approximately $28. While LLMs offer powerful capabilities, it is important for the research community to carefully consider the environmental impacts and strive to develop more computationally efficient approaches. Future work should explore techniques to reduce the carbon footprint of LLM usage without compromising performance. Judicious use of these models, along with transparency around the associated costs, can help balance the research benefits with the broader sustainability implications. Through mindful practices and continued innovation, we aim to harness the potential of LLMs in an ethically responsible manner.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

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.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

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.

Viktoria Filippatou. 2024. Finding Meaning in a Haystack: On How Vision and Language Models Process Figurative Language. Master's thesis, University of Gothenburg.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. 2022. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5078–5088.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin

Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

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.

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.

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.

Runchuan Wang, Zhao Zhang, Fuzhen Zhuang, Dehong Gao, Yi Wei, and Qing He. 2021. Adversarial domain adaptation for cross-lingual information retrieval with multilingual BERT. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3498–3502.

A Parameters and prompts

Claude 3.5 Sonnet	
max_tokens	8192
temperature	0
OpenAI GPT-4o	
No additional parameters were provided to the model.	

Table 4: Parameters provided to the models

Subtask A	
System	You are a skilled linguist with deep knowledge of idiomatic expressions. You can easily distinguish between idiomatic and non-idiomatic uses of phrases in English and Portuguese.
User	In the following sentence, is the expression *expression* used idiomatically or literally? Expression: *expression* Sentence: *sentence* Answer only with 'idiomatic' or 'literal'
System (I+T)	You are an expert in semantic analysis and image relevance evaluation. Given a classification of an expression as idiomatic or literal, your task is to: Assign each of five provided images to one of the following categories: 1. Synonym for the idiomatic meaning of the expression. 2. Synonym for the literal meaning of the expression. 3. Related to the idiomatic meaning, but not synonymous. 4. Related to the literal meaning, but not synonymous. 5. A distractor unrelated to either meaning. Rank the images based on their relevance to the identified meaning of the expression: - Synonyms should be ranked highest. - Related images should be ranked next. - Distractors should always be ranked lowest.
User (I+T)	Rank the following images for the expression *expression* used in a *idiomatic/literal* way, from most relevant to least relevant. Return an array of five numbers that correspond to the image numbers, like [1,4,3,2,5]. *image data*
User (T)	Rank the following sentences for the expression *expression* used in a *idiomatic/literal* way, from most relevant to least relevant. Return an array of five numbers that correspond to the sentence numbers, like [0,3,2,1,4]. 1. *caption1* 2. *caption2* 3. *caption3* 4. *caption4* 5. *caption5*
System	You are an expert in linguistic analysis with a deep understanding of idiomatic and literal expressions in *English/Portuguese*. Your task is to provide a clear explanation of an idiomatic or literal expression.
User	Explain *expression* used in a *idiomatic/literal* way.
User (I+T)	Given the following explanation of the expression *expression* used in a *idiomatic/literal* way, rank the images from most relevant to least relevant. Return an array of five numbers that correspond to the image numbers, like [1,4,3,2,5]. Explanation: *explanation*, *image data*
System (I+T)	Você é um especialista em análise semântica e avaliação de relevância de imagens. Dada a classificação de uma expressão como idiomática ou literal, sua tarefa é: Atribuir cada uma das cinco imagens fornecidas a uma das seguintes categorias: 1. Sinônimo para o significado idiomático da expressão. 2. Sinônimo para o significado literal da expressão. 3. Relacionado ao significado idiomático, mas não sinônimo. 4. Relacionado ao significado literal, mas não sinônimo. 5. Um distrator não relacionado a nenhum dos significados. Classificar as imagens com base na sua relevância para o significado identificado da expressão: - Os sinônimos devem ser classificados como os mais relevantes. - As imagens relacionadas devem ser classificadas em seguida. - Os distratores devem sempre ser classificados como os menos relevantes.
User (I+T)	Dada a seguinte explicação da expressão *expression* usada de forma *idiomatic/literal*, classifique as imagens da mais relevante para a menos relevante. Retorne um array de cinco números que correspondem aos números das imagens, como [1,4,3,2,5]. Explicação: *explanation*, *image data*

Table 5: Prompts for Subtask A. The first block describes the DR approach. The second block describes the additional prompts used for DER. The third block shows the translations used for Portuguese. Prompts only used for Image+Text are marked with (I+T), while prompts used only for text are marked (T)

