Proceedings of MSS 2025 Symposium: Automation and Smart Technologies in Petroleum Engineering
DOI: 10.54254/2755-2721/142/2025.KL21691

A Prompt That Can Stimulate the Self-Correction Ability of the VLM, Improving the Performance on VQA Task

Tiange Lyu[1,a,*]

[1]School of Communication & Information Engineering, Shanghai University, Shanghai, China
a. ltg03@shu.edu.cn
**corresponding author*

Abstract: Visual language Models (VLMS) are revolutionizing multimodal understanding by bridging the gap between vision and language, with great potential in diverse applications. How to improve its performance, so that it can better complete a variety of tasks has become the goal of researchers. To improve the performance of VLM on Visual Question answering (VQA) tasks, this paper proposes an innovative method, Self-correction Prompt, which integrates self-correction into prompt engineering by looping questions to improve accuracy while avoiding additional model training. Experiments on three typical VLMs show that Self-correction Prompt is effective, and the accuracy can be improved by 1.7% at most. It can also stimulate the self-detection ability of the model to find the previously made errors, and its efficiency can reach an average of 72%. The paper also discusses the types of tasks VLMs are not good at, which often prevent models from further improving their accuracy. The proposed method is simple and can be seamlessly integrated into a variety of VLMs, which provides a new idea for the following research on prompt engineering.

Keywords: Prompt Engineering, VLM, VQA, self-correction

1. Introduction

Large language models (LLMs) have shown unparalleled abilities in the domain of natural language processing (NLP) [1,2,3]. They can complete various contextual tasks effectively and accurately. Due to the success of LLMs in NLP, studies have begun to extend it to multimodal tasks, especially in the development of visual language models (VLM), by integrating visual encoders with pre-trained LLMs [4,5,6]. LLM provides strong support for handling tasks that involve the combination of images and text, while visual encoding offers a channel for understanding images. VLMs have demonstrated significant success on a variety of multimodal tasks.

As an important indicator for assessing the comprehension abilities of VLMs, Visual Question Answering (VQA) is a fundamental task within the domain of multimodal learning [7]. VQA evaluates the abilities of models to understand visual content and generate appropriate answers to the questions. To enhance the performance of models, various approaches have been explored, including the utilization of enhanced visual feature extractors [8,9], the design of cross-modal fusion architectures [10], and the integration of knowledge graphs [11]. However, these methods usually require adjustments in the model architecture or require a large amount of labeled data for training, which poses challenges in terms of both computational resources and data acquisition.

In recent years, Prompt Engineering, as a new method, has provided a new way to solve the above problems [12,13]. The core idea of Prompt Engineering is to guide models by carefully designing prompts to make better use of their pre-trained knowledge, thus significantly improving model performance without modifying parameters or requiring additional training. This method not only has the advantage of low computational cost but also has high flexibility.

For instance, studies show that the accuracy of VLM on VQA tasks can be improved by designing an appropriate few-shot prompt [13] or prompting strategies such as Chain-of-Thought (CoT) [14,15]. The design of the prompt directly affects the attention and generation of the model. When more information is added to the prompt, such as extending or filling the prompt, adding relevant context descriptions, etc., it can guide the model to focus on the relevant visual information and generate more accurate answers. Therefore, how to design efficient prompts is an important direction.

In addition, the Self-Correction capability of LLMs has garnered significant interest, enabling models to refine their outputs through iterative reflection [16]. Researchers are now exploring the adaptation of this concept to VLMs. For example, Self-Correction Learning (SCL) aims to enhance VLM performance in visual language reasoning by learning from its own corrections [17]. However, such methods often involve a training phase, leading to increased computational demands.

As a result, integrating Self-Correction with Prompt Engineering may offer a more efficient solution. Prompt Engineering can avoid additional model training and thus save many computational resources and the well-designed prompt can guide the model to self-reflection, which can further improve the performance of VLM.

This paper focuses on the VQA tasks and proposes an approach to incorporate Self-Correction into Prompt Engineering. This method uses the cyclic questioning architecture to add the first generated answer to the subsequent prompt as additional context information, and the VLM will reprocess the image and evaluate the previous answer to determine the correctness of the answer and correct the answer if necessary. This method aims to leverage the efficiency of Prompt Engineering and the accuracy of Self-Correction to improve the performance of VLM in VQA tasks without the need for additional training.

