Multimodal Chain-of-Thought Prompting for Metaphor Generation

Sofia Lugli[1,*], Carlo Strapparava[2]

[1]University of Trento, Italy
[2]Fondazione Bruno Kessler, Trento, Italy

Abstract

This paper introduces an exploratory approach in the field of metaphorical and visual reasoning by proposing the Multimodal Chain-of-Thought Prompting for Metaphor Generation task aimed to generate metaphorical linguistic expressions from non-metaphorical images by using the multimodal LLaVA 1.5 model and the two-step approach of multimodal chain-of-thought prompting. The generated metaphors were evaluated in two ways: using BERTscore and by five human workers on Amazon Mechanical Turk. Concerning the automatic evaluation, each generated metaphorical expression was paired with a corresponding human metaphorical expressions. The overall BERTscore was the following: precision= 0.41, recall= 0.43, and F1= 0.42, suggesting that generated and human metaphors might not have captured the same semantic meaning. The human evaluation showed the model's ability to generate metaphorical expressions, as 92% of them were classified as metaphors by the majority of the workers. Additionally, the evaluation revealed interesting patterns in terms of *metaphoricity*, *familiarity* and *appeal* scores across the generated metaphors: as the metaphoricity and appeal scores increased, the familiarity score decreased, suggesting that the model exhibited a certain degree of creativity, as it has also generated novel or unconventional metaphorical expressions. It is important to acknowledge that this work is exploratory in nature and has certain limitations.

Keywords

metaphor generation, large language models, pragmatics, creativity, multimodality

1. Introduction

The scope of this paper is to introduce an alternative approach to multimodal metaphor generation. As metaphors are not only pervasive in language but also in everyday life, influencing our thoughts and actions [1], and as human meaning representations relies on multiple modalities [2], it became relevant to study metaphors in more than one modality, in particular in the vision domain. Recent research has indeed explored multimodal metaphors generation in a variety of ways: from visual metaphor to literal language [3, 4, 5]; and from metaphorical language to visual metaphor [3, 6]. Nevertheless, the common aspect across these studies is that the metaphorical quality was already present either in the linguistic or in the visual input employed. Therefore, this paper proposes an alternative approach that involves generating metaphorical linguistic expressions from non-metaphorical images, which lack inherent metaphorical qualities. To accomplish this, we employed the new multimodal model LLaVA 1.5 [7] and adopted a two-step approach known as multimodal chain-of-thought prompting [8]: given the first prompt, the model generates the content of the picture; then, the model is provided with both the generated output and a specific prompt to facilitate metaphor generation. The metaphors generated by the model were evaluated through BERTscore [9] and by human workers on Amazon Mechanical Turk. The results show the model's ability to generate metaphorical expressions, with 92% of the generated expressions being classified as metaphors. Additionally, the evaluation revealed interesting patterns in terms of the metaphoricity, familiarity and appeal scores of the generated expressions. Interestingly, as the metaphoricity score increases, the familiarity score decreases while the appeal score increases. This suggests that the model was able to create novel or uncommon metaphorical expressions which may differ from the more conventional metaphors, which the evaluators might have been more familiar with. Despite being less familiar, the metaphorical expressions were preferred over the non-metaphorical ones. It is important to acknowledge that this is an exploratory work, which aims to offer a different approach in multimodal metaphor generation. As such, it is essential to point out the presence of some limitations, in particular concerning the choice of the visual inputs and the constraints of human evaluation.

2. Background

2.1. Metaphor Theory

For most people, metaphor is merely a rhetorical device restricted to poetic language; however, according to the Conceptual Metaphor Theory (CMT) [1] metaphor is per-