	Subtask B
System	You are a skilled linguist with deep knowledge of idiomatic expressions.
User	Given the following sentences, is the expression most likely used literally or idiomatically? Answer only with 'idiomatic' or 'literal'! Expression: *expression* Sentences: *sentences*
System (I+T)	You are a skilled visual artist specialized in images that convey idiomatic or literal meanings. You can easily rank images in terms of relevance to idiomatic and non-idiomatic uses of phrases in English. Respond only with a number.
User (T)	Given the following expression used in a *idiomatic/literal* way, and the following description, which of the following four sentences best continues the description. Respond only with the sentence number (1,2,3,4). Expression: *expression* Description: *sentences* 1. *caption1* 2. *caption2* 3. *caption3* 4. *caption4*
User (I+T)	Given the following expression used in a *idiomatic/literal* way, and the following two images, which of the following four images best continues the description. Respond only with the image number (1,2,3,4). Expression: *expression*, *image data*

Table 6: Prompts for Subtask B. Prompts only used for Image+Text are marked with (I+T), while prompts used only for text are marked (T)

B Results: Plots

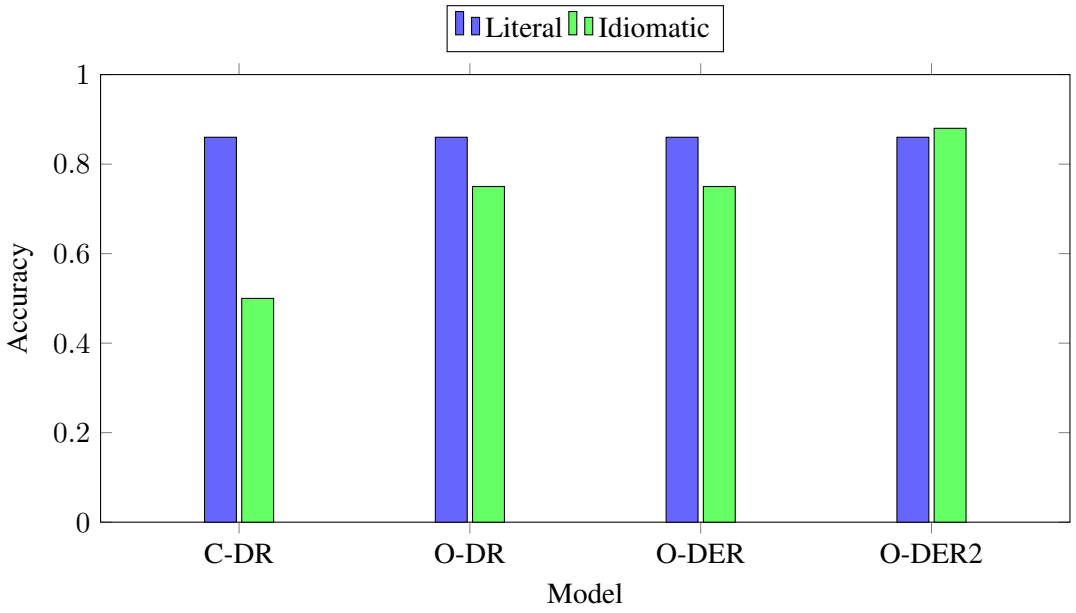

Figure 1: Comparison of accuracy for literal vs. idiomatic expressions on English dataset (Image+Text)