2. Method

2.1. Dataset

This paper employs the TextVQA dataset, a benchmark specifically designed for VQA tasks [18]. This dataset requires VLMs to have the ability of both reading comprehension and reasoning about textual and other visual elements within images.

The dataset contains 28,408 images and 45,336 question-answer pairs. These questions cover many types, for instance, recognizing specific text (e.g., brand names), understanding spatial relationships (e.g., "What does it say on the top right logo?), inferring the answer based on the text and image content, or determine whether the text itself is the answer (e.g., copy-paste a number). Furthermore, each image in the dataset has 10 corresponding standard answers. These multiple answers serve to avoid biases problem introduced by a single ground truth, thereby allowing for a more comprehensive evaluation of the VLM's ability of visual understanding. The diversity of images, questions, and answers, and the requirement for reading and reasoning make TextVQA an appropriate and versatile benchmark for testing VLM abilities. It facilitates a more thorough evaluation of the performance of VLM in scenarios that require a strong fusion of visual and textual understanding.

2.2. Self-correction Prompt

This chapter is divided into three parts, introducing: 1) structure, 2) Self-correction Prompt, and 3) Method Feasibility (Figure 1).

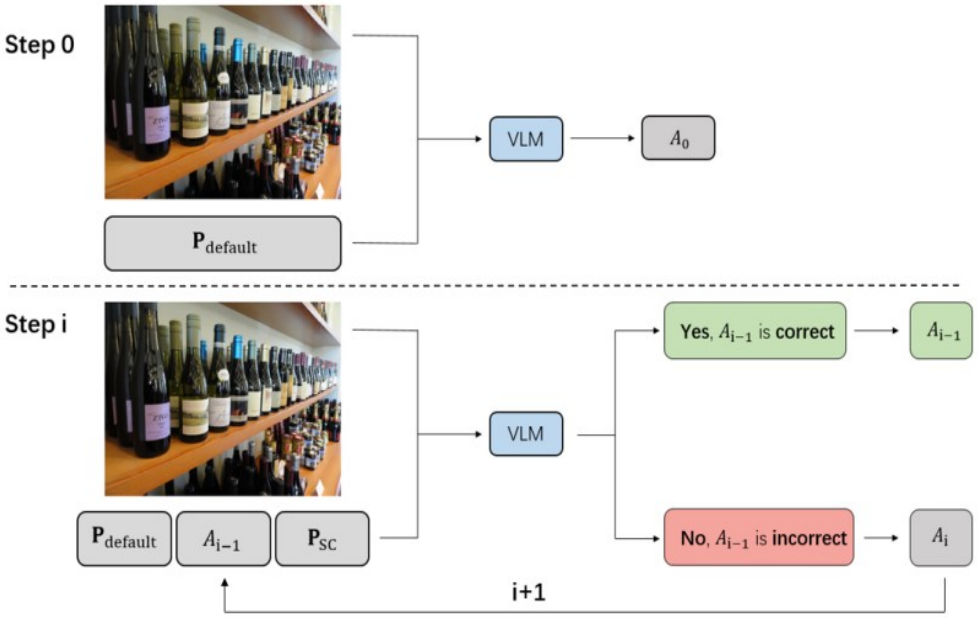


Figure 1: The structure of Self-correction Prompt

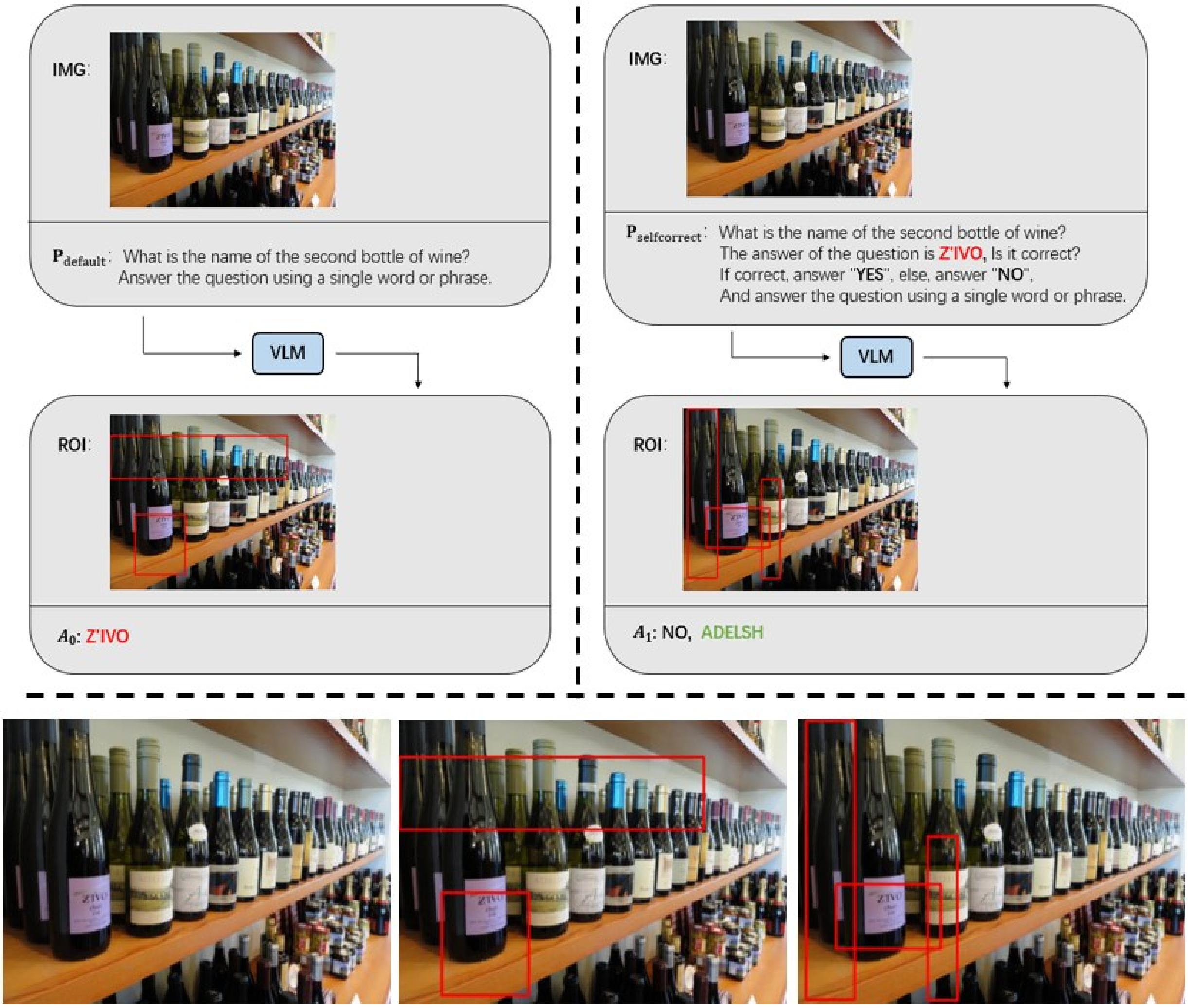
2.2.1. Structure

