Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA37294

# *SRFE: Measuring Chain-of-Thought Faithfulness with Independent Stepwise Visual Probes*

**Xinyi Wei**

*School of Computer Science, The University of Sydney, Sydney, Australia*
*xwei0594@uni.sydney.edu.au*

***Abstract.*** Using stepwise probe as an external reference or Chain-of-Thought (CoT) to evaluate visual-language model (VLM) analysis ability is common, but the faithfulness of PASS/FAIL verdicts in VLM produced by explicit CoT and independent SRFE probe needs further work to figure out. Prior works on CoT faithfulness are mainly about large language model (LLM) and internal interpretability signals, with limited VLM, Vision Prompt Distillation (VPD), structured verdict settings and behavioral probes. This paper focuses on evaluating Stepwise Reasoning Faithfulness Evaluation (SRFE), a five-step protocol (object->color->size->relation->composition) with prompt-defined order, and CoT faithfulness after distillation, trying to align them using probe Supervised Fine-Tuning (SFT) to reduce two different methods to a single faithfulness evaluating method. To get features of SRFE, the paper compares the independent and conditional probe version and find weak chain effect in model then order ablation experiment find SRFE is order invariance. After VPD distillation on Qwen2-VL, teacher and student have a high agreement (97.6% stepwise, 88% exact bits), supporting SRFE probe is procedure faithfulness but not internal reasoning faithfulness. To find Chain-of-thought procedure faithfulness, the paper uses VLM to generate explicit CoT then separate the answer and compare each one under SRFE to get the similarity, finds that baseline CoT-SRFE probe partial misalignment. To improve alignment, SRFE probe supervised SFT on CoT is implemented, but finds a trade-off between independent probe agreement, chain consistency, and say-do &do-say.

***Keywords:*** Chain-of-Thought Faithfulness, Visual Language Model, Visual Prompt Distillation, Probes, Large Language Model

## 1. Introduction

Recent work shows that Chain-of-Thought (CoT) is often not a faithful representation of a model's reasoning process and CoT fails to align with the model's internal computation path [1, 2]. Metrics such as CoT-Interpretability Alignment (CIA) quantify parametric faithfulness using mechanistic interpretability tools to detect similarity between CoT text and internal reasoning but mainly focuses on text large language models (LLMs) and white-box settings [2]. Current vision-language models (VLMs) often struggle to perform systematic and structured reasoning, especially when handling complex visual question-answering tasks [3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA37294

CoT is difficult to align with internal reasonings, but its parsed PASS/FAIL verdicts can be checked against Stepwise Reasoning Faithfulness Evaluation (SRFE), an external five-step protocol (object-> color-> size-> relation->composition) with prompt-defined order. Whether CoT verdict result can align with SRFE probes to get same faithfulness remains an open question to further explore. SRFE should be procedure faithfulness which means after Visual Prompt distillation (VPD), SRFE procedure remains stable and serve as an external reference. If baseline CoT verdicts are misaligned with probes, this paper applies probe Supervised Fine-tuning (SFT) to improve alignment.

This paper studies alignment between CoT-probe in multimodal VLM(Qwen-2VL) without defining either side with "true internal reasoning". SRFE evaluates model reasoning stepwise using probes on CLIP model, and the outputs are PASS/FAIL judgments and margins on controlled-probe geometric artificial synthesized images with annotation. The canonical step order is fixed by the protocol and to further discover chain effect in SRFE, Exp-1 compares Independent vs. Conditional protocol. To further find whether order variance or not, Exp-2 is order ablation. Then use VPD distillation as a compression test: if SRFE probe is stable and reliable external evaluation procedure, it should remain consistent after transferring from a frozen teacher VLM to a lightweight student. (On Qwen2-VL, external SRFE is stable) Exp-3 designed to find whether CoT-probe alignment or not via parsed step-wise verdicts, evaluating from CoT_bits, probe_bits, order_violation and say-do/do-say ratio. Exp-4 is probe-supervised SFT to improve Qwen-2VL CoT-probe alignment, evaluating from delta step_agree/ CoT_agree/ order_violation/ say-do. Contributions in this paper is SRFE probe is stable under distillation on Qwen2-VL means it can be used as external reference of procedure faithfulness. In addition, SRFE protocol probe run approximately parallel, conditional add weak chain effect. To get CoT faithfulness, the paper finds CoT-probe misalignment, verdicts phrased from CoT are partially misaligned with PASS/FAIL answered by independent SRFE probe. Then, probe-supervised SFT improves CoT-probe alignment, but will reach trade-offs in specific criteria.

