Distilling Small Vision Language Models with Structured Reasoning

Xiwen Chen[1], Hongbin Dong[1], Jinbo Liu[1], Yanqing Lu[1], and Ruitao Mao[1]

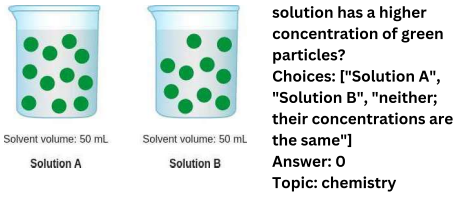
[1]Department of Computer Science, University of Southern California
{xiwenc,hdong372,jinboliu,ylu62702,ruitaoma}@usc.edu

Abstract

Visual Question Answering (VQA) is a challenging task that requires models to integrate visual understanding with reasoning capabilities. However, deploying such models in real-world scenarios is often hindered by the computational demands of large-scale models. To address this, we propose a Structured Visual Reasoning Distillation (SVRD) framework, leveraging chain-of-thought prompting and structured logical reasoning. The framework consists of three main stages. First, the teacher model generates structured rationales. Second, the structured rationales, combined with the original ground-truth labels, serve as the training targets for the student model to help the student model learn to capture the distilled knowledge from the teacher. Finally, the student model generates structured rationales using a greedy decoding strategy. We evaluated our approach with six baseline models, where LLaVA-CoT (11B) serves as the teacher and Moondream (2B) as the student, while the others act as references. The results, measured by accuracy on the ScienceQA benchmark dataset, demonstrate that our method achieves a performance of 53.54% accuracy after knowledge distillation, with a significantly reduced inference time of 279 minutes, compared to the teacher model's 371 minutes. These results highlight that the proposed method achieves acceptable performance while significantly reducing model size and computational overhead. The code is publicly available at: https://github.com/yqlu1015/vllm_distillation.

1 Introduction

In recent years, large language models(LLMs) have demonstrated remarkable performance across a wide range of natural language processing(NLP) tasks. Building upon the advancements of LLMs in understanding and generating text, vision-language models (VLMs) extend this capability by integrating visual and textual modalities to enable more comprehensive multimodal reasoning. However VLMs deployment faces significant challenges, including resource limitations, data security concerns, etc. For instance, the computational and memory requirements of VLMs are often prohibitively high, making them impractical for real-world applications, especially in resource-constrained environments. Additionally, sharing sensitive data with external APIs or cloud-based VLMs raises privacy and security issues.



(a) Model Input

Model Response:
<SUMMARY>
The problem is to determine which solution has a higher concentration of green particles.
</SUMMARY>
<CAPTION>
The image shows two flasks with green particles, each labeled as 'Solution A' and 'Solution B' with a 'Solvent volume: 40 mL' indicated below them.
</CAPTION>
<REASONING>
Since both solutions have the same volume of solvent and the same number of green particles, their concentrations are the same.
</REASONING>
<CONCLUSION>
1
</CONCLUSION>

(b) Model Output

Figure 1: Model demo: Input and Output

To mitigate these problem, knowledge distillation has emerged as an effective method for model

compression. The pioneering work by Hinton et al. (2015) on knowledge distillation introduced the idea of transferring knowledge from a large teacher model to a smaller, more efficient student model. The central idea is for the student to learn not only from the ground truth (hard labels) but also from the predictions (soft labels) of the teacher model.

When employing small models for visual reasoning tasks, we observed that these models often have poor reasoning capabilities. Several key challenges hinder their performance: 1) Small models have fewer parameters, making it difficult to store extensive features and patterns. 2) Small models are prone to overfitting superficial patterns in the training data and lack true reasoning capabilities. 3) Important contextual information might be lost during the fusion of visual and language features due to computational constraints.

