Evaluating the effectiveness of prompting methods on VQA-AnswerTherapy

Anush Kumar Venkatesh

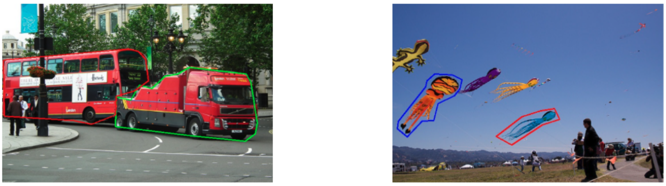
University of Colorado Boulder

Abstract. It is common to have humans provide multiple answers for the same question in the task of VQA. The answer differences come from ambiguity, subjectiveness, etc. In this work we ask the question can vision language foundation model do the same without additional finetuning. Foundation models have proven to produce high zero shot and few shot performance for various downstream applications via prompting. However, their ability to produce multiple answers or mapping answers back to visual evidence remain under-explored. In this work, we probe the models for same. Through our experiments we find that foundation models, as of today, lacks the ability to generate varied answers via prompting. All of our code to replicate our experiments have been released here.

1 Introduction

Visual Question Answering (VQA) [1–3], represents a well-explored challenge in the realm of machine learning, involving the identification of the correct answer to a visual question based on an associated image. The evolution of top-performing models in VQA has transitioned from CNN-RNN architectures [4–6], to transformer-based models [7–9]. This progress owes much to the collaborative efforts within the research community, particularly in constructing high-quality datasets [2, 10–12], often leveraging crowd-sourcing methods. During the dataset creation process, a recurrent observation is the potential for multiple plausible correct answers to be associated with a given visual question. Many of the past studies [13, 14] have aimed to identify why this is the case. A more recent study [12] investigated to test whether the visual evidence for those multiple different answers overlap or not. Notably, their findings reveal that approximately 15% of their constructed dataset exhibits distinct groundings. In other words, annotators provided answers referencing different parts of the image. The variance in answers can be attributed to factors like ambiguous questions, subjective inquiries, and varying levels of answer granularity. Examples illustrating these nuances are presented in Figure 1.

While the pursuit of enhancing models for optimal performance in Visual Question Answering (VQA) persists, a notable shift within the machine learning community has been the growing interest in prompting large language models to perform well on downstream tasks, as opposed to the conventional pre-train, fine-tune, and predict paradigm [15]. The emerging approach involves pre-training,



Question	What type of vehicle is here?
Answer 1	Tow truck
Answer 2	Bus

Question	What color is his kite?
Answer 1	Blue
Answer 2	Orange

Fig. 1. Examples from VQA-AnswerTherapy Dataset

prompting, and then predicting. As the research community increasingly embraces prompting-based methodologies, recent investigations, such as the work by [16], have sought to understand its effectiveness specifically in the context of VQA. Building upon this foundation, our study extends the insights gained from [16] and seeks to ascertain whether pre-trained models can (1) discern whether different answers correspond to the same visual evidence or not, and (2) generate distinct correct answers, from a different perspective, for the same visual question, as noted in Figure 1.

In summary, through this work, our contributions include

1. Use different prompting techniques to investigate whether large pre-trained foundation models such as Blip2 [17], InstructBlip [18], and Flamingo [19] are able to identify the different answers refer to the same visual evidence or answer grounding in the image.
2. Use different prompting techniques to understand whether the same foundation models are capable of providing multiple correct answers for the same visual question.
3. Analyse which prompting technique performs reasonably better at the given two tasks that the rest.

2 Related Work

2.1 Prompting

The recent times have seen a tremendous increase in the popularity of prompting large language models to achieve phenomenal zero shot performance in downstream applications. The process involves introducing a prompt designed to elicit a response in the desired format for the specific downstream task. Common prompting methods include manually creating prompts [20–23], or automating the learning of the best prompt for a given task [24, 25]. Prompts can take the form of either discrete [25–27] or continuous inputs [28, 29]. [15] provide an overview of various prompting strategies. While the concept of prompting initially gained popularity in the textual domain, it has now extended its application to the multimodal domain. For instance, Flamingo [19], MAPL [30] employ few-shot in-context learning to facilitate knowledge transfer to unseen tasks. [16] study the effects of zero and few shot prompting particularly on the VQA domain.