## 2. Literature review

Generating a chain of thought—a series of intermediate reasoning steps—significantly improves the ability of large language models to perform complex reasoning [4]. Besides accuracy improvement, CoT is also claimed to "provide an interpretable window into the behavior of the model" [4]. However, faithfulness is not guaranteed: Turpin et al. claimed that CoT can systematically misrepresent the true reason for a model's prediction, as the final answer does not necessarily follow from the generated reasoning chain. Because these explanations remain fluent and coherent, they can raise confidence in the model's reasoning even when the underlying decision differs [5]. Lanham et al. intervene on the CoT and then examine the model's predictions when adding mistakes or paraphrasing [6]. They find that CoT faithfulness decreases as models become larger, on most tasks they study. Chen et al. extend these findings to reasoning models, showing that they frequently fail to verbalise the cues that determine their answers [1]. Hong et al. instead adopt a white-box perspective, proposing CoT-Interpretability Alignment (CIA) to compare CoT text against internal reasoning strategies recovered with mechanistic interpretability tools [2].

Two limitations are common to the prior work above. First, evaluation is text-only: faithfulness is assessed on language tasks rather than multi-modal tasks. Second, references are mainly from final result or the model's internal state, both of which are difficult to obtain at the level of individual steps.

DOI: 10.54254/2755-2721/2026.BA37294

To obtain a step-level reference, process reward models (PRMs) score individual reasoning steps and have become a test-time scaling, but step supervision is expensive to collect [7]. ThinkPRM trains a verifier that produces a verification chain-of-thought every step and performs well while using only 1% of the process labels in PRM800K [7]. The verifier selects better solutions under best-of-N selection, however, in this paper, stepwise probe is used for auditing the agreement between the verdicts parsed from the CoT and the probe bits.

General visual reasoning requires a model not only to understand visual content but also to reason about it [8]. In this paper, the object, color, size, relation and composition are the reasoning intermediates.

Many benchmarks evaluate this ability, from commonsense to multi-discipline indicators [8]. However, most of them score only the outcome and ignore the importance of detecting steps along the way.

Previous works make multimodal reasoning more explicit. Visual CoT provides 438k question-answer pairs annotated with and intermediate bounding boxes, together with a pipeline that produces interpretable intermediate thoughts [9]. LLaVA-CoT instead decomposes reasoning into explicit stages and searches over them at inference time [3].

These methods make multimodal reasoning more interpretable but do not test whether a stated step agrees with an independent measurement of the same visual property. Interpretability is not equal to faithfulness. In this paper, the stepwise probe gives the model's own non-generative judgement, which is then compared with verdicts parsed from chain-of-thought.

Synthetic images with annotation are widely used in compositional reasoning. CLEVR renders images on daily objects with controlled scenes, so the correct answer is known in advance [10]. This paper follows this setting and introduces a benchmark of 500 two-object images, in which each image violates only one attribute among object, colour, size, relation and composition. In addition, every image has a known expected failure step, which general VQA benchmarks cannot provide.

Knowledge distillation is used to transfer model reasoning from large models to small ones. Socratic CoT decomposes complex problem to small problems, and the distillation target is the reasoning ability, and the aim is better accuracy. In this paper, the distillation target is the fixed stepwise probe: the student is trained to reach same step verdicts as the teacher [11].

## 3. Method

### 3.1. Methodology

#### 3.1.1. Dataset

The paper evaluates on a controlled synthetic benchmark of 500 geometric-scene images with programmatically verified annotations. The dataset combines 51 hand-designed， comprising one fully correct canonical scene and ten failure scenes for each of the five SRFE steps, and is expanded to 500 with procedurally generated variants (random seed 42) that vary layout, color, size, and spatial configuration while preserving ground-truth structure.

Each image describes two objects (cube, sphere, and triangle) with known attributes and is labeled with: a category (canonical, object, color, size, relation, or composition) and expected_failures --SRFE steps should receive FAIL under the intended perturbation.

