**SCIENCE CHINA**
Information Sciences

• **LETTER** •

January 2025, Vol. 68, Iss. 1, 119101:1–119101:2
https://doi.org/10.1007/s11432-024-4211-9

# PS-CoT-Adapter: adapting plan-and-solve chain-of-thought for ScienceQA

Qun LI[1], Haixin SUN[1], Fu XIAO[1]*, Yiming WANG[1], Xinping GAO[2] & Bir BHANU[3]

[1]*School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;*
[2]*Purple Mountain Laboratories, Nanjing 211111, China;*
[3]*Department of Electrical and Computer Engineering, University of California at Riverside, Riverside 92521, USA*

Received 20 April 2024/Revised 16 September 2024/Accepted 27 October 2024/Published online 19 November 2024

**Citation** Li Q, Sun H X, Xiao F, et al. PS-CoT-Adapter: adapting plan-and-solve chain-of-thought for ScienceQA. Sci China Inf Sci, 2025, 68(1): 119101, https://doi.org/10.1007/s11432-024-4211-9

Large language models (LLMs) have recently shown remarkable performance in a variety of natural language processing (NLP) tasks. To further explore LLMs' reasoning abilities in solving complex problems, recent research [1–3] has investigated chain-of-thought (CoT) reasoning in complex multimodal scenarios, such as science question answering (ScienceQA) tasks [4], by fine-tuning multimodal models through human-annotated CoT rationales. However, collected CoT rationales often miss the necessary reasoning steps and specific expertise. Our motivation stems from the limitations observed in existing methods such as the Multimodal-CoT reasoning paradigm [2] and LLaMA-Adapter [5]. These approaches, while pioneering, face challenges in maintaining the integrity of multimodal information and providing clear, step-by-step rationales for complex tasks. To address these problems, this paper optimizes the original manually annotated rationales with high-quality plan-and-solve CoT (PS-CoT) that are generated by LLMs. Furthermore, a semantic adapter is introduced, which allows the understanding of high-order image semantics, achieving the fusion and alignment of visual and textual modalities. The combination of these two innovations constitutes the proposed method in this paper called PS-CoT-Adapter. By connecting the semantic adapter with both the visual encoder and the language model, the proposed PS-CoT-Adapter leverages the strengths of each component. The PS-CoT-Adapter is evaluated on the ScienceQA dataset, achieving an accuracy rate of 95.35%, which outperforms the strongest fine-tuned baseline by 2.82%, while its model size is only 5.7% of the fine-tuned baseline. We also present the experimental results on the A-OKVQA dataset to showcase the generalization capability of our model. Our model shows remarkable performance in both direct-answer and multi-choice tasks compared to current advanced models. This significant improvement of performance by PS-CoT adapter demonstrates the effectiveness of the proposed approach. The code and models are available at the website[1)].

The proposed PS-CoT-Adapter contains three stages, including pre-training a semantic adapter, generating PS-CoT rationales, and fine-tuning the model with the integrated planned rationales.

*Pre-training semantic adapter.* The proposed semantic adapter is a trainable module and designed to bridge the gap between the frozen image encoder and LLMs. The semantic adapter consists of two main components, including the context decoder and the semantic aligner. The context decoder is composed of transformer submodules and uses a cross-attention module to guide the model to focus on image content related to the text, while the semantic aligner is a linear layer after the context decoder to refine and align visual and textual modalities.

For pre-training the semantic adapter, we combine this adapter with a frozen image encoder and a frozen language model, and train it with pair-wise image-text data. Given the input image $X_v$, we make use of a pre-trained visual encoder to generate the visual feature $Z_v = g(X_v)$. And for the input text question $X_q$, context $C$, and options $O$, these three are denoted as $X_t = \{X_q, C, O\}$, we adopt the text encoder of a language model to obtain the text feature $Z_t = f(X_t)$. Then we map the image features and text features to the same space for fusion through the context decoder, getting the fused feature $Z_{\mathrm{CD}}$. Specifically, the context decoder consists of transformer submodules, including self-attention layers, cross-attention layers and feed forward layers. The text embeddings from the text encoder interact with each other through self-attention layers, and interact with frozen image features through cross-attention layers. Formally, it can be formulated as follows:

$$Z_{\mathrm{CD}} = \text{Context Decoder}((Z_t, Z_v), \theta), \tag{1}$$

where $\theta$ is the parameter of the context decoder, and it can be learned automatically in the training process.

Then, we apply the semantic aligner to align the text feature $Z_t$ and the visual feature $Z_v$. The refined features can be represented as

$$Z_{\mathrm{SA}} = \text{Semantic Aligner}(Z_t, Z_{\mathrm{CD}}) = Z_t + \lambda Z_{\mathrm{CD}}, \tag{2}$$

* Corresponding author (email: xiaof@njupt.edu.cn)
1) https://github.com/Sunhxxin/PS-CoT-Adapter.

where $\lambda$ is the parameter of the residual link, and it can be learned automatically during the training process.

Finally, we feed the multi-modal features $Z_{\mathrm{SA}}$ into the decoder of the language model to predict the output $Y$. We pre-train the semantic adapter using its original auto-regressive training objective towards prediction tokens, specifically, for generating target text $Y$ of length $L$, we compute the probability as

$$\mathcal{P}(Y|X_v, X_t) = \prod_{i=1}^{L} \mathcal{P}_{\phi}(Y_i|X_v, X_t, Y_{<i}), \tag{3}$$

where $\phi$ is the trainable parameter in the pre-training stage of the semantic adapter, $Y_{<i}$ represents the set of tokens before the current prediction token $Y_i$.