Initially, the VLM is provided with an image and a corresponding text prompt, and then give an initial answer, denoted as A_0. Subsequently, the same VLM is provided with the same image, A_0, and a Self-correction Prompt. This VLM maintains the same parameters as before but without relying on prior inferences. In this stage, the VLM output is a classification of the answer A_0, whether it is correct or not. If deemed correct, the process terminates and A_0 is outputted as the final answer. If classified as incorrect, a new answer A_1 is generated base on the incorrect answer A_0 as additional context. This cycle is repeated until the VLM classifies the previously generated answer A_{i-1} as correct.

Figure 1 illustrates the reasoning process of the VLM, where the VLM is the same model employed throughout, but with each iteration lacking any mutual memory or awareness of prior steps. $P_{default}$ Is the Prompt given in TextVQA. In Step 0**,** VLM will first generate a baseline answer. A_0. In Step i**,** VLM will judge the correctness of the answer generated last time and update the answer continuously (red part in the figure) until it self-approves and outputs the answer after the cycle (green part in the figure).

2.2.2. Self-correction Prompt

consists of three parts (in gray at the bottom left of Figure 1):$P_{default}$, A_i, and P_{SC}. $P_{default}$ is the original prompt given in TextVQA. A_i is the answer from the last iteration or the baseline answer A_0. P_{SC} requires the model to make a judgment on the correctness of the answer A_i. According to [16], the key to stimulate the self-correction of LLM lies in zero temperature and fair prompts. Therefore, the design of P_{SC} needs to ensure absolute neutrality without any tendency. An example of P_{SC} is shown in Figure 2.



$P_{default}$: What is the name of the second bottle of wine?
Answer the question using a single word or phrase.

