PoliTo at SemEval-2025 Task 1: Beyond Literal Meaning: A Chain-of-Though Approach for Multimodal Idiomacity Understanding

# PoliTo at SemEval-2025 Task 1: Beyond Literal Meaning: A Chain-of-Though Approach for Multimodal Idiomacity Understanding

**Lorenzo Vaiani**
Politecnico di Torino
lorenzo.vaiani@polito.it

**Davide Napolitano**
Politecnico di Torino
davide.napolitano@polito.it

**Luca Cagliero**
Politecnico di Torino
luca.cagliero@polito.it

## Abstract

Idiomatic expressions present significant challenges for natural language understanding systems as their meaning often diverge from the literal interpretation. While prior works have focused on textual idiom detection, the role of visual content in reasoning about idiomaticity remains underexplored. This study introduces a Chain-of-Thought reasoning framework that enhances idiomatic comprehension by ranking images based on their relevance to a compound expression used in a reference sentence, requiring the system to distinguish between idiomatic and literal meanings. We comprehensively evaluate our approach by quantitatively analyzing the performance improvements achieved integrating textual and visual information in the ranking process through different prompting settings. Our empirical findings provide insights into the capabilities of visual Large Language Models to establish meaningful correlations between idiomatic content and its visual counterpart, suggesting promising directions for multimodal language understanding.

## 1 Introduction

Idiomatic Expressions (IEs) are a unique and challenging natural language aspect. Their meaning often cannot be inferred directly from their individual words, making them particularly difficult for computational models to process (Mi et al., 2024). Unlike literal phrases, IEs require understanding linguistic conventions, context, and sometimes even cultural background (Hajiyeva, 2024). Given their widespread use, accurately interpreting idioms is crucial for many natural language processing (NLP) tasks, including fact-checking, hate speech detection, sentiment analysis, machine translation, and question-answering (Yosef et al., 2023; Tan and Jiang, 2021), Misunderstanding the idiomatic meaning can lead to significant errors in these applications, affecting accuracy and usability.

The recent success of Large Language Models (LLMs) has significantly advanced the field. They have demonstrated strong performance in several NLP tasks through zero-shot and few-shot prompting (Wei et al., 2022b,a), showcasing their ability to handle complex reasoning challenges with minimal supervision. However, their ability to effectively process IEs remains an open question (De Luca Fornaciari et al., 2024). Additionally, as visual LLMs become widespread, it is worth investigating whether these models can effectively associate visual information with the IEs' meaning.

In this work, we propose a novel Chain-of-Thought (CoT) framework to explore multimodal idiomaticity understanding. Specifically, we investigate how visual LLMs can integrate textual and visual information to establish meaningful connections between IEs and their corresponding visual representations. Our approach leverages structured reasoning to guide LLMs in ranking images based on their relevance to a given either literal or idiomatic compound, assessing whether visual-text content relations contribute to a more accurate interpretation of IEs' meaning.

Our contributions are twofold: (i) We introduce a novel multimodal idiomaticity understanding framework using Chain-of-Thought prompting, and (ii) We investigate how visual LLMs can leverage step-by-step reasoning to accurately rank images based on the meaning of a compound word as used in a given sentence, distinguishing between idiomatic and literal interpretations.

The code and some examples are available at PoliTo-AdMIRe repository[1].

## 2 Related Works

While previous research has explored idiom detection and interpretation from text, the role of mul-

[1] https://github.com/DavideNapolitano/Beyond-Literal-Meaning-A-Chain-of-Though-Approach-for-Multimodal-Idiomacity-Understanding/tree/main



*Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2071–2076

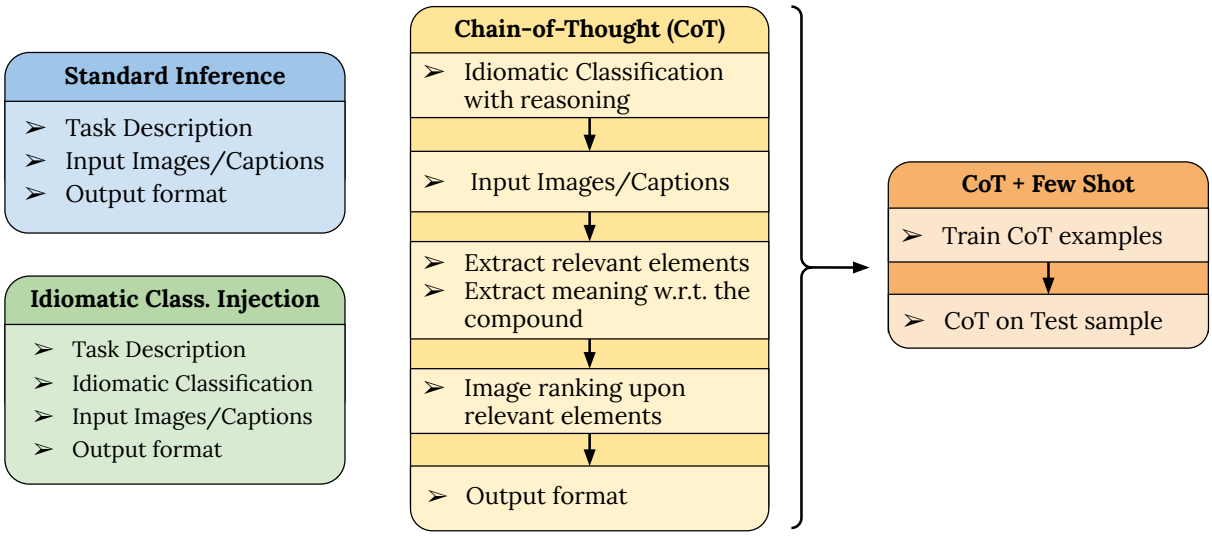


Figure 1: Different Prompt settings. Each titled block describes a different tested approach. Each inner block represents an interaction with the LLM. Consecutive inner blocks connected by an arrow represent consecutive interactions in the same LLM chat.

timodality in idiomaticity understanding remains largely unexplored. Most related studies have addressed figurative language understanding and disambiguation of a mix of visual and textual content. Specifically, visual figurative meaning understanding (Saakyan et al., 2024) aims to asses visual premise entails or contradicts a textual hypothesis. Instead, given a target word with limited textual context, visual-word sense disambiguation (Raganato et al., 2023) focuses on selecting, among a set of candidate images, those corresponding to its intended meaning. The AdMIRe challenge (Pickard et al., 2025) extends this idea to idiomatic expressions, considering idioms as textual descriptions and images that capture their intended meaning.

Transformer architectures, such as CLIP (Radford et al., 2021), and visual LLMs, such as LLaVA (Liu et al., 2023), have already shown promising performance on several multimodal tasks (Kulkarni et al., 2024; Vaiani et al., 2023; D'Amico et al., 2023; Napolitano et al., 2024) Furthermore, more advanced visual LLMs, such as Qwen2.5-VL (Bai et al., 2025) or Gemini (Team et al., 2023), have been trained to handle multiple images. Finally, combining visual models and LLMs with multimodal Chain-of-Thoughts (CoT) already demonstrated state-of-the-art performance in many vision-language tasks (Shao et al., 2024; Mondal et al., 2024; Zhang et al., 2024). Accordingly, our approach relies on CoT to improve visual LLMs' understanding of idiomatic expressions.

## 3 Methodology

Our approach investigates whether visual LLMs can accurately rank candidate images based on their relevance to a given textual compound used in a reference sentence. Since the compounds can assume either an idiomatic or a literal meaning, disambiguating their usage is crucial for meaningful image ranking. To address this, we propose a Chain-of-Thought (CoT) framework that integrates explicit reasoning steps to improve the ranking process. Figure 1) depicts the proposed solution.

As a starting point, we evaluate off-the-shelf visual LLMs by providing them with a sentence, the target compound expression, and five candidate images, prompting them to rank the images based on how well they reflect the compound's meaning in context. The *Standard Inference* block in Figure 1 refers to this approach. This setup allows us to assess whether these models can naturally align visual content with textual semantics.

We introduce a stepwise CoT framework incorporating structured reasoning into the ranking process to improve performance. The first step involves text-based idiomaticity classification, where a standard text-only LLM is used to determine whether the compound is being used literally or idiomatically in the given sentence. This binary classification provides crucial disambiguation before any visual reasoning takes place. Once the compound's usage is classified, this information is explicitly injected into the visual LLM's prompt,

helping the model contextualize the ranking task correctly. Equipped with this knowledge, the visual LLM is then asked to rank the images, now with a clearer understanding of whether the compound should be interpreted in a literal or idiomatic sense. The *Idiomatic Classification Injection* block in Figure 1 describes this strategy.