Building on previous research into effective prompting methods, we apply these approaches to assess whether foundation models can provide varied responses to the same visual question and recognize if multiple answers point to the same visual evidence in the image.

2.2 Multiple answers in VQA datasets

Previous work has focused on identifying the factors that can cause human annotators to provide different answers to the same question. For instance, [13] show us the presence of low quality images, invalid questions, subjective questions etc can influence multiple answers. Another work [14] tries to build a model that predicts whether a visual question can elicit disagreement from different annotators. The advantage here is to motivate data collection process to gather consistent data rather than designing algorithms to overcome ambiguity.

As foundation models gain popularity, it is crucial for us to explore their ability to mimic humans and produce answers from a different perspective. In this work, we prompt vision language foundation models to test for the same.

2.3 Visual Language Models for VQA

Popular Visual Language Models (VLMs) like Blip-2 [17], InstructBlip [18], CM3Leon [31], GPT4 [32], and Flamingo [19] have excelled in various tasks, especially Visual Question Answering (VQA). However, research like [16] suggests they struggle with complex compositional questions and adapting to new domains. In our study, we add to the existing rich literature with test results of zero-shot and few-shot answering capabilities to produce multiple outputs. We use the VQA-AnswerTherapy dataset [12] for our experiments, focusing on Blip-2, InstructBlip, and Flamingo as they are publicly available.

3 Methods

Our primary aim is to extend the work by [16] to probe for (1) can vision language foundation models recognize if multiple answers to the same visual question point the same visual evidence or not and (2) can they provide varied correct answers for the same visual question. The study by [16] aims to assess the effectiveness of different prompting methods for the application of VQA. A brief overview of the prompting methods they used are given below.

1. **Standard** - Here the foundation models are prompted with question as is only using the image as context. That is, the model is provided only the image and the visual question is asked in the below format.

   (a) ”Question: {question} Answer:”
2. **Image captions as context** - Here the foundation models are prompted with the question using both image and it’s caption as context. The model is first prompted to generate a caption for the image and subsequently the generated caption is used as additional context to probe for VQA. The prompt format is given below.
   (a) ”Context: {caption} Question: {question} Answer:”
3. **Chain of thought** - [33] show that chain of thought prompting helps improve the performance of foundation models on various downstream tasks. Thus [16] adapt it for VQA using the below format.
   (a) ”Please answer the following question by reasoning step by step. Question: {question} Answer:”

We use the exact same prompting templates mentioned above to test for two different capabilities.

3.1 Recognizing visual evidence for multiple answers

In this test, we include all the answers for the visual question in the prompt and ask the model to identify if answers point to the same visual content in the image. The model is expected to answer ”yes” if that is the case else ”no”. This allows us to gauge the model’s ability to ground different answers to its corresponding location in the image. The prompting templates used for this test is given below.

1. **Standard** - ”Question: {question} Answer: {answer} Question: Do all the given answers for the question point to the same visual content in the image?”
2. **Image captions as context** - ”Context: {caption} Question: {question} Answer: {answer} Question: Do all the given answers for the question point to the same visual content in the image?”
3. **Chain of thought** - ”Please answer the following question by reasoning step by step. Question: {question} Answer: {answer} Question: Do all the given answers for the question point to the same visual content in the image?”

3.2 Generating multiple answers

Contrary to the previous test, here we probe the models ability to provide us different possible answers for the same visual question. This allows us to test whether the model is capable of providing correct answers to the same visual question from multiple perspectives. The prompting templates used for this test is given below.

1. **Standard** - ”Indicate every possible answer to the given question. Question: {question} Answer:”
2. **Image captions as context** - ”Context: {caption} Indicate every possible answer to the given question. Question: {question} Answer:”
3. **Chain of thought** - ”Please indicate every possible answer to the visual question by reasoning step by step. Question: {question} Answer:”

3.3 Models used

We use Blip2 [17], InstructBlip [18] and Flamingo [19] to test for the above mentioned capabilities. Blip2 makes a good choice as it keeps it consistent with the models used by MILA and Blip2 achieved state of the art for VQA. In addition to Blip2, we also use an instruction tuned version of Blip2 which was released as InstructBlip. While Blip2 and InstructBlip help us assess zero shot prompting capability of the model, Flamingo helps us assess few shot prompting capabilities of the model.