To address these challenges, we propose a novel distillation method leveraging chain-of-thought(CoT) prompting and structured logical reasoning, called **Structured Visual Reasoning Distillation (SVRD)**, which enhances the reasoning capabilities of the student model while minimizing resource consumption and maintaining data security. Key steps in SVRD include: 1) Employ structured CoT reasoning to understand relationships and derive meaningful insights. Teacher model process the input data and generate question summary, image caption and rationale for later knowledge distillation. 2) Transfer reasoning capabilities from the large teacher model to smaller task-specific models for improved efficiency and scalability. An example is shown in Figure 1, which consists of our distilled model's input and output.

Our contributions are summarized as follows:

1. We propose SVRD framework: A three-stage distillation process where the teacher model generates structured rationales, and the student model learns step-by-step reasoning.

2. We apply CoT prompting: The student model replicates the teacher's reasoning process for improved interpretability.

3. Improved performance and efficiency: Our method outperforms baseline models on the ScienceQA benchmark and significantly reduces inference time.

4. Domain-specific insights: Performance analysis across domains highlights the impact of training data distribution.

2 Related Work

2.1 Visual reasoning with VLM

Visual reasoning requires models to combine visual perception with high-level cognitive abilities to interpret and analyze complex visual inputs. Tasks such as Visual Question Answering (VQA) (Ishmam et al., 2024; Li et al., 2022a) and Visual Entailment (Song et al., 2022) are widely used benchmarks to evaluate a model's ability to reason about visual content and textual inputs. Earlier methods often relied on neural-symbolic approaches (Amizadeh et al., 2020), explicitly modeling the reasoning process through symbolic representations and simulation. However, these methods are typically resource-intensive and limited in scalability.

With the rapid development of LLMs, VLMs now integrate advanced reasoning abilities to address visual tasks. Recent methods such as LLaVA-CoT (Let Vision-Language Models Reason Step-by-Step) (Xu et al., 2024), inspired by the CoT paradigm, enhance visual reasoning by employing a step-by-step reasoning strategy. This method focused on enhancing visual reasoning by refining the visual encoding strategies (Jin et al., 2024; Li et al., 2024; Liu et al., 2023), producing high-level cognition-focused visual tokens. Similarly, it optimizes reasoning performance through sequential instruction tuning steps, such as prompt tuning (Zamfirescu-Pereira et al., 2023), in-context learning, and supervised fine-tuning (Shen, 2024). Each step builds upon the previous one, enabling more interpretable and effective reasoning.

2.2 Chain-of-Thought in LLMs

Chain-of-Thought reasoning has emerged as an effective approach to improve the reasoning capabilities of LLMs by breaking complex problems into step-by-step reasoning chains. CoT prompting explicitly decomposes the reasoning process into intermediate steps, guiding the model to tackle challenging tasks such as commonsense reasoning, logical reasoning, and multimodal reasoning (Wei et al., 2022b; Sap et al., 2020; Geva et al., 2021).

Recent studies have explored two major paradigms of CoT reasoning: Zero-Shot-CoT and Few-Shot-CoT. In Zero-Shot-CoT, Kojima et al. (2022) demonstrated that LLMs could achieve reasonable performance by simply adding intuitive prompts like "Let's think step by step" to invoke reasoning. On the other hand, Few-Shot-CoT

Wei et al. (2022a); Zhang et al. (2023) uses carefully crafted or automatically generated reasoning demonstrations to condition the model during inference. Futhermore, works such as Multimodal Chain-of-Thought Reasoning (Zhang et al., 2024) extend the CoT paradigm to vision-language tasks, enabling LLMs to reason over multimodal inputs by incorporating intermediate steps that process both textual and visual information.

2.3 Reasoning Distillation

Reasoning distillation aims to transfer the reasoning capabilities of Large Language Models to smaller models, enabling them to achieve comparable performance on reasoning tasks while maintaining computational efficiency. Recent works (Li et al., 2022b; Ho et al., 2023; Magister et al., 2023) have explored distilling reasoning abilities, often leveraging CoT prompting (Kojima et al., 2022; Wei et al., 2022a) to guide smaller models in producing intermediate reasoning steps.