To increase the quality of this approach, we also propose an iterative CoT framework, where the visual LLM undergoes a multi-step reasoning process before producing a final ranking. This method retains the idiomaticity classification step, where the visual LLM explicitly informs whether the compound is used literally or idiomatically. Then, the model is prompted to extract key visual elements from each image that may be relevant to the compound's meaning. This step allows the model to focus on meaningful visual details that align with the expected interpretation. After identifying these elements, the model is instructed to rank the images based on the extracted features and the previously injected idiomaticity classification. Notably, the ranking prompt is adapted depending on whether the compound is used literally or idiomatically, ensuring that images are evaluated based on the correct semantic perspective. This technique is depicted in the *Chain-of-Thought* block of Figure 1. Each arrow represents a subsequent interaction in the visual LLM chat.

Finally, we incorporate a few-shot learning approach to improve idiomaticity-aware image ranking further. This approach provides the visual LLMs with in-context examples before performing the ranking task, where each example follows the entire Chain-of-Thought pipeline. The *Chain-of-Thought + Few Shot* block in Figure 1 visually describes this technique.

## 4 Experimental Results

### 4.1 Dataset

The AdMIRe challenge organizers released a dataset designed for understanding idiomatic expressions in a multimodal context. Each sample consists of a reference sentence, a compound expression used in the sentence, and five candidate images with their relative captions that could represent the compound's meaning within the given context.

The dataset is divided into three subsets:

- *Train Set*, it contains 70 samples, enriched with target annotations, including the compound's usage type (literal or idiomatic) and the image ranking based on its relevance to the compound's meaning.
- *Test Set*, contains 15 samples, released without target annotations at challenge time.
- *Extended Test Set*, it expands the previous Test Set with 100 additional samples, offering a more extensive evaluation benchmark.

Given the relatively small size of these sets, training a model could be challenging. However, the dataset still provides enough diverse examples to explore effective CoT prompting strategies and construct heterogeneous in-context learning demonstrations, allowing for a meaningful analysis of multimodal idiomaticity understanding.

### 4.2 Experimental Setup

To evaluate the effectiveness of our Chain-of-Thought (CoT) framework in multimodal idiomaticity understanding, we conduct experiments using several visual LLMs:

- **Qwen2.5-VL** (Bai et al., 2025), an open-source vision-language model that integrates visual perception with large-scale text generation. We employ the 7B instruction-tuned version of this model[2].
- **Gemini Flash (F)** (Team et al., 2023), both 1.5 and 2.0 versions, a Google proprietary multimodal model designed for fast inference while maintaining strong performance across vision-language tasks[3].
- **Gemini Flash Thinking (FT)** (Team et al., 2023), version 2.0 only, a variant of Gemini Flash designed to enhance complex reasoning, making it particularly relevant for structured multimodal reasoning tasks[4].

As a baseline, we frame the task as a text-to-image retrieval problem, using the large[5] version of CLIP (Radford et al., 2021) to measure the semantic similarity between the sentence containing the compound and the five candidate images, ranking them accordingly.

In our CoT approach, we first determine whether the compound in the sentence is used in a literal

[2]https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
[3]*Gemini 2.0 Flash Exp* on Gemini API
[4]*Gemini 2.0 Flash Thinking Exp 01-21* on Gemini API
[5]https://huggingface.co/openai/clip-vit-large-patch14