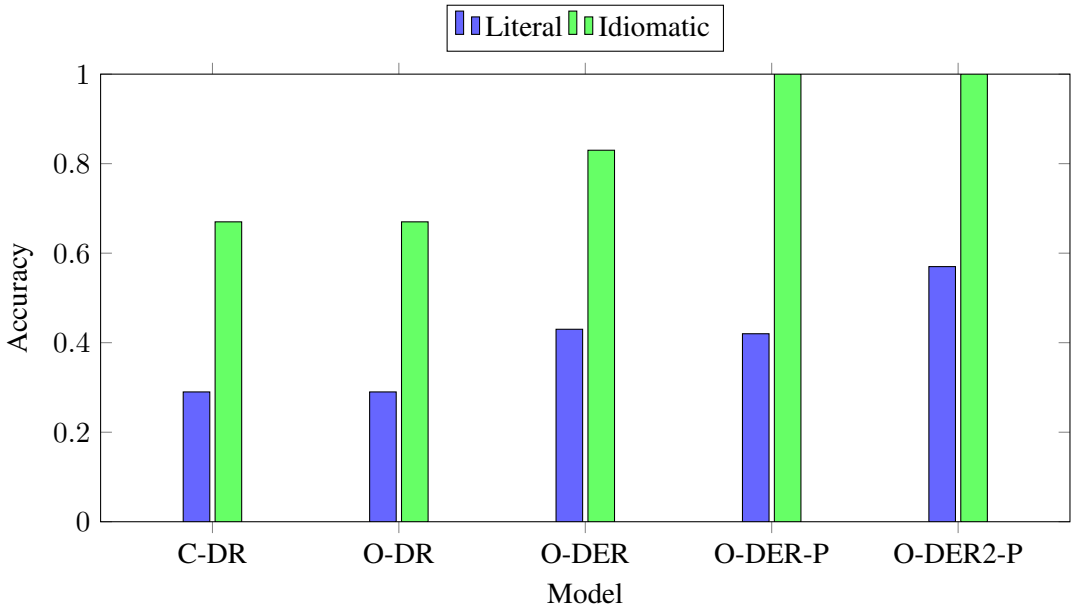

Figure 2: Comparison of accuracy for literal vs. idiomatic expressions on Portuguese dataset (Image+Text)

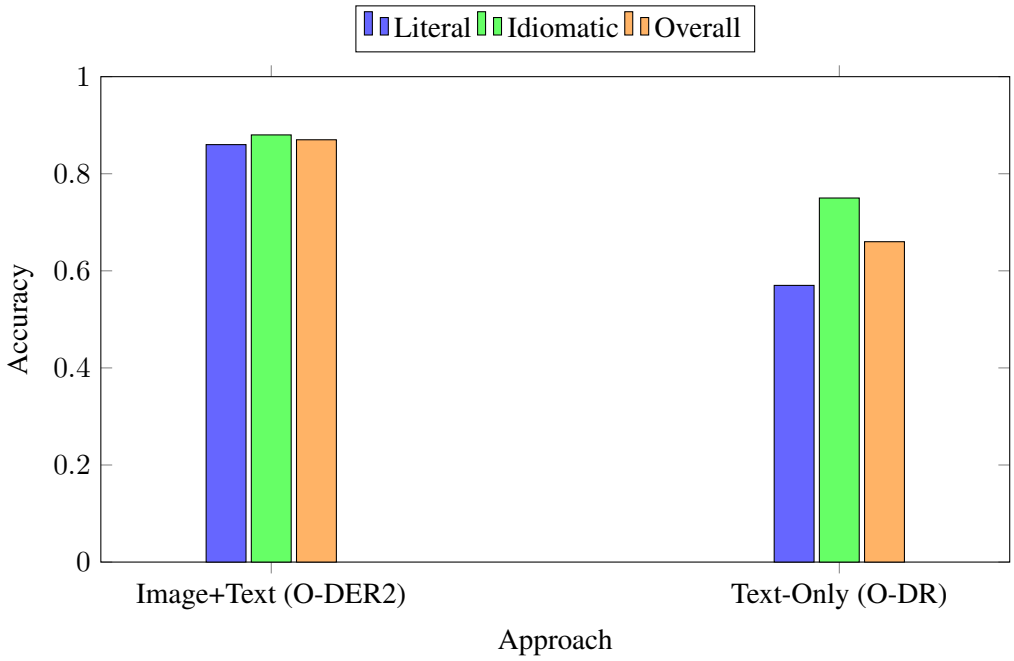

Figure 3: Comparison of performance between Image+Text and Text-Only approaches on English dataset