Blip2 and InstructBlip has been released for use via Huggingface [34], the original model of Flamingo has not been released, we therefore use a community based open source[1] implementation of the model.

4 Experimental Design

In our work, we evaluate the prompt based methods on the publicly available train and validation datasets for VQA-AnswerTherapy [12]. VQA-AnswerTherapy makes a perfect fit for us as the dataset not only provides multiple valid correct answers to a given visual question but also supplements them with groundings for each of the given answers. Additionally, we are also provided with a binary label that indicates whether the answers point to the same visual content or different. This feature helps us to easily evaluate our models outputs. Table 1 provides us a distribution of datapoints that mark the binary label as ”same” versus that of ”different” across train and validation sets. We choose accuracy as our evaluation metric and it’s calculation for each of the experiments is summarized below.

Binary Label	Val	Train
Same	309	3226
Diff	98	538
Total	407	3764

Table 1. Distribution same versus different labels in VQA-AnswerTherapy Dataset

1. **Recognizing visual evidence for multiple answers -** In this test, we expect the model to output ”yes” if all the answers point to the same visual evidence else ”no”. As the dataset provides this label by default calculating accuracy is straight forward. It is a one to one direct comparision between the model’s output to ground truth.
2. **Generating all answers -** This situation gets tricky since the model might provide sequences with fewer or more words than those in the ground truth,

[1] Open flamingo

making direct comparison challenging. To address this, we suggest calculating the percentage of answers from the ground truth that the model includes in its prediction.For instance, if the model captures all the answers for a given ambiguous visual question, it gets a score of 1. If it includes half of the answers from the ground truth, it gets a score of 0.5, and so on. Finally we calculate the mean over all scores across the dataset to get the final score. For instance, if the model captures all the answers for a given ambiguous visual question, it gets a score of 1. If it includes half of the answers from the ground truth, it gets a score of 0.5, and so on. Finally we calculate the mean over all scores across the dataset to get the final score. Table 2 provides additional examples.

In both of the above mentioned cases, a higher score means a better performance by the models.

	Ground truth			
Visual Question	**Answer 1**	**Answer 2**	**Predicted**	**Score**
What type of vehicle is here?	Tow truck	Bus	Bus and tow truck	1.0
			Bus	0.5
What kind of car is this?	Compact	Hatchback	Compact and hatchback	1.0
			Compact	0.5

Table 2. Sample accuracy calculation for generating multiple answers

5 Experimental Results

We summarize out results in Tables 3, 4 and 5 respectively.

		Recognise Visual Context			**Producing Multiple Answer**		
	Accuracy for Binary Label	**Standard**	**Image captions**	**Chain of thought**	**Standard**	**Image captions**	**Chain of thought**
Val	Overall	0.739	0.729	0.759	0.598	0.625	0.614
	Same	0.974	0.957	1.000	0.623	0.663	0.656
	Diff	0.000	0.102	0.000	0.518	0.508	0.481
Train	Overall	0.839	0.837	0.856	0.569	0.587	0.581
	Same	0.978	0.976	0.996	0.599	0.595	0.590
	Diff	0.001	0.001	0.014	0.416	0.542	0.525

Table 3. Accuracy results for Blip2

		Recognise Visual Context			Producing Multiple Answer		
	Accuracy for Binary Label	**Standard**	**Image captions**	**Chain of thought**	**Standard**	**Image captions**	**Chain of thought**
Val	Overall	0.756	0.751	0.756	0.555	0.531	0.558
	Same	0.987	0.970	0.990	0.599	0.579	0.600
	Diff	0.030	0.061	0.020	0.416	0.380	0.426
Train	Overall	0.841	0.828	0.842	0.304	0.305	0.304
	Same	0.981	0.961	0.981	0.319	0.320	0.319
	Diff	0.005	0.026	0.007	0.215	0.215	0.215