The final split contains 1 canonical, 100 images per single-step failure category (object, color, size, relation), and 99 composition-failure images (500 total). All images are used for VPD

DOI: 10.54254/2755-2721/2026.BA37294

distillation, CoT supervised fine-tuning (SFT), and downstream evaluation. Synthetic control enables exact probe-level ground truth; generalization to natural image is the future work.

### 3.1.2. Stepwise reasoning faithfulness evaluation

In figure 1, SRFE is a five-step protocol in fixed prompt-defined order (object → color → size → relation → composition). For each step, the paper generates a ground-truth caption and a distractor caption from annotation templates. A vision–language scorer assigns softmax probabilities to both captions. The margin calculates by $\delta = score(gt) - score(distractor)$. The step passes if $\delta > 0$ (bit 1) and fails otherwise (bit 0). Concatenating the five bits yields a 5-bit probe signature (e. g., 11100).

SRFE has two Execution versions, independent version is all five steps are scored regardless of earlier outcomes. (Exp2,3,4), and conditional version is different with independent version on chain effect, if one step fails, then skip downstream. (Exp1)

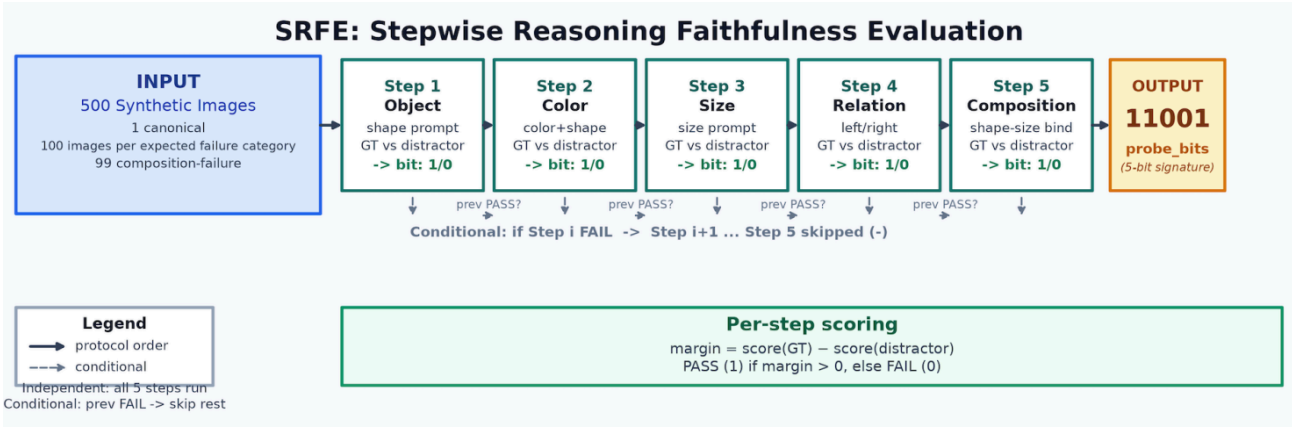



Figure 1. SRFE workflow (picture credit: original)

### 3.1.3. VPD

Directly running a full vision–language model (VLM) teacher on every probe is costly at scale. The paper therefore distills a frozen Qwen2-VL-2B-Instruct teacher into a lightweight student that preserves probe-level scoring behavior while keeping the backbone fixed.

Student add a visual prompt adapter (4 learnable prompt tokens fused into pooled vision features), and a linear projector mapping student vision features into the teacher text-embedding space.

Because teacher text features are cached and shared, only vision-side adapter and projector are trained.

Given an image and a probe caption pair (gtdistractor), the student computes the same pairwise margin as the teacher and assigns PASS/FAIL accordingly.

Training objective is $\mathcal{L}_{VPD} = \mathcal{L}_{KL} + \lambda_p \mathcal{L}_{probe}$ $\lambda_p$ =1, where $\mathcal{L}_{KL}$ distills vocabulary logits from the teacher (T=2), and $\mathcal{L}_{probe}$ is the MSE between student and teacher probe margins over all images and probe steps.

The paper trains on all 500 images for 30 epochs with AdamW (learning rate $5 \times 10^{-4}$, batch size 1). Only the visual prompt adapter and projector are updated.

DOI: 10.54254/2755-2721/2026.BA37294

Because loading two Qwen2-VL copies exceeds device memory, the paper pre-computes and caches teacher vocabulary logits and probe margins, then unload the teacher during student optimization.