However, prior studies have shown that while reasoning distillation is effective for tasks such as arithmetic or symbolic reasoning, it performs less effectively for knowledge-intensive reasoning tasks, where external factual knowledge plays a critical role (Ho et al., 2023; Li et al., 2022b). To address this challenge, techniques like Knowledge-Augmented Reasoning Distillation (Kang et al., 2023) augment small models with external knowledge sources, such as retrieved documents, to generate more accurate rationales and improve reasoning outcomes. By incorporating external knowledge, smaller models are able to compensate for their limited capacity compared to larger counterparts.

3 Structured Reasoning Distillation for Vision Language Models

Our goal is to 1) leverage the advanced zero-shot reasoning ability of large VLM to elicit structured rationales from a teacher model, and 2) train a student VLM to reason systematically. We consider the task of VQA which requires not only image recognition but also text comprehension capabilities of a model (Małkiński and Mańdziuk, 2023). Specifically, given an image and a question related to it, a model is asked to predict the gold answer. In this section, we mainly discuss our SVRD framework for VQA.

3.1 Generating Structured Rationales

Wei et al. (2022b) demonstrate that CoT prompting significantly improves the reasoning ability of LLMs. Our teacher model, based on LLaVA-CoT (Xu et al., 2024), generates systematic and structured reasoning chains exclusively for visual reasoninf tasks. The teacher model takes the question text, its corresponding image and answer from the training data as input. It is pretrained to be able to generate high quality rationales in zero-shot setting. Specifically, the structured rationale decomposes the process of knowledge generation into three structured reasoning stages to ensure an explicit breakdown of the task and a clear recognition of each thinking step:

- **Question Summary Stage**: Teacher model generates a concise summary of the question, which contains the key points and main issues of the question.
- **Image Caption Stage**: Teacher model extracts visual elements related to the question and provide a description of the image which will improve contextual understanding.
- **Rationale Generation Stage**: Teacher model conducts structured logical rationale to show the relationships between various aspects of the question and the answer.

3.2 Knowledge Transfer From Teacher to Student

Knowledge transfer is the primary part to distill the knowledge from a larger, more powerful teacher model into a smaller, more efficient student model. The upper part of Figure 2 illustrates the knowledge transfer process. First, use a frozen teacher model to process the question text, its corresponding image and answer from the training data. Then use this teacher model to generate structured rationales, as illustrated in Section 3.1. The three-stage knowledge from the teacher model along with the answer to the question form the label of student model, while the input of student consists of the question text and its corresponding image. This training strategy enables the student model to use annotated examples during training, and automatically generate similar structured rationales free from the teacher model for new input questions.

The student model is trained to minimize the discrepancy between its outputs and the ground-truth references. We denote the parameters of the

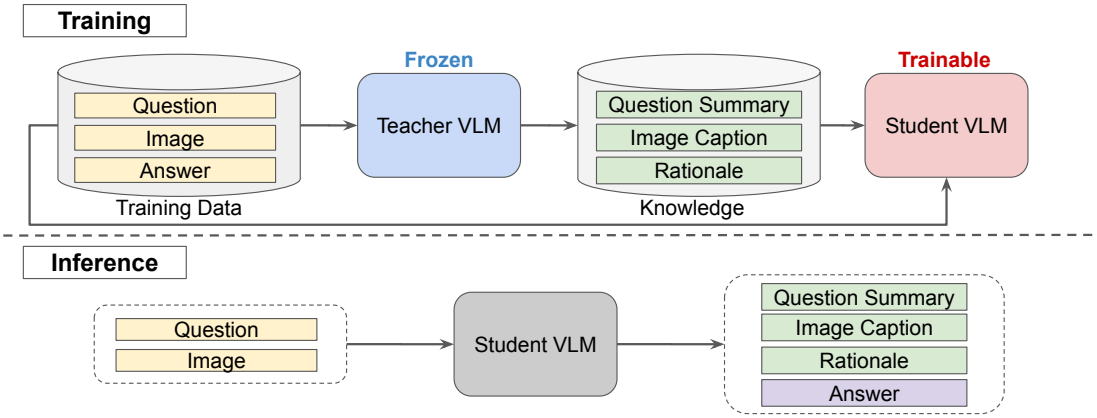


Figure 2: **Overview of our structured reasoning distillation framework.** The upper part shows that during training, a student VLM learns to generate multi-stage rationales obtained from a teacher. The lower part illustrates that during inference, the student VLM reasons step-by-step before giving its answer.