Table 4. Accuracy results for InstructBlip

		Recognise Visual Context		Producing Multiple Answer	
	Accuracy for Binary Label	**Standard**	**Image captions**	**Standard**	**Image captions**
4-shot	Overall	0.442	0.265	0.585	0.546
	Same	0.404	0.061	0.619	0.579
	Diff	0.561	0.908	0.465	0.443
8-shot	Overall	0.245	0.243	0.577	0.537
	Same	0.009	0.003	0.624	0.579
	Diff	0.989	1.0	0.430	0.406

Table 5. Accuracy results on Validation set for Flamingo

5.1 Are models capable of recognizing visual evidence for given answers?

Upon looking at results from Tables 3, 4 and 5 we recognize that it is not the case. We find a standard pattern that the Blip2 and InstructBlip usually answers ”yes” and Flamingo answers ”no” irrespective of the questions asked. This suggests to us that either the models arent able to co-relate the answers back to the image, that is the models lack some form of grounding answers capability. However, we do concede that a deeper exploration of various prompt templates are necessary to solidify the results.

5.2 Which prompting method performs best to recognize visual evidence?

Comparing the results from Tables 3, 4 we slight yet noticeable improvement in the performance of Blip2 and InstructBlip to recognize that answers come from different visual evidences. On the contrary we do not find the same applied in the case of Flamingo. Flamingo inherently has the bias to answer ”no” to all the questions asked. Adding image captions do not seem to help it recognize that some answers do come from the same visual evidence. There seems to an exception however when the standard template is used. In this case, the models randomly guesses between ”yes” and ”no”. For the rest of the cases however, we do not see the performances differing. A surprising result is that chain of thought prompting hurts the model performance more when compared to standard style prompting and prompting with image captions. A similar observation was made by [16]. The authors there also find chain of thought prompting hurts model performance when it comes to VQA.

5.3 Are models able to generate multiple answers for the same question?

Although these state of the art models provide us the correct answer for a given visual question, they aren't able to do so from multiple different perspectives as humans do. This is evident by the fact that the accuracy of all the models, that is BLip2, InstructBlip, and Flamingo stays close to 50%. This suggests that the models either get one correct answer among the given ground truth answer or do not get them correct at all. Observing the datapoints from VQA-AnswerTherapy, we find that most visual question have two different ground truth answers, with very few datapoints containing more than two different ground truth answers. This further explains the accuracy number being close to 50%.

5.4 which prompting method performs best at giving multiple answers?

The results from Tables 3, 4 and 5 suggests that no particular method necessarily improve the model's capability to produce multiple answers. This result is important as it helps us to determine the future direction that we need to take in order to get the models to produce a variety of answers. The language models component Blip2, InstructBlip, and Flamingo is pre-trained to produce results that are closely in line with image captions. The tend to produce results that are most probable. In the future, we aim to either try different sampling techniques or Instruction tune these foundation models to produce varies outputs.

5.5 Does few shot prompting help?

Few shot prompting doesn't seem to help in either of our experiments. Additionally contrary to belief, increasing the number of few shot exemplars hurt the model performance more. This is evident comparing the numbers from Table 5

6 Conclusion

In our study, we showcased the potential of prompting vision language foundation models to discern visual content for different answers and generate diverse responses for a given visual question. However, our experiments revealed subpar performance in these tasks when employing various prompting methods ranging from chain of thought prompting to few-shot prompting. Consequently, we cautiously assert that naively prompting these models for such tasks does not yield satisfactory results. This, in turn, underscores the need for future work to explore and enhance performance in this domain. We see two possible steps. First to use more variety of prompting templates to confidently ascertain our observed results. Second, to Instruction tune these various foundation models to perform well in this domain.