The distilled student provides independent SRFE probe bits used as supervision targets in CoT faithfulness experiments. VPD does not supervise natural-language rationales.

### 3.1.4. CoT template and surpervised

Giving Qwen2-VL prompts to produce explicit chain-of-thought reasoning in a fixed five-block template: step kth <step name>: <rationale> Verdict: PASS or FAIL for $k \in \{1, \ldots, 5\}$, matching the SRFE step order (Objects, Colors, Sizes, Relation, Composition). At evaluation time (Experiments 3–4), CoT verdicts is generated with greedy decoding (do_sample=False, max_new_tokens=384).

When implementing CoT SFT, each training example pairs an image with a target verdicts which is the distilled student probe bits (independent mode). Step rationales are taken from the teacher's Exp3 CoT generations (rationale_source = teacher_cot). This combines the step-wise decision (verdict alignment) and step-wise observation (teacher text).

Training uses token-level cross-entropy on the full target CoT in each batch. Each token $t$ receives weight $w(t)$ :

$$w(t) = \begin{cases} wv \cdot mstep(t) & if\ t\ belongs\ to\ a\ verdict\ token, \\ wr & if\ t\ belongs\ to\ rationale\ text, \end{cases} \tag{1}$$

with $w_v = 4$, $w_r = 1$ and step multipliers $m_{\text{relation}} = 1.5$, $m_{\text{composition}} = 2.0$ (all other steps: $m = 1$ ), so the verdict weights are therefore 6× for relation and 8× for composition. The batch loss is the weighted mean cross-entropy over supervised tokens. The paper reported results use $w_r = 1$ and no chain penalty. Verdict-only and chain-penalty variants are left for future ablation.

The paper fine-tune Qwen2-VL-2B with LoRA ( $r = 16$, $\alpha = 32$ ) on all 500 supervised examples for 5 epochs (AdamW, learning rate $2 \times 10^{-4}$, batch size 1). The vision tower remains frozen.

Experiment 3 measures CoT–probe alignment before SFT; Experiment 4 repeats the same evaluation after SFT to quantify faithfulness gains.

### 3.1.5. Evaluation indicator

At the probe layer, which is used in Experiment 1 and in VPD distillation, the paper reports the step pass rate, that is, the proportion of PASS verdicts among the steps that are actually executed and the detection rate, which measures how often the probe returns FAIL on the step that a scene is designed to violate and PASS on the canonical scene. To compare two probes, the paper uses bits-exact agreement, the proportion of images whose full five-bit verdict patterns are identical, and the step agreement mean is the average rate of per-step when the two probes return the same verdict. Because the purpose of figuring out continuity of margin consistency, the paper also reports the MAE Pearson of the probe margins. Experiment 1 additionally reports four chain-specific quantities, which are chain integrity rate, the skip rate, the first-breakage alignment rate, and the difference in pass rate between the two modes.

At the CoT–probe alignment layer, which is used in Experiments 3 and 4, the paper first reports the parse success rate, the proportion of all five step verdicts can be successfully parsed. Alignment

DOI: 10.54254/2755-2721/2026.BA37294

is then measured by the mean step agreement between the CoT verdict and the probe verdict and by the bits agreement rate, the proportion of images on which all five verdicts match. Additionally, the paper focuses on the two directions of disagreement, because they carry different consequences. A say_do error occurs when the CoT reports PASS while the probe returns FAIL, an optimistic claim that the scene does not support, while a do_say error occurs when the CoT reports FAIL while the probe returns PASS, an unnecessarily pessimistic verdict. Finally, the order violation rate records how often a chain reports PASS on a step that follows an earlier FAIL, which is an explicit inconsistency within the stated chain.

## 3.2. Experimental setup

All experiments use Qwen2-VL-2B-Instruct as the backbone for both the teacher and the student, with LoRA adapters of rank r = 16, and run on either Apple MPS or CUDA hardware. Evaluation is carried out on the 500-image benchmark described in Section 3.1.1, which supports SRFE probing, CoT generation, and CoT–SRFE alignment alike. Fine-tuning uses five epochs, a learning rate of $2 \times 10^{-4}$, and a generation budget of 384 new tokens.