student model as θ, and formulate its generation as a probability distribution P_θ. Let t be the target token sequence comprising rationale and answer, t_i represent the target token at time step i, $t_{<i}$ denote the sequence of tokens generated prior to step i, and q be the token sequence of question and image. The loss function is defined using a standard language modeling loss:

$$L(\theta) = -\sum_i \log P_\theta(t_i \mid q, t_{<i}) \tag{1}$$

3.3 Inference

After training, we obtain a well-trained student model with parameter $\theta^* \in \arg\min_\theta L(\theta)$, which is able to mimic the teacher model to generate structured rationales. As depicted in the lower part of Figure 2, the input to the student model during inference only consists of the text question and its corresponding image. Then it generates three-stage rationales before giving its final answer.

For a new visual question q, the student model employs a greedy decoding strategy to generate a structured rationale r^*, followed by an answer a^*. Let r_i^*, $r_{<i}^*$ denote the generated rationale token at time step i and prior to step i, a_i^*, $a_{<i}^*$ denote the generated answer token at step i and prior to step i. The greedy decoding is formulated as

$$r_i^* = \arg\max_{r_i} P_{\theta^*}(r_i \mid q, r_{<i}^*) \tag{2}$$

$$a_i^* = \arg\max_{a_i} P_{\theta^*}(a_i \mid q, r^*, a_{<i}^*) \tag{3}$$

4 Experiments

4.1 Dataset

ScienceQA (Lu et al., 2022) is used as our VQA dataset. It consists of a comprehensive collection of multimodal science-related questions, featuring text, images, and detailed explanations, designed to test a model's reasoning capabilities. The ScienceQA dataset is structured as follows: each question is paired with accompanying images, text-based questions, multiple-choice answers, and optional lecture notes or solution explanations for detailed knowledge generation. The data is stored in a JSON format where each entry contains the question ID, question text, answer choices, correct answer, optional image paths, and explanations. Our preprocessing step is as follow:

1. Filter out entries that do not include images.
2. Utilize both the "lectures" and "solution" fields to construct additional knowledge context from teacher model for reasoning generation.
3. Normalize and tokenize the text components for consistency, ensuring compatibility with the model's input format.
4. Split the dataset into training, validation, and testing subsets, maintaining a balanced distribution of questions across science topics and modalities.

4.2 Evaluation

The evaluation of this project leverages the test set from the ScienceQA dataset to measure model

performance. The primary evaluation metric is accuracy, which quantifies the proportion of correctly answered questions out of the total number of questions in the test set. The evaluation process is as follows:

- **Model Predictions**: Feed the test set questions, along with their associated images, into the student model. For each question, the model outputs a predicted answer.
- **Ground Truth Comparison**: Extract the ground truth answers from the test set annotations.Compare the model's predictions to the ground truth answers on a question-by-question basis.
- **Accuracy Calculation**: We calculate the overall accuracy across all samples in the test set and also compute the accuracy separately for each topic.

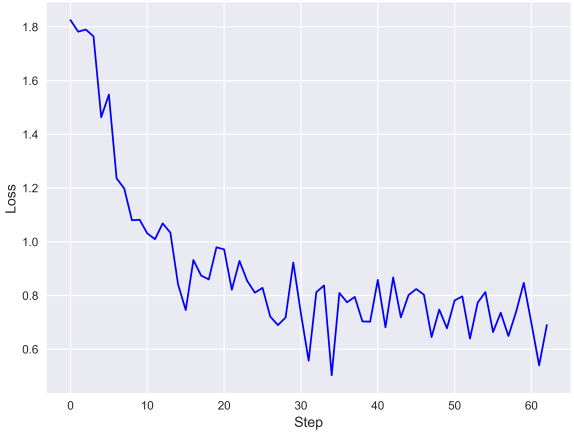


Figure 3: SVRD training loss

4.3 Implementation Details

The project involves testing various VLMs, as detailed in Section 4.4, for image-to-text question answering tasks with reasoning processes, integrating both inference and fine-tuning capabilities.