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D.: Vqa: Visual question answering. International Journal of Computer Vision **123** (2015) 4 – 31
2. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. International Journal of Computer Vision **127** (2016) 398 – 414
3. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017) 4971–4980
4. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. ArXiv **abs/1606.00061** (2016)
5. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Conference on Empirical Methods in Natural Language Processing. (2016)
6. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 21–29
7. Lu, K., Fang, B., Chen, K.Y.: A transformer-based cross-modal fusion model with adversarial training for vqa challenge 2021. ArXiv **abs/2106.13033** (2021)
8. Siebert, T., Clasen, K.N., Ravanbakhsh, M., Demir, B.: Multi-modal fusion transformer for visual question answering in remote sensing. In: Remote Sensing. (2022)
9. Yang, Y., Jin, J., Li, D.: A study of visual question answering techniques based on collaborative multi-head attention. 2023 3rd Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS) (2023) 552–555
10. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123** (2016) 32 – 73
11. Gurari, D., Li, Q., Stangl, A., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 3608–3617
12. Chen, C., Anjum, S., Gurari, D.: Vqa therapy: Exploring answer differences by visually grounding answers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (October 2023) 15315–15325
13. Bhattacharya, N., Li, Q., Gurari, D.: Why does a visual question have different answers? In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (October 2019)
14. Gurari, D., Grauman, K.: Crowdverge: Predicting if people will agree on the answer to a visual question. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. CHI '17, New York, NY, USA, Association for Computing Machinery (2017) 3511–3522

15. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing (2021)
16. Awal, R., Zhang, L., Agrawal, A.: Investigating prompting techniques for zero- and few-shot visual question answering (2023)
17. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023)
18. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
19. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Bińkowski, M.a., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., eds.: Advances in Neural Information Processing Systems. Volume 35., Curran Associates, Inc. (2022) 23716–23736
20. Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.: Language models as knowledge bases? In Inui, K., Jiang, J., Ng, V., Wan, X., eds.: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Association for Computational Linguistics (November 2019) 2463–2473
21. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020)
22. Schick, T., Schütze, H.: Exploiting cloze-questions for few-shot text classification and natural language inference. In Merlo, P., Tiedemann, J., Tsarfaty, R., eds.: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, Association for Computational Linguistics (April 2021)
23. Schick, T., Schütze, H.: It's not just size that matters: Small language models are also few-shot learners. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y., eds.: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, Association for Computational Linguistics (June 2021)
24. Shin, R., Lin, C., Thomson, S., Chen, C., Roy, S., Platanios, E.A., Pauls, A., Klein, D., Eisner, J., Van Durme, B.: Constrained language models yield few-shot semantic parsers. In Moens, M.F., Huang, X., Specia, L., Yih, S.W.t., eds.: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, Association for Computational Linguistics (November 2021)
25. Jiang, Z., Xu, F.F., Araki, J., Neubig, G.: How can we know what language models know? Transactions of the Association for Computational Linguistics **8** (2020)

26. Haviv, A., Berant, J., Globerson, A.: BERTese: Learning to speak to BERT. In Merlo, P., Tiedemann, J., Tsarfaty, R., eds.: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, Association for Computational Linguistics (April 2021)
27. Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. In Zong, C., Xia, F., Li, W., Navigli, R., eds.: 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), Online, Association for Computational Linguistics (August 2021)
28. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In Zong, C., Xia, F., Li, W., Navigli, R., eds.: 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), Online, Association for Computational Linguistics (August 2021)
29. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning (2022)
30. Mañas, O., Rodriguez Lopez, P., Ahmadi, S., Nematzadeh, A., Goyal, Y., Agrawal, A.: MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In Vlachos, A., Augenstein, I., eds.: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, Association for Computational Linguistics (May 2023)
31. Yu, L., Shi, B., Pasunuru, R., Muller, B., Golovneva, O., Wang, T., Babu, A., Tang, B., Karrer, B., Sheynin, S., Ross, C., Polyak, A., Howes, R., Sharma, V., Xu, P., Tamoyan, H., Ashual, O., Singer, U., Li, S.W., Zhang, S., James, R., Ghosh, G., Taigman, Y., Fazel-Zarandi, M., Celikyilmaz, A., Zettlemoyer, L., Aghajanyan, A.: Scaling autoregressive multi-modal models: Pretraining and instruction tuning (2023)
32. OpenAI: Gpt-4 technical report (2023)
33. Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., eds.: Advances in Neural Information Processing Systems. Volume 35., Curran Associates, Inc. (2022) 24824–24837
34. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface's transformers: State-of-the-art natural language processing. CoRR **abs/1910.03771** (2019)