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
sofia.lugli@studenti.unitn.it (S. Lugli); strappa@fbk.eu (C. Strapparava)

vasive in everyday language, playing a significant role in communication, cognition and decision making. More precisely, we talk about *conceptual metaphor* and *linguistic metaphor*. Conceptual metaphors consist of systematic sets of mappings across conceptual domains, whereby a target domain, which is usually a more abstract and complex concept, is partly structured in terms of a different source domain, which usually defines a more concrete and common concept. Conceptual metaphors are then reflected in our everyday language by a wide variety of linguistic metaphors. For instance, ARGUMENT IS WAR is a conceptual metaphor, where ARGUMENT is the target domain and WAR is the source domain; examples of its linguistic metaphors are e.g. Your claims are *indefensible*. He *attacked every weak point* in my argument. You disagree? Okay, *shoot*! [1]. Some of these metaphorical mappings can be defined as *conventional* metaphors, as they are so deep-rooted in our everyday thought and language that they might have become the dominant way of framing a specific concept, and they represent the *commonsense* [10]; while other metaphorical mappings, i.e. *novel* metaphors, are more creative, and they are not (yet) used in everyday discourse, but may become conventionalized if frequently used.

2.2. Related Works

Over the past years, NLP research has been focusing on literal and lower-level linguistic information, while humans excels at high-level semantic task, involving also the use of figurative language [11]. Moreover, statistical corpus analysis [12] indicates that in corpora, metaphors occur in approximately one-third of the sentence. Therefore, metaphor gradually became an important topic in computational linguistics and NLP. Numerous studies have been conducted to investigate metaphors, resulting in three main sub-tasks: metaphor identification [11, 13, 14, 15], metaphor interpretation [16, 17, 18], and metaphor generation [19, 20, 21].

As human meaning representations rely not only on linguistic exposure, but also on perceptual system and sensory-motor experience, [2, 22]; and as metaphors are not merely a matter of language but also of thought and action [1], it became relevant to study metaphors through different modalities. In NLP, the shift towards multimodality happened once computational approaches started adding sensory and contextual features which led to a better performance in metaphor processing [23, 24]. Because of the grounded nature of metaphors, metaphors can occur in different modalities: visual and multimodal metaphors are typically used in mass media communication (e.g., advertising, newspaper) [25]. Visual metaphors are monomodal and expressed through vision, whereas multimodal metaphors are expressed at least through two modalities. Compared to textual metaphors, there has been less research in computational modelling of visual and multimodal metaphors, in particular works accounting for metaphor localization, understanding and generation [26, 27, 5, 4]. In particular, [3] introduced MetaCLUE, a collection of vision tasks on visual metaphor which enables comprehensive evaluation and development of visual metaphor research. Concerning metaphor generation, [3] proposed a task that involves generating an image that effectively conveys the metaphorical message provided as the text prompt; however, the generated images perform poorly compared to real images in conveying metaphorical messages. Additionally, [27] proposed an alternative task for generating visual metaphors from linguistic metaphors using Chain-of-Thought prompting, showing improvements in the quality of visual metaphors generated by diffusion-based text-to-image models. Nevertheless, the common aspect across these studies is that the metaphorical quality was already present either in the textual or in the visual input employed. Interestingly, [28] and [29] dealt with literal images and textual metaphors; however their tasks focused on association between the text and images, rather than on metaphor generation. Therefore, this paper aims to propose an alternative approach involving generating metaphorical linguistic expressions from non-metaphorical images, which lack inherent metaphorical qualities.

2.3. Chain-of-Thought Prompting

The advent of large language models has inevitably changed the NLP field [30], in particular they opened the prospect to the new paradigm of "prompt-based learning" [31]. [30] introduced the concept of chain-of-thought (CoT) prompting, which improves the ability of large language models to perform complex reasoning tasks by employing intermediate reasoning steps. They combined this approach with few-shot prompting (Few-shot-CoT), which enables the language model to generate chains of thought when examples of those are provided. Another approach, known as Zero-shot-CoT [32] consists in adding the simple prompt *Let's think step by step* to the original prompt. The advantage of this method is that it eliminates the need for hand-crafted few-shot examples, resulting in greater versatility. Recently, [8] introduced a multimodal chain-of-thought prompting approach (Multimodal-CoT), which incorporates language (text) and vision (images) modalities into a two-stage framework. The rationale generation and answer inference are separated in two different steps, allowing the answer inference to benefit from well-generated rationales that are based on multimodal information.

3. Experimental Setup

All the data used and the complete results obtained are publicly available at the following repository: https://github.com/SofiaLugli/Multi_COT_meta_gen.git.