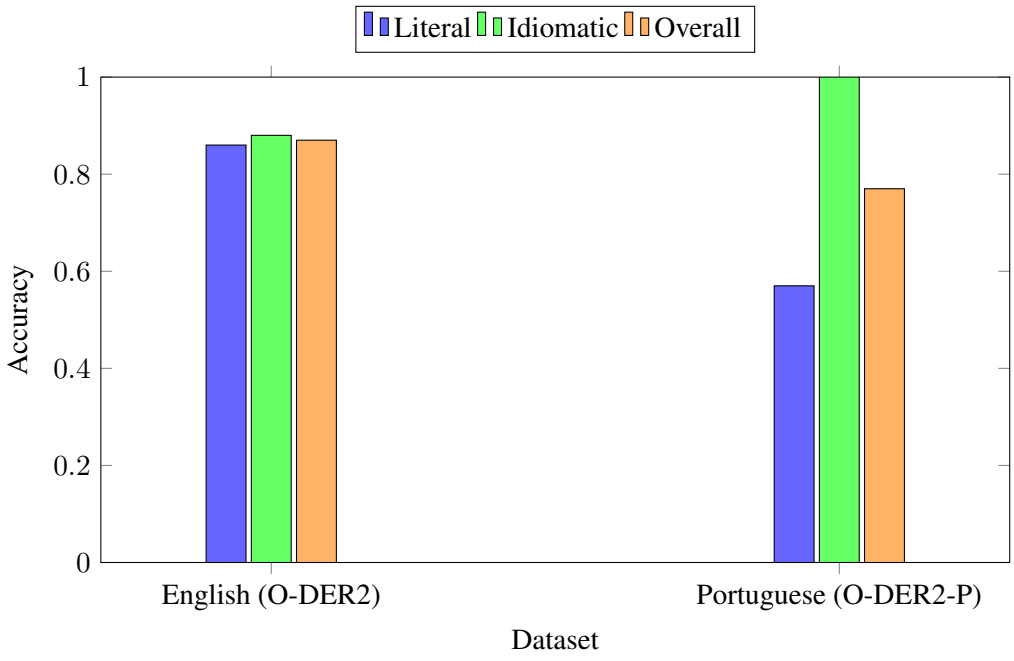

Figure 4: Comparison of best-performing models on English and Portuguese datasets

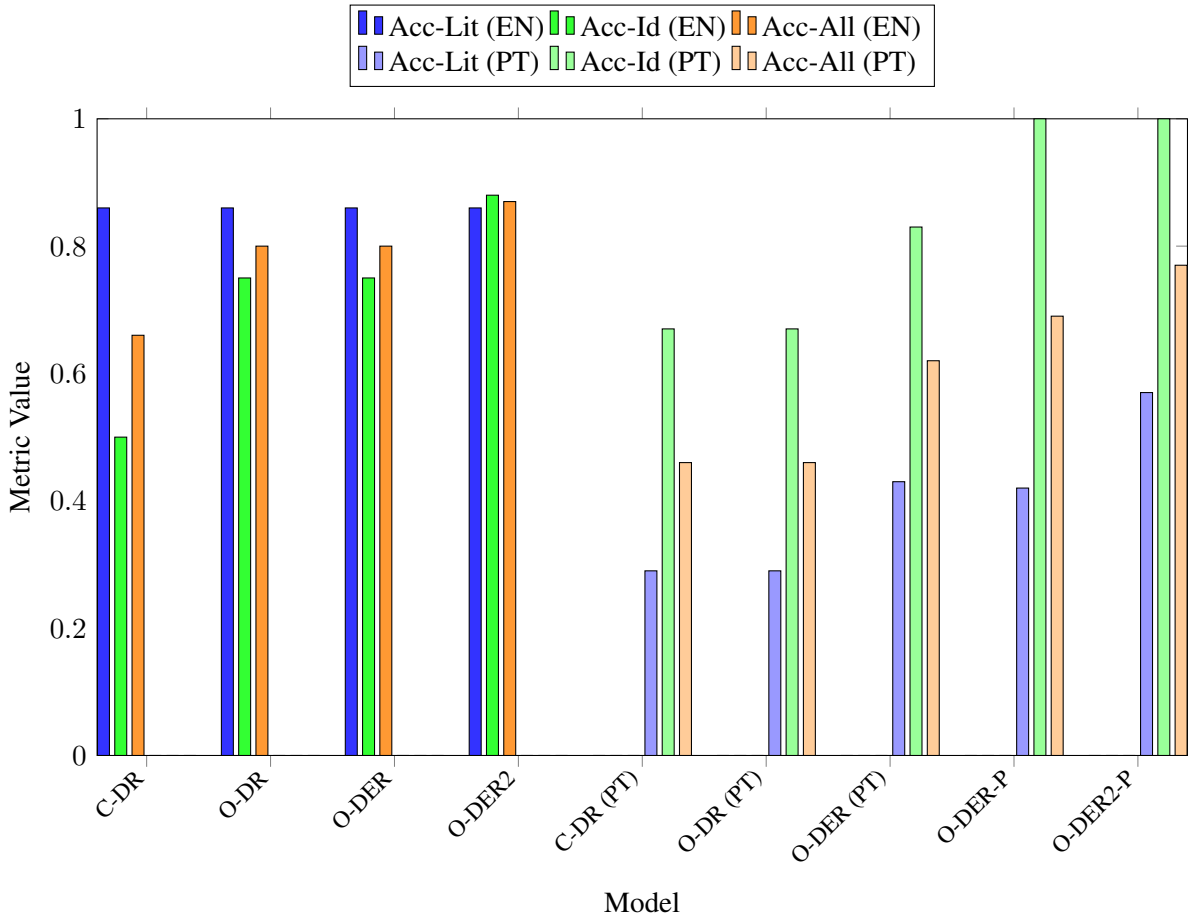

Figure 5: Comparison of model performance across English (EN) and Portuguese (PT) datasets with Image+Text modality