*Generating PS-CoT rationales.* In order to generate high-quality CoT rationales by using LLMs, we introduce the process of zero-shot prompting. This process specifically employs the PS-CoT method, which decomposes complex problems into a series of manageable steps for solution.

Considering that existing annotated data are sufficient for effective training for simpler problems, thus complex planning is not required. Our intervention is focused on those situations where performance is poor due to suboptimal rationales. Based on the previously pre-trained model, we initially assess the performance across various skill categories, identifying specific categories where performance is hindered by suboptimal rationales. Then, for these specific categories, we implement PS-CoT optimization. The generation process of PS-CoT rationales is divided into two primary steps, including plan generation based on skill category and rationale generation based on the specific plan, which are detailed in Appendix C.3.

*Fine-tuning using planned rationales.* The fine-tuning process follows the fine-tuning framework of Multimodal-CoT [2], which includes rationale generation and answer inference. The same model architecture is used in both stages. The image encoder stays frozen to retain pre-learned visual representations, while the language model UnifiedQA is unfrozen. The difference between these two stages lies in their respective inputs and outputs, which are specifically designed to aid in generating an accurate rationale and subsequently deriving an accurate answer based on this rationale.

During the rationale generation stage, we train the model $\mathcal{F}_{\text{rationale}}$ to predict the rationale based on the input, which can be denoted as $R = \mathcal{F}_{\text{rationale}}(X)$, where $R$ is the rationale predicted by $\mathcal{F}_{\text{rationale}}$. As mentioned in the stage of pre-training semantic adapter, the input $X$ is composed of $X_t$ and $X_v$, where $X_t = \{X_q, C, O\}$, which is composed of the question, context, and options. In summary, the input for the rationale generation stage is denoted as $X = \{X_t, X_v\} = \{X_q, C, O, X_v\}$. The probability of generating the rationale $R$ is expressed as follows:

$$\mathcal{P}(R|X_v, X_t) = \prod_{i=1}^{L} \mathcal{P}_{\psi}(R_i|X_v, X_t, R_{<i}), \tag{4}$$

where $\psi$ is the trainable parameter in the rationale generation stage, $L$ is the length of target text, and $R_{<i}$ represents the set of tokens before the current prediction token $R_i$.

In the answer inference stage, we train the model $\mathcal{F}_{\text{answer}}$ to predict the correct answer based on the input, which can be expressed as $A = \mathcal{F}_{\text{answer}}(X')$, where $A$ is the answer that is predicted by $\mathcal{F}_{\text{answer}}$. The input $X'$ is denoted as $X' = \{X'_t, X_v\}$, where $X'_t$ can be represented as $X'_t = X_t \oplus R$. $\oplus$ stands for the concatenation operation. In summary, the input for the answer inference stage is denoted as $X' = \{X'_t, X_v\} = \{X_q, \mathrm{C}, \mathrm{O}, R, X_v\}$. The probability of generating the answer $A$ is expressed as follows:

$$\mathcal{P}(A|X_v, X'_t) = \prod_{i=1}^{L} \mathcal{P}_{\psi^*}(A_i|X_v, X'_t, A_{<i}), \tag{5}$$

where $\psi^*$ is the trainable parameter in the answer inference stage, $L$ is the length of target text, and $A_{<i}$ represents the set of tokens before the current prediction token $A_i$.

In these two fine-tuning stages, we independently fine-tune two models with the same architecture and perform supervised learning on the data integrated with planned rationales. Specifically, the first stage is the rationale generation stage, where we train the model to map the input $X$ to the rationale $R$. The second stage is the answer inference stage, aiming to train the model to map the input $X$ and its corresponding rationale $R$ to the answer $A$. In the inference process of the model, we first use the model that is trained during the rationale generation stage to generate rationales for the test questions. The generated rationale $R$ is then input into the model that is trained during the answer inference stage, which utilizes $X$ and its corresponding rationale $R$ to derive the answer $A$.

*Conclusion.* The novel step-by-step PS-CoT-Adapter demonstrated the vast potential of LLMs in addressing complex reasoning tasks within multimodal scenarios. By utilizing LLMs to generate high-quality PS-CoT rationales for model fine-tuning, we overcame the challenges of missing reasoning steps and specific domain expertise in human manual reasoning. In addition, the integration of the semantic adapter offers an effective mechanism for integrating and aligning visual information with textual information, thereby further enhancing the model's capabilities in semantic reasoning. In the future, we will continue investigating fine-tuning with LLMs and exploring the incorporation of locality-enhanced visual features to enhance language and visual interaction. This approach will enable our model to better tackle complex tasks such as map reading and counting.

**Acknowledgements** This work was supported by National Natural Science Foundation of China (Grant Nos. 62276143, 62302233), Natural Science Foundation of Jiangsu Province (Grant No. BK20231287), and National Science Fund for Distinguished Young Scholars of China (Grant No. 62125203).

**Supporting information** Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

## References

1 Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the Advances in Neural Information Processing Systems, 2022. 824–837

2 Zhang Z, Zhang A, Li M, et al. Multimodal chain-of-thought reasoning in language models. 2023. ArXiv:2302.00923

3 Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. 2023. ArXiv:2303.08774

4 Lu P, Mishra S, Xia T, et al. Learn to explain: multimodal reasoning via thought chains for science question answering. In: Proceedings of the Advances in Neural Information Processing Systems, 2022. 2507–2521

5 Zhang R, Han J, Liu C, et al. LLaMA-Adapter: efficient fine-tuning of large language models with zero-initialized attention. In: Proceedings of the International Conference on Learning Representations, 2024. 1–22