For inference, we use PyTorch (Paszke et al., 2019) and Hugging Face's Transformers (Wolf et al., 2019) to evaluate Kosmos-2(Microsoft, 2023), Phi-3.5(Microsoft, 2024), and ViLT(Kim et al., 2021) on image-to-text question-answering tasks. The inference hyperparameters include a batch size of 4, a maximum of 500 new tokens, a temperature of 0.0, and greedy decoding. Kosmos-2, a 2-billion parameter model, utilizes a vision

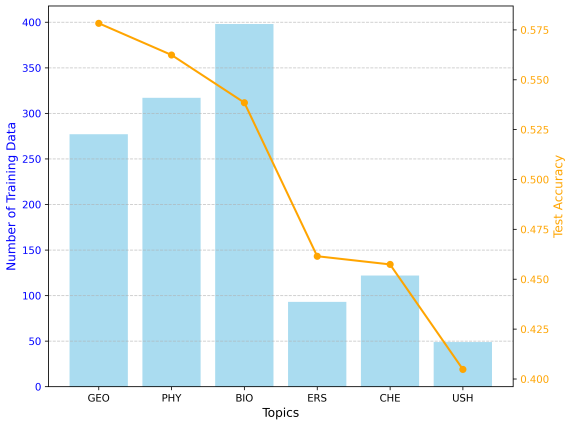


Figure 4: Number of SVRD training data vs accuracy

encoder and text decoder to combine image features with chain-of-thought textual queries. Phi-3.5, a 4-billion parameter model, employs a similar encoder-decoder structure but focuses on structured reasoning stages, achieving a higher accuracy but with a longer inference time. ViLT adopts a encoder only transformer to process images and text in a unified representation space, scoring multiple-choice answers using masked language modeling techniques. All models are evaluated on a filtered version of the ScienceQA dataset containing image-based questions, and predictions are assessed for accuracy with results logged and analyzed for topic-wise performance.

For distillation, we also leverage PyTorch and Hugging Face's Transformers library to fine-tune our Moondream-2B student model. The vision encoder of Moondream is kept frozen to preserve its visual perception capabilities, while fine-tuning is performed solely on the text model. Using a single A100-80GB GPU, we conduct full-parameter fine-tuning of the text model on the first 2000 samples from the ScienceQA training set, supplemented with the structured rationales generated by our teacher model. The detailed experimental configurations are provided in Table 3.

4.4 Results and Analysis

Table 1 shows the accuracies of six baseline models and distilled moondream on ScienceQA dataset. Among the results of the six baseline models, the teacher model LLaVA-CoT achieved the highest scores in five of the seven topics (PHY, BIO, ERS, SAEP, CHE, and USH), while the accuracies of the student model Moondream are very low in

Table 1: Baseline and knowledge distillation models accuracies (%) on ScienceQA datset. LLaVA-CoT to ViLT are six baseline models and Moondream-kd represents knowlegde distilled Moondream. Question topics: GEO = geography, PHY = physics, BIO = biology, ERS = earth science, SAEP = science and engineering practices, CHE = chemistry, USH = US history. Overall accuracy is based on 15 topics and we only list the 7 most sample numbers topics. "min" means minutes.

Model	Size	GEO	PHY	BIO	ERS	SAEP	CHE	USH	Overall	Inference time
LLaVA-CoT	11B	62.50	**84.23**	**85.60**	**84.61**	89.84	**74.46**	**95.23**	78.82	371 min
LLaVA-NeXT	7B	72.33	61.17	74.44	63.46	60.93	50.00	42.85	64.94	70 min
Phi-3.5-vision-instruct	4B	**88.17**	75.06	78.41	82.05	**93.75**	65.96	86.90	**81.71**	156 min
Moondream	2B	45.17	48.47	50.87	44.87	44.53	43.62	41.67	46.20	158 min
KOSMOS-2	2B	9.67	20.47	13.90	22.44	13.28	20.21	11.90	14.97	17 min
ViLT	113M	19.83	44.94	41.44	50.64	46.88	42.35	23.81	35.91	3 min
Moondream-kd	2B	57.83	56.24	53.85	46.15	53.91	45.74	40.48	53.54	279min

Table 2: Baseline BLEU, ROUGE, and BERT-Score on VQAonline dataset. "Avg" means average and "min" means minutes.