3.1. Model

For the purpose of this study, we employed the new multimodal model LLaVA 1.5 (Large Language and Vision Assistant) [7] which is the next iteration of LLaVA [33], considered as the first attempt to use language-only GPT-4 to generate multimodal language-image instruction-following data. LLaVA 1.5 is a end-to-end trained large language model combining a pre-trained CLIP-ViT-L-336px visual encoder with an MLP projection [34] and large language model Vicuna [35] for general purpose visual and language understanding. The model achieved new SoTA performance across 11 benchmarks, thanks to new academic-task-oriented VQA data with simple response formatting prompts. One of the main reason for choosing this model is its impressive multimodal chat abilities; additionally, it is worth noting it is the first open-source project to GPT-V alternative. More precisely, we used the llava-v1.5 13B-4bit and the parameters were set as follows: temperature=0.2, max_new_tokens=1024.[1]

3.2. Dataset Collection

In order to select the metaphors for our research, we retrieved 300 conceptual metaphors from the MetaNet Metaphor Wiki, [2] a comprehensive repository of conceptual metaphors based on years of research on the Conceptual Metaphor Theory. These metaphors follow the standard format, where a target domain is compared to a source domain, e.g., ACHIEVING POWER IS MOVING UPWARDS, CANCER IS A JOURNEY, ENVIRONMENTAL HARM IS PHYSICAL INJURY. To ensure an effective visual representation for the metaphors, we collected two images for each metaphor: one representing the target domain and the other representing the source domain. Given the fact that "LLaVA-1.5 is not yet capable of processing multiple images" [7], for each metaphor, the two images corresponding to the two domains have been pasted together in one image with the target domain image at the top and the source domain image at the bottom. The images were sourced from Google Image and they vary in style, ranging from realistic to cartoon-like pictures.

[1]https://github.com/haotian-liu/LLaVA
[2]https://metaphor.icsi.berkeley.edu

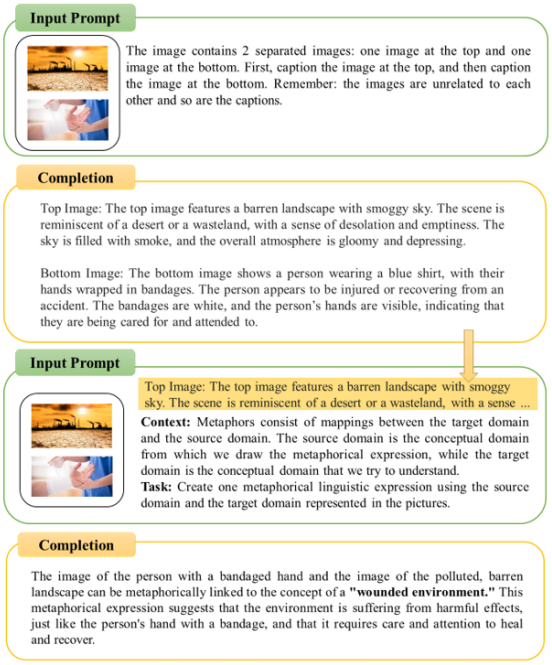

Figure 1: Visual representation of the task for the metaphor ENVIRONMENTAL HARM IS PHYSICAL INJURY.

3.3. Task

In this section, we will provide an explanation of the task at hand. We propose an alternative approach for multimodal metaphor generation by using both language and non-metaphorical visual inputs. Our approach is based on the multimodal CoT prompting technique [8, 36]. Our approach follows a two-step process, as shown in Fig.1. Firstly, the model is fed with the non-metaphorical image containing both the images of the target and source domains. The model's task is to generate captions describing each of these images. We provide the prompt: `The image contains 2 separated images: one image at the top and one image at the bottom. First, caption the image at the top, and then caption the image at the bottom. Remember: the images are unrelated to each other and so are the captions.` Once the content of the picture has been generated, it is then used as input for the second prompt, which involves generating metaphorical expressions based on the source and target domains. For this, we employ the following prompt: `Context: Metaphors consist of mappings between the source domain and the target domain.The source domain is the conceptual domain from which we draw the metaphorical expression, while the target domain is the conceptual domain that we try`