$P_{selfcorrect}$: What is the name of the second bottle of wine?
The answer of the question is Z'IVO, Is it correct?
If correct, answer "YES", else, answer "NO",
And answer the question using a single word or phrase.

Figure 2: Example of Self-correction prompt and ROIs in different situations

2.2.3. Method Feasibility

This paper suggests that the Self-correction Prompt enhances VLM accuracy by offering more context compared to the initial prompt. By including A_i in the prompt, the VLM is encouraged to re-evaluate the input based on established information rather than generating answers from scratch, thus increasing the likelihood of error correction and generating a new answer. Figure 2 demonstrates VLM's Regions of Interest (ROIs) for Step 0 and Step 1 using various prompts. The VLM was able to identify and correct an error in Step 0 during Step 1.

In Figure 2, the top-left shows an example from Step 0, the top-right shows an example from Step 1, and the bottom shows a zoomed-in view of the ROIs. In Step 0, VLM gives the wrong answer "Z 'ivo" and incorrect ROIs (bottom middle of Figure 2), but under the guidance of the Self-correction Prompt in Step 1, it pays attention to the correct region (bottom right of Figure 2) and order relation and gives the correct answer "ADELSH".

3. Experiment

This section evaluates the performance of the Self-correction Prompt on TextVQA base on several advanced VLMs [18]. It begins with a brief introduction to the involved VLMs, followed by a presentation and discussion of the experimental results.

3.1. Baseline VLMs

In this experiment, three well-established VLMs are evaluated, including: LLaVA-1.5-7B, Qwen-VL-plus, and Gemini-2.0-flash [5,19]. This paper utilized the official implementations for inference with these VLMs and all experiments are conducted under zero-temperature [16]. Given that the approach methodology centers on Prompt Engineering, there is no necessity to modify on model architectures during inference.

Soft Accuracy is used as the metric, which assesses the model's performance by quantifying the degree of matching between the predicted answer and the reference answer [20]. A brief overview of each VLM is provided below.

(1) LLaVA-1.5-7B

LLaVA is a representative of an earlier VLM, which was primarily concerned with how to integrate visual information into large language models effectively. The core idea of LLaVA is to learn a linear mapping that projects visual features into the embedding space of LLMS. It employs CLIP-ViT-L as a visual encoder, while the LLM Vicuna serves as the language decoder [21,22]. Its structure is relatively simple but validates the effectiveness of aligning vision and language modalities.

(2) Qwen-VL-plus

Qwen-VL is an improvement and extension of LLaVA. It adopts a more complex visual receptor architecture, including a language-aligned visual encoder and a position-aware adapter, which aims to capture visual information and spatial relationships in images. The model indicates that VLMs is gradually shifting from simple feature alignment to more complex modeling of visual information. Qwen-VL also benefits from larger scale training data and more powerful LLM Qwen-7B to achieve better performance [19].

(3) Gemini-2.0-flash

Gemini-2.0 is one of the latest generation of multi-modal models introduced by Google and represents the current state of the art in the field of VLM. Although the specific architecture and training details of Gemini are not yet public, its excellent performance suggests that it may adopt more advanced techniques. In this paper, Gemini-2.0-flash are employed, which is recognized as one of its most powerful and responsive versions.

3.2. Results

In Table 1, this paper evaluates the performance of the three baseline VLMS, as well as the performance after the integration of the Self-correction Prompt. Tested on 5000 question-answers pairs. All models are cycled twice, terminating at Step 2.

Table 1: Comparison of the three VLM and Self-correction Prompt on the TextVQA benchmark

Method	Correct	Incorrect	Accuracy
LLaVA-1.5-7B	3062	1938	61.25
LLaVA-1.5-7B + SC Prompt Step_1	3152	1847	63.06
LLaVA-1.5-7B + SC Prompt Step_2	3148	1852	62.96
Qwen-VL-plus	3444	1556	68.88
Qwen-VL-plus + SC Prompt Step_1	3446	1554	68.92

Table 1: (continued).