Model	Size	BLEU	ROUGE1	ROUGE2	Avg F1	Avg precision	Avg recall	Inference time
LLaVA-CoT	11B	**0.0125**	0.2163	0.0314	**0.8840**	0.8446	0.8242	276min
LLaVA-NeXT	7B	0.0125	**0.2439**	**0.0394**	0.8263	0.8233	**0.8297**	171min
Phi-3.5-vision-instruct	4B	0.0028	0.1812	0.0235	0.8366	**0.8538**	0.8206	49min
Moondream	2B	0.0097	0.2113	0.0295	0.8306	0.8410	0.8209	74min
KOSMOS-2	2B	0.0013	0.1144	0.0160	0.5833	0.5963	0.5711	33min
ViLT	113M	0.0000	0.0020	0.0020	0.7826	0.8077	0.7594	2min

Table 3: Experimental configurations. OP = optimizer, LR = learning rate, BS = batch size, GA = gradient accumulation steps, DS = decoding strategy, #IT = number of image tokens, #MNT = number of max new tokens.

Param.	Value	Param.	Value
OP	Adam	DS	Greedy
Epoch	1	GA	1
LR	1.4×10^{-5}	#IT	729
BS	16	#MNT	512

all topics. Table 2 shows the results of BLEU, ROUGE, and BERT-Score of six baseline models on VQAonline dataset. After we completed baseline testing of six models on two data sets, we found that compared with the results of ScienceQA, the VQAonline data set could not well reflect the performance difference between the student model and the teacher model. In the results of VQAonline, the performance of the Moondream model are almost the same as those of the LLaVA-CoT in terms of the ROUGE and BERT-Score evaluation metrics, except for a slight difference in BLEU. Therefore, we chose to train a distilled student model on the ScienceQA dataset.

Figure 3 shows the decrease in training loss over 60 steps in the distillation process. The loss initially drops sharply from about 1.8 to just over 1.0 within the first 10 steps, then gradually decreases and stabilizes between 0.6 and 0.8 after step 30. The fluctuations toward the end indicate minor adjustments in the model's learning process, but the overall trend suggests effective learning and performance improvement. It can be seen from the results in Table 1 that after SVRD, Moondream-kd has made great progress in six of the seven topics. The biggest improvement was in topic geography, with accuracy increased by 12.66%. The overall accuracy increased from 46.20% to 53.54%. Figure 4 shows the relationship between the number of rationales of training data for different topics and the final accuracy during the distillation process. From the figure, it can be noticed that the accuracy is generally proportional to the number of training rationales, because the number of rationales and their accuracies of GEO, PHY, and BIO are much higher than those of ERS, CHE, and USH.

Compared to the teacher model LLaVA-CoT, although the distilled student model can not perform as well as it's teacher model, it has relatively much lower inference time. Compared to the three baseline models LLaVA-CoT, LLaVA-Next, and Phi-3.5-vision-instruct, the Moondream-kd has a significant advantage that the size of the student model is much smaller, which means it can be deployed in many devices or software that have lim-

ited storage space, such as embedded systems in IoT devices and lightweight applications running on edge computing platforms.

5 Conclusion

In this work, we explore knowledge distillation for VLMs, with a focus on the reasoning-intensive VQA task. We propose Structured Visual Reasoning Distillation (SVRD), a method that efficiently transfers the systematic reasoning capabilities of large VLMs to smaller VLMs. Experimental results on standard VQA benchmark datasets demonstrate that SVRD significantly enhances the structured reasoning ability of small VLMs while maintaining inference efficiency. Additionally, performance variations across questions from different domains highlight the critical role of training data in the effectiveness of SVRD.