Metaphoricity Agreement	Generated Metaphor	Conceptual Metaphor
5	*Wounded environment*	ENVIRONMENTAL HARM IS PHYSICAL INJURY
	House of thoughts	MIND IS A BUILDING
	She is wearing a bandage on her heart	PSYCHOLOGICAL HARM IS PHYSICAL INJURY
4	*Climbing the stairs of success*	ACHIEVING POWER IS MOVING UPWARDS
	Fighting the battle against cancer	CANCER PATIENT IS PHYSICAL COMBATANT
	The burden of the virus is weighing heavily on the man's shoulders	DISEASES ARE BURDENS
3	*Digesting knowledge*	ACQUIRING IDEAS IS EATING
	Battle of words	ARGUMENT IS WAR
	Walking down a road to recovery	CANCER IS A JOURNEY
2	*A financial heart attack*	ADDRESSING ECONOMIC PROBLEMS IS TREATING AN ILLNESS
	Embracing the warmth of friendship	AFFECTION IS WARMTH
	Their love was as hot as the sun	PASSION IS HEAT
1	*Shaking hands over a book of contracts is like a marriage of business and legal agreements*	AGREEMENT IS PHYSICAL PROXIMITY
	A family's journey through life, with the man as the guide and the woman and child as his companions	BEING IN A LOW SOCIAL CLASS IS BEING LOW ON A SCALE
	A political body is like a human body	GOVERNMENT IS A PERSON

Table 1
Some examples of metaphorical linguistic expressions generated by the model and their corresponding conceptual metaphors. The first column shows the workers agreement on the metaphoricity (with 5 being the highest and 1 the lowest) when evaluating the generated expressions.

`to understand. Task: Create one metaphorical linguistic expression using the source domain and the target domain represented in the pictures.` For instance, Fig. 1 provides a visual representation of the task in the case of the conceptual metaphor ENVIRONMENTAL HARM IS PHYSICAL INJURY. In this example, the model was able to successfully generate two distinct captions for the target domain image and the source domain image. Subsequently, given the second prompt, the model was able to generate a corresponding metaphorical expression such as *wounded environment*. Additionally, the model provided a correct explanation of the new generated metaphor. To prove the utility of the method, the task was performed on a subset of the dataset without using CoT prompting. In this case, only the second prompt of generating the metaphor was used, without first the image captioning prompt. The results were less satisfactory. For instance, for the conceptual metaphor ENVIRONMENTAL HARM IS PHYSICAL INJURY, the model generated the expression *The sun shines brightly over the barren landscape, illuminating the industrial complex like a beacon of hope.* This output, compared to the metaphor generated through CoT prompting (e.g., *wounded environment*), does not involve a metaphor and fails to consider the images of both source and target domains.

3.4. Evaluation setup

The evaluation of the generated metaphorical expressions has been conducted in two ways: through BERTscore and by five human workers through Amazon Mechanical Turk.

Concerning the automatic metaphor evaluation trough BERTscore [9], each generated metaphorical expression (*candidate*) was paired with a corresponding human metaphorical expression retrieved from MetaNet (*reference*), which provides real world examples of linguistic metaphors, sourced from various contexts (e.g., newspapers, books, etc.). However, the MetaNet does not provide examples for all the metaphors in their repository, as such 75 metaphors were excluded from this evaluation, as they lacked example references. Compared to traditional commonly used evaluation metrics [37, 38, 39], which relied on n-gram count, BERTscore [9] computes token similarity using contextualized token embeddings, which have been shown to be effective for paraphrase detection [40]. It then calculates Recall and Precision, which are combined into an F1 score.

Concerning human evaluation, each generated expression was evaluated by five Amazon Mechanical Turk workers from English speaking countries (Australia, Canada, Ireland, New Zealand, United Kingdom, and United States). The workers were required to had an approval rate greater than 95% on 1000 prior approved HITs; their reward was $0.12 per task. To ensure the quality of the evaluation, the workers were given background knowledge regarding the Conceptual Metaphor Theory, as well as positive and negative examples for the task. The workers had to chose whether the generated linguistic expression (e.g., *Wounded environment*) could be accepted as a linguistic metaphor for its corresponding conceptual metaphor (e.g., ENVIRONMEN-

TAL HARM IS PHYSICAL INJURY) with the following Yes or No question: *Can the linguistic expression be considered as a linguistic metaphor for the provided conceptual metaphor?*. Additionally, they were asked other two yes/no questions regarding the familiarity and appeal of the expressions: *Have you encountered this linguistic expression before?* and *Is this linguistic expression appealing to you?*. To consider an expression as metaphorical, it had to be evaluated as such by at least three out of the five workers. It is worth noting that it was not mentioned that the metaphors were not human-generated in order to prevent any potential bias.