Qwen-VL-plus + SC Prompt $Step_2$	3446	1554	68.92
Gemini-2.0-flash	3693	1307	73.86
Gemini-2.0-flash + SC Prompt $Step_1$	3704	1296	74.08
Gemini-2.0-flash + SC Prompt $Step_2$	3701	1299	74.02

Firstly, compared with the baseline accuracy of three models: 61.25%, 68.88%, 73.86%, after integrating Self-correction Prompt, the accuracy of single-step is increased by 1.81%, 0.04%, 0.22% respectively, and the accuracy of the twice-steps is increased by 1.71%, 0.04%, 0.16%. It can be seen that Self-correction Prompt does improve the accuracy of the three VLMS on TextVQA to certain extent. The findings support paper's main idea that the Self-correction Prompt can improve VLMs' performance, by adding more information into the prompt.

Secondly, among the three models, LLaVA-1.5-7B has the lowest baseline accuracy 61.25%, but it also has the highest improvement in single-step Self-correction Prompt, compared with Qwen-VL-plus: 0.04% and Gemini-2.0-flash: 0.16%, LLaVA-1.5-7B has achieved 1.81% accuracy improvement. This result may be due to the additional information in the prompt fully leveraging the inherent understanding capabilities of the original model.

Thirdly, independently comparing the accuracy of each model after single-step and twice-steps Self-correction Prompt, it is found that after one additional loop, the accuracy of LLaVA-1.5-7B drops by 0.10%, Gemini-2.0-flash drops by 0.06%. Only Qwen-VL-plus maintains the same accuracy as single-step, which means that a single Step of Self-correction Prompt can exert the maximum effect. This paper suggests that the observed phenomenon may be attributed to the VLM's inherent limitations in providing accurate answers for certain types of questions. This limitation introduces a degree of uncertainty in the model's responses, potentially leading to considerable variations in the generated outputs even when presented with same input prompts. Consequently, the VLMs may exhibit a lack of confidence in its last answer, leading it to revise the response iteratively, which, in turn, can contribute to a decrease in accuracy.

Table 2: The judgments made by the three VLMS in Step1, which based on the answers they output in Step 0. The accuracy all compare with the standard answer

Method	Model Judgment	Total	Actual Correct	Actual Incorrect	Accuracy
LLaVA-1.5-7B	YES, last answer correct	4254	2862	1392	67.28
	NO, last answer incorrect	746	200	546	26.81
Qwen-VL-plus	YES, last answer correct	4989	3443	1546	69.01
	NO, last answer incorrect	11	1	10	9.1
Gemini-2.0-flash	YES, last answer correct	4515	3590	925	79.51
	NO, last answer incorrect	485	103	382	21.24

In Table 2, the paper evaluates the correct rates of the three VLMS for the judgments given in the Step 0, after a single Self-correction Prompt (Step 1).

Firstly, the high accuracy in the classification of “YES” and low accuracy in the classification of “NO” proves that VLM does have the strong ability to self-detect errors, but lacks the strong ability to self-correct, which leads to the small accuracy improvement shown in Table 1. Taking Gemimi-2.0-flash as an example, through Self-correction Prompt, the model divided the answer in Step 0 into two parts. In the “YES” part, the correct rate of Step 0 was 79.51%. In the “NO” part, the correct rate of Step 0 is as low as 21.24%, the error rate of this part is 78.76%, indicating that the Self-correcting Prompt can stimulate the self-detection ability of VLM to some extent.

Secondly, Qwen-VL-plus shows high level of confidence, with a strong inclination to agrees with its first answer in Step 0. The model only give 11 “NO” answers, but almost all of it is indeed incorrect. Combing with smallest accuracy improvement of Qwen-VL-plus in Table 1, it can be attributed to the fact that Qwen-VL is trained on TextVQA, causing it to be overly familiar with their questioning styles and image features [19].

3.3. Limitations

As mentioned above, Self-correction Prompt may not improve the accuracy of VLM by a large margin. However, this approach enables us to identify specific types of questions that VLMs struggle to address effectively, which often represent significant impediments to achieving higher accuracy. By leveraging this method, we can circumvent the need to train a separate classifier—a common, but resource-intensive, strategy for identifying hard samples—thereby streamlining the process of improving VLM performance.

In Figure 3, a word cloud of the tasks plaguing LLaVA-1.5-7B and Gemini-2.0-flash are provide. For example, in the complex time understanding tasks. For both models, it is observed that the iterative process frequently looped to Step N, yet failed to recognize the answer provided in Step N-1. This phenomenon suggests that future research can deeply explore solving strategies for these tasks and optimize the performance of the model on temporal understanding.