The four experiments build on one another. Experiment 1 compares independent and conditional probing on the student model in order to select an external reference. Experiment 2 tests whether SRFE verdicts depend on the order in which the probes are applied, comparing the canonical order against a reversed order, a relation-first order, and a per-image seeded random shuffle. Experiment 3 measures the alignment between the parsed CoT verdicts and the independent probe before any fine-tuning, and Experiment 4 applies probe-supervised SFT and evaluates the resulting alignment, taking the Experiment 3 student as the baseline.

## 3.3. Result

### 3.3.1. VPD distillation

Table 1. Distillation fidelity

| metric | T-S agreement |
|---|---|
| Step agree mean | 96.7% |
| Bits exact agreement | 85.0%(425/500) |
| Pass-rate MAE | 0.022 |
| Margin Person | 0.993 |

Table 1 reports teacher-student agreement for distillation fidelity (step agree, bits exact, margin consistency), these indicators show solid evidence of the success of distillation.

Table 2. Probe behavior (teacher vs student)

| step | Teacher pass rate | Student pass rate | delta |
|---|---|---|---|
| Object | 66.4% | 63.0% | -3.4% |
| Color | 82.2% | 78.6% | -3.6% |
| Size | 32.4% | 28.4% | -4.0% |

DOI: 10.54254/2755-2721/2026.BA37294

Table 2. (continued)

| | | | |
|---|---|---|---|
| Relation | 49.4% | 49.6% | +0.2% |
| composition | 44.6% | 44.6% | 0 |

Table 2 compares teacher and student pass rates in each step to verify that the student preserves probe behavior rather than only matching aggregate margins.

### 3.3.2. Experiment 1

Table 3. Independent vs. conditional result

| Step | Independent success | Skip rate | conditional success |
|---|---|---|---|
| Object | 63.0% | 0.0% | 63.0% |
| Color | 78.6% | 37.0% | 55.2% |
| Size | 28.4% | 44.8% | 12.8% |
| Relation | 49.6% | 87.2% | 9.2% |
| Composition | 44.6% | 90.8% | 3.8% |
| overall | 52.8% (mean) | 52.0% (mean) | 28.8% |

Table 3 compares the two probing modes. Conditional chaining enforces that once step $k$ fails, all subsequent steps are skipped, so averaged across the five steps 52.0% of step evaluations are never executed.

Table 4. Chain performance

| metric | value |
|---|---|
| exact five-bit agreement between modes | 9.2% (46/500) |
| chain completeness | 3.8% (19/500) |
| first-breakage alignment with the designed failure step | 25.6% (123/481) |

This weakens the conditional mode as a reference in two ways. First, table 4 shows the two modes rarely agree on the full verdict pattern, with only 9.2% of images having identical five-bit verdicts. Second, among the 481 images for which conditional probing has at least one FAIL, the first failure matches the designed failure step in only 25.6% of cases, so the FAIL point carries little information about where the scene actually breaks. Chain completeness is 3.8%, and this is expected because 499 of the 500 images contain at least one designed failure. The paper concludes that conditional chaining does not provide a reliable probe reference, and the paper uses independent probing in all subsequent experiments.

### 3.3.3. Experiment 2

In ablation experiment, all 500 images output identical 5-bits strings and mean probe-margins MAE is 0.0.

DOI: 10.54254/2755-2721/2026.BA37294

SRFE verdicts are order-invariant under independent probing—step order reflects protocol design, not recovered reasoning order.

### 3.3.4. Experiment 3

Table 5. Baseline CoT vs independent probe (n=500)

| indicator | value |
|---|---|
| Parse success rate | 100.0% |
| Step agreement mean | 58.1% |
| Bits agreement rate | 4.8% (24/500) |
| Order violation rate | 30.6% |
| Say_do | 1009 |
| Do_say | 38 |

Baseline CoT is partially aligned with independent SRFE probes: In table 5, step agreement mean is 58.1%, but only 4.8% of images have fully matching 5-bits verdicts. Errors of say_do highly take over do_say, which means Co tends to over-accept (especially relation/composition).

### 3.3.5. Experiment 4

Table 6. CoT-probe alignment before and after probe-supervised SFT

| indicator | Baseline[1] (before SFT) | Trained (after SFT) | delta |
|---|---|---|---|
| Parse success rate | 100.0% | 100.0% | - |
| Step agreement mean | 57.0% | 89.1% | +32.1pp |
| Bits agreement rate | 4.6% (23/500) | 59.6% (298/500) | +55.0pp |
| Order violation rate | 30.6% | 78.0% | +47.4pp |
| say_do | 1050 | 90 | -960 |
| do_say | 25 | 182 | +157 |

Baseline[1] : CoT is regenerated in experiment 4; metrics are consistent with baseline in experiment 3, only run-to-run variation.