4. Results

In this section, we present the results derived from the automatic and the human evaluation. Regarding the automatic evaluation, it is important to note that, overall the BERTscore between the generated and the human metaphors was low, the average scores were the following precision= 0.41, recall= 0.43, and F1= 0.42. The highest score was achieved in the metaphor SAD IS DOWN, where the generated metaphor *feeling down in the dumps* and the real-world example *I'm feeling down* achieved the scores precision= 0.67, recall= 0.84, and F1= 0.74. The low BERTscore suggests that there is a discrepancy between the model's generations and human examples, which may indicate that the generated metaphors may not be capturing the same semantic meaning as the human-generated ones. Additionally, this might be due to the difference in contexts. Human-generated metaphors often reference real-world examples, including real people and events; whereas the generated metaphors tend to be more generic and less nuanced compared to the human-generated ones. Moreover, another reason behind the low BERTscore is that, while robust, it might still have limitations in capturing the subtle and nuanced differences and similarities in metaphorical language, which are typically subjective and context-dependent.

Concerning the human evaluation by five MTurk workers, it was conducted on three criteria: *metaphoricity*, *familiarity* and *appeal* of the generated linguistic expressions. First of all, the expressions obtained a metaphoricity mean score of 3.8, which means that, on average, the generated expressions were considered as metaphorical by the majority of the workers. A total of 92% of the linguistic expressions were evaluated as metaphors by at least three workers. Among these, 92 expressions were unanimously recognized as metaphors by all five evaluators, for instance *Wounded environment* generated for the conceptual metaphor ENVIRONMENTAL HARM IS PHYSICAL INJURY. Additional examples of the generated expressions and their corresponding metaphoricity agreement scores can be found in Table 1, while the complete results are available in our repository. Furthermore, 108 expressions were considered as metaphors by four workers and 76 expressions by three workers. Out of the 300 metaphors, only 24 generated expressions were not evaluated as metaphors as they were recognized as metaphors by either two (21 expressions) or only one worker (3 expressions). It is worth noting that none of the expressions were evaluated as non metaphors by any of the workers. These results can be considered as positive, suggesting that LLaVA 1.5 successfully generated metaphorical expressions from non-metaphorical visual inputs.

Now let us examine the remaining two criteria. In terms of *familiarity*, the average score is 2.95, and 67% of the expressions were considered as familiar by at least three workers. Only 22 expressions were considered as familiar by all five workers; for instance the expression *A journey through life* for PROGRESSING THROUGH LIFE IS MOVING ALONG A PATH. Additionally, 73 metaphors were familiar to four evaluators, while 106 expressions were familiar to three evaluators. On the other hand, there were 71 metaphors that were not familiar to all but two workers, 24 that were only familiar to one worker, and 4 that were not familiar to any worker. In other words, out of 300 expressions, 99 expressions can indeed be considered unfamiliar, as they are only rated as familiar by two or fewer workers. These findings regarding familiarity indicate that the model generated not only familiar expressions but also novel, or uncommon expressions. This suggests that the model exhibits a certain degree of creativity in this task.

Moving on to the *appeal* criterion, the average score is 3.32, and 78% of the generated expressions were liked by at least three workers. Among the expressions, 37 were liked by all five workers, e.g., *Walking down a road to recovery* for CANCER IS A JOURNEY. Furthermore, 98 expressions appealed to four workers, 99 to three workers, 57 to two workers and 9 to only one worker. These results indicate that the generated expressions were mostly appreciated.