6 Limitations and Future Works

Despite the promising results, this work has certain limitations that require further exploration. First, while our distillation method improves the structured reasoning capabilities of smaller VLMs, the distilled model sometimes performs reasoning-to-error. i.e., it gives incorrect answers while the rationales are correct. Figure 5 shows such an example. Addressing this issue remains an important direction for improvement. Second, evaluating whether the rationales produced by the large VLM are truly effective for knowledge distillation remains challenging. Developing robust methods to assess the quality and relevance of these rationales would further strengthen the distillation process.

For future work, we plan to conduct experiments on a broader range of datasets to validate the generalizability of our SVRD approach. Additionally, we aim to incorporate more comprehensive evaluation metrics beyond accuracy, such as measuring the quality and consistency of generated rationales, to better capture the reasoning capabilities of distilled models.

References

Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. 2020. Neuro-symbolic visual reasoning: Disentangling. In *International Conference on Machine Learning*, pages 279–290. PMLR.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Md Farhan Ishmam, Md Sakib Hossain Shovon, Muhammad Firoz Mridha, and Nilanjan Dey. 2024. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, page 102270.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2023. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *arXiv preprint arXiv:2305.18395v2*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Hao Li, Xu Li, Belhal Karimi, Jie Chen, and Mingming Sun. 2022a. Joint learning of object graph and relation graph for visual question answering. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 01–06. IEEE.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jingu Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022b. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jianke Zhu, and Lei Zhang. 2024. Token-packer: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Mikołaj Małkiński and Jacek Mańdziuk. 2023. A review of emerging research directions in abstract visual reasoning. *Information Fusion*, 91:713–736.

Microsoft. 2023. Kosmos-2: Grounding multimodal large language models to the world.

Microsoft. 2024. Phi-3.5: Vision-language instruction tuning.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8024–8035.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33.

Ming Shen. 2024. Rethinking data selection for supervised fine-tuning. *arXiv preprint arXiv:2402.06094*.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. 2022. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022a. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

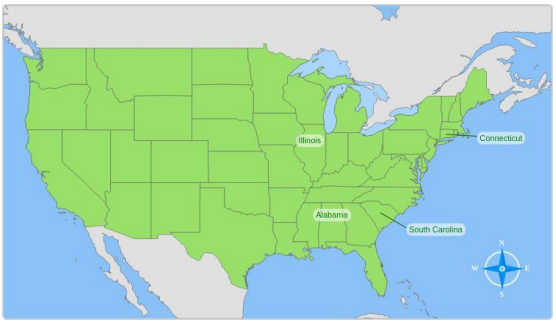
Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440v2*.

JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923v5*.

A Generation Examples



Question: Which of these states is farthest west?
Options: A. "Alabama", B. "Illinois", C. "South Carolina", D. "Connecticut"

Answer: B
Solution: To find the answer, look at the compass rose. Look at which way the west arrow is pointing. Illinois is farthest west.

Generation of distilled student model:

<SUMMARY> To determine which state is farthest west, I will analyze the map's layout and the relative positions of the labeled states. I will identify the state that is located farthest to the west among the options provided. </SUMMARY>

<CAPTION> The image is a map of the United States with four states labeled: Illinois, Alabama, South Carolina, and Connecticut. The map includes a compass rose indicating the cardinal directions. </CAPTION>

<REASONING> To find the state that is farthest west, I will examine the map's layout and the relative positions of the labeled states. Illinois is located in the central part of the map, Alabama is to the south of Illinois, South Carolina is to the south of Alabama, and Connecticut is to the north of Illinois. Therefore, the state that is farthest west is Illinois. </REASONING>

<CONCLUSION> A </CONCLUSION>

Figure 5: An example of reasoning-to-error for our distilled model. Even though the model correctly identifies "Illinois" as the farthest west state in the reasoning stage, it fails to choose the correct option in the conclusion stage.