After SFT, trained model probe-CoT alignment improved (bits agreement: 4.6% → 59.6%; step agreement: 57.0% → 89.1%; say_do: 1050 → 90). However, in table 6, say_do is 1050 and do_say is 25 before SFT, but when say_do declined to 90, do_say increased to 182, so a trade-off exists between say_do and do_say. In addition, order violations increase (30.6% → 78.0%), indicating the model becomes more probe-aligned but less chain-consistent in its explicit CoT.

### 3.3.6. Experiment results summary

Experiment 2 shows independent probe is order-invariant, experiment 1 shows conditional chaining is unreliable, so independent probe can be used as external reference reasonably. Experiment 3

DOI: 10.54254/2755-2721/2026.BA37294

evaluates the model's CoT ability to match with independent probe, and then experiment 4 shows probe-supervised SFT improves CoT-probe misalignment but encounters trade-offs between say-do/do-say, order violation. Verdicts category in experiement 3&4 can be found in table 7 below.

Table 7. Verdicts category (compare model probe-CoT alignment in exp3&4)

| Experiment | Verdict |
|---|---|
| Experiment 3 | Partially aligned |
| Experiment 4 baseline | Partially aligned |
| Experiment 4 trained | cot_aligned (bits ≥ 50% and step ≥ 75%) |

## 4. Conclusion

The paper introduces SRFE, an external 5-step probe protocol, as a reasoning reference for auditing explicit CoT verdicts. Protocol validation (Exp1-2) favors independent SRFE over conditional. Probe-supervised SFT can improve probe-CoT alignment but at the cost of chain-order consistency and a shift toward more pessimistic verdicts (do_say increases, say_do decreases). Future work should focus on training CoT to maintain chain-order consistency when preserving verdict faithfulness, reducing do_say steps and expanding SRFE on a broader visual reasoning task.

## References

[1] Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). *Reasoning models don't always say what they think* (arXiv:2505.05410). arXiv. https://doi.org/10.48550/arXiv.2505.05410.

[2] Hong, Y., Ravfogel, S., Zhao, C., & Choi, E. (2026). *Making LLMs say what they think: Measuring and improving CoT-interpretability alignment* [Paper presentation]. Mechanistic Interpretability Workshop at ICML 2026. https://openreview.net/pdf?id=iJnGfq4793

[3] Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., & Yuan, L. (2025). LLaVA-CoT: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2087–2098). IEEE. https://ieeexplore.ieee.org/document/11445580

[4] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837. https://www.proceedings.com/068431-1800.html

[5] Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems, 36*, 74952–74965. https://doi.org/10.48550/arXiv.2305.04388

[6] Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., ... Perez, E. (2023). *Measuring faithfulness in chain-of-thought reasoning* (arXiv:2307.13702). arXiv. https://doi.org/10.48550/arXiv.2307.13702

[7] Khalifa, M., Agarwal, R., Logeswaran, L., Kim, J., Peng, H., Lee, M., Lee, H., & Wang, L. (2025). *Process reward models that think* (arXiv:2504.16828). arXiv. https://doi.org/10.48550/arXiv.2504.16828

[8] Li, Y., Liu, Z., Li, Z., Zhang, X., Xu, Z., Chen, X., Shi, H., Jiang, S., Wang, X., Wang, J., Huang, S., Zhao, X., Jiang, B., Hong, L., Wang, L., Tian, Z., Huai, B., Luo, W., Luo, W., ... Zhang, M. (2025). *Perception, reason, think, and plan: A survey on large multimodal reasoning models* (arXiv:2505.04921). arXiv. https://doi.org/10.48550/arXiv.2505.04921

[9] Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., & Li, H. (2024). *Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning* (arXiv:2403.16999). arXiv. https://doi.org/10.48550/arXiv.2403.16999

DOI: 10.54254/2755-2721/2026.BA37294

[10] Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2901–2910). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8099698

[11] Shridhar, K., Stolfo, A., & Sachan, M. (2023). Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 7059–7073). Association for Computational Linguistics. https://aclanthology.org/2023.findings-acl.441.pdf