Let us now examine the distribution of the mean agreement scores for familiarity and appeal in relation to the agreement scores for metaphoricity. As illustrated in Fig. 2, the observed pattern seems to suggest that the mean familiarity and appeal scores exhibit contrasting trends across different metaphoricity scores. Interestingly, as the metaphoricity score increases, the familiarity score decreases while the appeal score increases. Metaphoricity scores 5 and 1 represent the extremes, with distinct differences in both familiarity and appeal. For the generated metaphorical expressions evaluated as such by all five workers, the mean score of familiarity is 2.92 and of appeal is 3.6; whereas for the expressions considered metaphorical only by one worker, the mean familiarity score is 3.67 and appeal is 3.0. With the exception of the

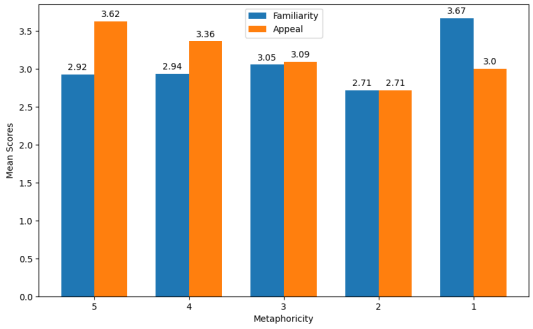

Figure 2: Mean familiarity and appeal scores for each metaphoricity score.

expressions with metaphoricity score 2, which registered the lowest score (2.71) both for familiarity and appeal, the pattern seems to indicate that metaphoric expressions with higher metaphoricity scores tend to have lower familiarity and higher appeal. This means that the evaluators found the literal generated expressions (metaphoricity scores 1 and 2) to be more familiar compared to the metaphorical ones. Hence, the results suggest that the model was able to create novel metaphorical expressions which may differ from the more conventional metaphors, which the evaluators might have been more familiar with. Despite being less familiar, the metaphorical expressions were preferred over the non-metaphorical ones. These findings show that the model exhibited a degree of creativity in metaphor generation, as it generated novel or unconventional metaphorical expressions which where appreciated by human evaluators.

5. Conclusion

This study aimed to explore an alternative approach for multimodal metaphor generation using the new LLaVA 1.5 model and Multimodal-CoT prompting. The results showed the model's ability to generate metaphorical expressions when provided with both linguistic and visual inputs which lack inherent metaphorical qualities. Additionally, the evaluation revealed interesting patterns across the metaphoricity, familiarity and appeal scores of the generated expressions. The model exhibited its creativity, as it generated novel or unconventional metaphorical expressions, which were also preferred over non-metaphorical ones. It is important to state again that this is an exploratory work with some limitations. One limitation to consider is the choice of the images used in the study. As manually selected from Google Image, their quality may influence the quality of the captions and metaphors generated by the model. Another limitation to consider is the subjectivity of the evaluation process, it is possible that Amazon MTurk workers may lack the necessary sensitivity and background knowledge to accurately recognize and evaluate metaphorical expressions, despite the instructions included background information about metaphor. Future works should aim to address these limitations by selecting more accurate images, as well as incorporating more diverse and expert annotators.

Despite these limitations, the task show promising results for future research in the field of metaphorical and visual reasoning.

Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] G. Lakoff, M. Johnson, Metaphors We Live By, University of Chicago Press, 2008. URL: https://books.google.it/books?id=r6nOYYtxzUoC.

[2] L. W. Barsalou, Grounded cognition, Annu. Rev. Psychol. 59 (2008) 617–645.

[3] A. R. Akula, B. Driscoll, P. Narayana, S. Changpinyo, Z. Jia, S. Damle, G. Pruthi, S. Basu, L. Guibas, W. T. Freeman, et al., Metaclue: Towards comprehensive visual metaphors research, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23201–23211.

[4] E. Hwang, V. Shwartz, Memecap: A dataset for captioning and interpreting memes, arXiv preprint arXiv:2305.13703 (2023).

[5] B. Xu, T. Li, J. Zheng, M. Naseriparsa, Z. Zhao, H. Lin, F. Xia, Met-meme: A multimodal meme dataset rich in metaphors, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2887–2899.

[6] T. Chakrabarty, Y. Choi, V. Shwartz, It's not rocket science: Interpreting figurative language in narratives, Transactions of the Association for Computational Linguistics 10 (2022) 589–606.