| Type | Model | CLS | CoT | Few Shot | Test Set | | Extended Test Set | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Top 1 Accuracy | DCG Score | Top 1 Accuracy | DCG Score |
| Multimodal | CLIP | - | - | - | 0.40 | 2.71 | 0.44 | 2.82 |
| | Qwen2.5-VL | - | - | - | 0.40 | 2.70 | 0.56 | 2.97 |
| | | ✓ | - | - | 0.53 | 2.80 | 0.58 | 3.00 |
| | | ✓ | ✓ | - | 0.27 | 2.41 | 0.33 | 2.62 |
| | | ✓ | ✓ | ✓ | 0.40 | 2.69 | 0.42 | 2.75 |
| | Gemini 1.5 Flash | - | - | - | 0.67 | 2.99 | 0.53 | 2.89 |
| | | ✓ | - | - | 0.60 | 2.95 | 0.53 | 2.89 |
| | | ✓ | ✓ | - | 0.73 | 3.34 | 0.64 | 3.15 |
| | | ✓ | ✓ | ✓ | 0.73 | 3.28 | 0.77 | 3.33 |
| | Gemini 2.0 Flash | - | - | - | 0.60 | 3.10 | 0.73 | 3.24 |
| | | ✓ | - | - | 0.60 | 3.07 | 0.75 | 3.27 |
| | | ✓ | ✓ | - | 0.80 | **3.40** | 0.69 | 3.25 |
| | | ✓ | ✓ | ✓ | 0.73 | 3.23 | **0.88** | **3.45** |
| | Gemini 2.0 Flash Thinking | - | - | - | 0.73 | 3.25 | 0.76 | 3.29 |
| | | ✓ | - | - | **0.87** | 3.35 | 0.73 | 3.24 |
| | | ✓ | ✓ | - | 0.73 | 3.23 | 0.81 | 3.36 |
| | | ✓ | ✓ | ✓ | **0.87** | **3.40** | 0.87 | 3.40 |
| Text | CLIP-Text | - | - | - | 0.47 | 2.69 | 0.49 | 2.83 |
| | Gemini 2.0 Flash Thinking | ✓ | ✓ | ✓ | **0.73** | **3.17** | **0.78** | **3.28** |

Table 1: Model Performance Comparison for both Multimodal and Text-only approaches. Best results are reported in bold

or idiomatic sense. To ensure high-quality classification, we rely on the best-performing LLM to generate this classification label. In our case, Gemini 2.0 FT is identified as the candidate compound usage type classifier (i.e, if the compound usage in the sentence is idiomatic or literal), providing the best classification result on the AdMIRe *Train Set*, with an accuracy of 90%. The resulting idiomaticity class is then injected as prior knowledge into the prompts of all tested visual LLMs.

Regarding the few-shot approach, we randomly select three different train samples as in-context examples for both the literal and the idiomatic use case. These samples undergo the same CoT pipeline described in Section 3. For the Qwen model, we decrease the number of in-context examples to one, due to memory constraints.

We also apply the proposed framework to the text-only version of the AdMIRe challenge. In the absence of visual input, we maintain identical prompt formats as delineated in Section 3, substituting image inputs with their corresponding textual captions provided in the dataset.

All tested models have been evaluated using the official AdMIRe competition test sets, denominated as *Test Set* and *Extended Test Set*, and metrics, i.e., top 1 accuracy and Discounted Cumulative Gain (DCG) score. In detail, Top-1 accuracy measures the model's ability to select the most appropriate image from five candidates. On the other hand, the DCG score evaluates the quality of the entire ranking of these images, providing insight into the model's overall ordering capabilities.

### 4.3 Results

Table 1 shows the obtained results. All employed visual LLMs in their default inference setting were prompted to directly provide a rank of the images and overcome the CLIP baseline, with Gemini 2.0 FT achieving the highest performance.

The three techniques constituting our framework, i.e., idiomatic classification, CoT, and few-shot learning, are evaluated as incremental steps. Injecting the result of idiomatic classification (CLS: ✓) only into the prompting of visual LLMs does not produce relevant performance changes and the results reflect those of the standard approach, with a slightly improved performance on the *Extended Test Set*. However, Qwen and Gemini 2.0 FT performance only increases on the *Test Set*. Although this improvement affects only two out of four tested models, it can be attributed to the composition of this specific evaluation set. The *Test Set* presumably contains more ambiguous usages of the com-

pounds, making the idiomatic classification result a piece of valuable information to address the ranking task. Notably, Gemini 2.0 FT obtains the higher Top-1 accuracy value on the *Test Set* with this approach.