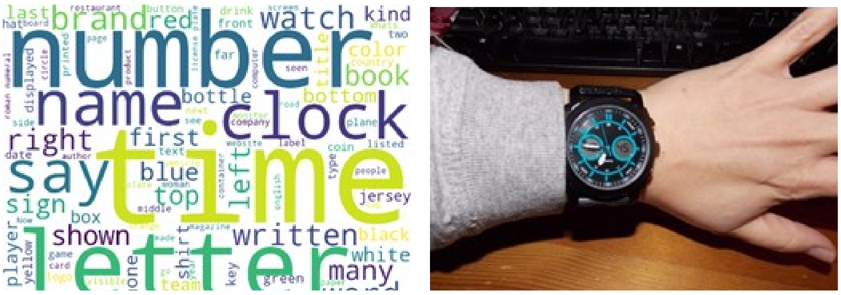


Figure 3: Limitations and example: the left side shows the word cloud of "NO, last answer incorrect", and the right side shows one of the typical question images

4. Conclusion

This paper introduces Self-correction Prompt, a Prompt that combines the advantages of LLMs’ self-correction capability and the efficient of Prompt Engineering, aiming to improve the accuracy of VLMs on VQA tasks. By adding additional information to the prompt, the method in this paper requires the model to make a judgment on the previous answer, which stimulates the self-correction ability of the model for the previous answer, to output a more correct answer. Self-correction Prompt can be seamlessly integrated in any VLMs without any architectural adjustments to the model.

On the TextVQA dataset, experiments on three typical models on the development process of VLMs prove that Self-correction Prompt can improve the accuracy of the model to a certain extent, from a minimum of 0.04% to a maximum of 1.71%. More significantly, Self-correction Prompt can stimulate the self-detection ability of the models, and the two parts classified by Self-correction Prompt achieved 72% accuracy and 81% error, which means that the model is aware of the previously made errors, but cannot modify most of them, and can only correct a small number of errors.

This paper also analyzes and summarizes some typical problems that trouble VLMs. Future research can deeply explore the solution strategies for these tasks to improve the accuracy of VLMs in VQA tasks.

References

[1] *Zhao, W. X., et al. (2023) A survey of large language models. arXiv preprint arXiv:2303.18223.*

[2] *OpenAI. (2024) GPT-4 technical report.*

[3] *Gao, Y., et al. (2023) Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.*

[4] *Chen, Z., et al. (2024) Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[5] *Liu, H., et al. (2024) Visual instruction tuning. Advances in Neural Information Processing Systems 36.*

[6] *Zhang, J., et al. (2024) Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[7] *Huynh, N. D., et al. (2025) Visual question answering: From early developments to recent advances—a survey. arXiv preprint arXiv:2501.03939.*

[8] *Anderson, P., et al. (2018) Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[9] *Zhang, P., et al. (2021) Vinvl: Revisiting visual representations in vision-language models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[10] *Li, Y., Q. Y., et al. (2023) Asymmetric cross-modal attention network with multimodal augmented mixup for medical visual question answering. Artificial Intelligence in Medicine, Volume 144.*

[11] *Li, L., et al. (2019) Relation-aware graph attention network for visual question answering. Proceedings of the IEEE/CVF International Conference on Computer Vision.*

[12] *Chung, H. W., et al. (2024) Scaling instruction-finetuned language models. Journal of Machine Learning Research 25.70: 1-53.*

[13] *Brown, T., et al. (2020) Language models are few-shot learners. Advances in Neural Information Processing Systems 33: 1877-1901.*

[14] *Wei, J., et al. (2022) Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35: 24824-24837.*

[15] *Luan, B., et al. (2024) TextCoT: Zoom in for enhanced multimodal text-rich image understanding. arXiv preprint arXiv:2404.09797.*

[16] *Liu, D., et al. (2024) Large language models have intrinsic self-correction ability. arXiv preprint arXiv:2406.15673.*

[17] *He, J., et al. (2024) Self-correction is more than refinement: A learning framework for visual and language reasoning tasks. arXiv preprint arXiv:2410.04055.*

[18] *Singh, A., et al. (2019) Towards VQA models that can read. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[19] *Bai, J., et al. (2023) Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966.*

[20] *Antol, S., et al. (2015) VQA: Visual question answering. Proceedings of the IEEE International Conference on Computer Vision.*

[21] *Radford, A., et al. (2021) Learning transferable visual models from natural language supervision. International Conference on Machine Learning. PMLR.*

[22] *Chiang, W. L., et al. (2023) Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. See https://vicuna.lmsys.org, 2(3):6*