[7] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, 2023. arXiv:2310.03744.

[8] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, A. Smola, Multimodal chain-of-thought reasoning in language models, arXiv preprint arXiv:2302.00923 (2023).

[9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[10] E. Semino, Metaphor in Discourse, Metaphor in Discourse, Cambridge University Press, 2008. URL: https://books.google.it/books?id=QT1uilVRDTYC.

[11] E. V. Shutova, Computational approaches to figurative language, Technical Report, University of Cambridge, Computer Laboratory, 2011.

[12] G. J. Steen, A. G. Dorst, J. B. Herrmann, A. A. Kaal, T. Krennmayr, Metaphor in usage, Cognitive Linguistics (2010).

[13] Y. Tsvetkov, L. Boytsov, A. Gershman, E. Nyberg, C. Dyer, Metaphor detection with cross-lingual model transfer, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 248–258.

[14] G. Gao, E. Choi, Y. Choi, L. Zettlemoyer, Neural metaphor detection in context, arXiv preprint arXiv:1808.09653 (2018).

[15] R. Mao, X. Li, M. Ge, E. Cambria, Metapro: A computational metaphor processing model for text preprocessing, Information Fusion 86 (2022) 30–43.

[16] E. Shutova, Automatic metaphor interpretation as a paraphrasing task, in: Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics, 2010, pp. 1029–1037.

[17] C. Su, S. Huang, Y. Chen, Automatic detection and interpretation of nominal metaphor based on the theory of meaning, Neurocomputing 219 (2017) 300–311.

[18] E. Liu, C. Cui, K. Zheng, G. Neubig, Testing the ability of language models to interpret figurative language, arXiv preprint arXiv:2204.12632 (2022).

[19] T. Veale, Round up the usual suspects: Knowledge-based metaphor generation, in: Proceedings of the Fourth Workshop on Metaphor in NLP, 2016, pp. 34–41.

[20] Z. Yu, X. Wan, How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation, 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), 2019, pp. 861–871.

[21] T. Chakrabarty, X. Zhang, S. Muresan, N. Peng, Mermaid: Metaphor generation with symbolism and discriminative decoding, arXiv preprint arXiv:2103.06779 (2021).

[22] M. M. Louwerse, Symbol interdependency in symbolic and embodied cognition, Topics in Cognitive Science 3 (2011) 273–302.

[23] P. Turney, Y. Neuman, D. Assaf, Y. Cohen, Literal and metaphorical sense identification through concrete and abstract context, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 680–690.

[24] E. Shutova, D. Kiela, J. Maillard, Black holes and white rabbits: Metaphor identification with visual features, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2016, pp. 160–170.

[25] C. Forceville, Pictorial metaphor in advertising, Routledge, 2002.

[26] D. Zhang, M. Zhang, H. Zhang, L. Yang, H. Lin, Multimet: A multimodal dataset for metaphor understanding, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 3214–3225.

[27] T. Chakrabarty, A. Saakyan, O. Winn, A. Panagopoulou, Y. Yang, M. Apidianaki, S. Muresan, I spy a metaphor: Large language models and diffusion models co-create visual metaphors, arXiv preprint arXiv:2305.14724 (2023).

[28] G. Özbal, D. Pighin, C. Strapparava, et al., A proverb is worth a thousand words: learning to associate images with proverbs, in: Proceedings of the 41st Annual Conference of the Cognitive Science Society (CogSci'19), Cognitive Science Society, 2019, pp. 2515–2521.

[29] R. Yosef, Y. Bitton, D. Shahaf, Irfl: Image recognition of figurative language, arXiv preprint arXiv:2303.15445 (2023).

[30] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in Neural Information Processing Systems 35 (2022) 24824–24837.

[31] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

[32] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, 2023. arXiv:2205.11916.

[33] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, arXiv preprint arXiv:2304.08485 (2023).

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. arXiv:2103.00020.

[35] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

[36] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, Q. Liu, et al., Language is not all you need: Aligning perception with language models, arXiv preprint arXiv:2302.14045 (2023).

[37] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[38] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[39] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[40] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).