The introduction of Chain of Thought (CoT) without adopting few-shot learning produces model-dependent effects with notable differences between the test and extended test sets. We can notice a significant decrease in the Qwen performance on both evaluation sets. This is probably due to the model size, which is too small to handle the entire reasoning pipeline completely. On the other hand, this approach is beneficial for all the Gemini family models: Gemini 2.0 FT improves on the *Extended Test Set*, Gemini 2.0 F improves on the *Test Set*, and Gemini 1.5 F improves on both evaluation sets. Noteworthy that Gemini 2.0 F obtains the higher DCG value on the *Test Set* with this approach.

Moving forward, including a few-shot learning technique (Few Shot: ✓) in our CoT pipeline leads to the best results. For Gemini 1.5 F, this combination maintains the improvements obtained using CoT only on *Test Set*, while substantially improving *Extended Test Set* performance. Gemini 2.0 F performance slightly decreases from CoT-only to CoT plus Few Shot on the *Test Set* but achieves its peak performance on the *Extended Test Set*, representing the best performance among all models. This combination of CoT and in-context learning allows Gemini 2.0 FT to achieve the highest overall results on both metrics for the *Test Set*. Also, the results on the *Extended Test Set* improve significantly, settling slightly below those of the Gemini 2.0 F. Moreover, thanks to introducing few-shot learning, Qwen can better understand the previous CoT pipeline, partially restoring its performance.

Although the effect of the proposed steps seems to be model-dependent, the results demonstrate the effectiveness of employing both CoT and few-shot learning, which always lead to the best result.

Finally, we use the best-performing model overall, i.e., Gemini 2.0 FT, to evaluate the proposed approach in a text-only fashion and compare it with a baseline model, i.e, CLIP-Text. The results reported in the bottom section of Table 1 provide valuable comparative insights. The textual encoder of CLIP performs similarly to its multimodal counterpart. Accordingly to the multimodal case, Gemini 2.0 FT achieves substantial improvement. However, it achieves lower metric values than its multimodal implementation, indicating that the visual modality provides essential information to accurately associate idiomatic expressions with the corresponding meaning.

## 5 Conclusions

In this work, we investigated the role of multimodal reasoning in idiomaticity understanding, introducing a novel Chain-of-Thought (CoT) framework to enhance the ranking of images based on their association with idiomatic expressions.

We observed that standard prompting of Visual LLMs is insufficient for correctly ranking images according to idiomatic meanings. The joint introduction of idiomatic classification, CoT reasoning, and few-shot learning consistently led to the best results, proving the effectiveness of step-by-step reasoning and in-context learning for this task. Furthermore, we evaluated the extent to which visual components contribute supplementary information that enhances task resolution, assessing its relevance.

Future work could explore more advanced prompting strategies, model fine-tuning, and larger-scale idiomaticity datasets to enhance the robustness of multimodal idiomaticity understanding further.

## Acknowledgments

This research has been carried out by the SmartData@PoliTO center for Big Data technologies, the HPC@POLITO Academic Computing Center. This study was partially carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013), as well as by the European Union's Horizon Europe research and innovation program EFRA (Grant Agreement Number 101093026). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

## References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye,

Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

Lorenzo D'Amico, Davide Napolitano, Lorenzo Vaiani, Luca Cagliero, et al. 2023. Polito at multi-fake-detective: Improving fnd-clip for multimodal italian fake news detection. In *EVALITA*.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Bulbul Hajiyeva. 2024. Challenges in understanding idiomatic expressions. *Acta Globalis Humanitatis et Linguarum*, 1(2):67–73.

Shreyas Kulkarni, Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2024. A report on the FigLang 2024 shared task on multimodal figurative language. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 115–119, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2024. Rolling the dice on idiomaticity: How llms fail to grasp context. *Preprint*.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18798–18806.

Davide Napolitano, Lorenzo Vaiani, and Luca Cagliero. 2024. On leveraging multi-page element relations in visually-rich documents. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 360–365. IEEE.

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.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. SemEval-2023 task 1: Visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2227–2234, Toronto, Canada. Association for Computational Linguistics.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. V-flute: Visual figurative language understanding with textual explanations. *arXiv preprint arXiv:2405.01474*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, ZHUOFAN ZONG, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 8612–8642.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Lorenzo Vaiani, Luca Cagliero, and Paolo Garza. 2023. PoliTo at SemEval-2023 task 1: CLIP-based visual-word sense disambiguation based on back-translation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1447–1453, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.