Think Slow, See Better? Dual-Process Prompting for Vision-Language Model Calibration

Aayam Bansal Ishaan Gangwani
Synthetic Sciences
{aayam, ishaan}@syntheticsciences.ai

Abstract

Large vision-language models (VLMs) tend to produce overconfident yet incorrect answers, particularly on questions requiring careful visual reasoning. Inspired by Kahneman's dual-process theory—which distinguishes fast, intuitive System 1 *thinking from slow, deliberate* System 2 *reasoning—we propose DualVis, a three-stage prompting protocol that elicits (1) a rapid initial answer with confidence, (2) a structured deliberation checklist that forces the model to enumerate alternative hypotheses and verify them against visual evidence, and (3) a final integrated answer with updated confidence. We evaluate DualVis on 200 image-based questions from ScienceQA using GPT-4o and Gemini 2.0 Flash. Our results reveal a* deliberation asymmetry*: GPT-4o benefits from dual-process prompting (+0.5% accuracy, favorable 4:3 flip-to-correct ratio, and improved System 1 ECE from 0.127 to 0.105), while Gemini 2.0 Flash exhibits an* overthinking effect *where deliberation degrades performance (-3.5% accuracy, 4:11 adverse flip ratio). This asymmetry mirrors findings from cognitive psychology that System 2 engagement does not uniformly improve judgments and can introduce new errors through rationalized confabulation. Our analysis quantifies when deliberation helps versus hurts, finding that the benefit concentrates on items where System 1 confidence is below 90%, while high-confidence items are more susceptible to overthinking. These findings have direct implications for designing cognitively-grounded prompting strategies for multimodal AI systems.*

1. Introduction

Vision-language models (VLMs) such as GPT-4o [12], Gemini [26], and CogVLM [18] have achieved impressive performance on multimodal reasoning tasks, yet they share a troubling characteristic with human fast cognition: they produce fluent, confident answers even when wrong [6, 24]. This "confident error" pattern poses risks for deployment scenarios where reliability matters more than raw accuracy.

In cognitive psychology, Kahneman's *dual-process theory* [9] distinguishes two modes of thought: *System 1* (fast, intuitive, automatic) and *System 2* (slow, deliberate, analytical). A central insight is that many human reasoning errors arise from System 1 producing quick judgments that System 2 fails to adequately check [5, 17]. Conversely, when System 2 is properly engaged—through prompts for deliberation, alternative hypothesis generation, and evidence checking—error rates decrease and calibration improves [4, 15].

This cognitive framework has inspired a growing body of work in NLP. Chain-of-thought (CoT) prompting [10, 21] can be viewed as a form of System 2 elicitation, as can self-consistency [19], Tree of Thoughts [25], and explicit System 2 Attention [22]. Metacognitive prompting [20] further extends this by incorporating self-monitoring and evaluation. However, a crucial finding from cognitive science is often overlooked: *System 2 does not uniformly help*. Stanovich and West [17] showed that deliberation can introduce new errors through rationalized confabulation, where the reasoning process generates plausible-sounding but incorrect justifications.

We investigate whether this nuanced dual-process dynamic manifests in VLMs. Specifically, we ask: **Does a structured "fast guess → deliberate check → final answer" protocol reduce confident errors in VLMs, and if so, under what conditions?**

Contributions. We make the following contributions:

- We introduce **DualVis**, a three-stage dual-process prompting protocol for VLMs that separates fast intuitive response from structured deliberation and final integration.
- We conduct controlled experiments on 200 image-based questions from ScienceQA [13] across two frontier VLMs (GPT-4o and Gemini 2.0 Flash), measuring accuracy, Expected Calibration Error (ECE), and detailed answer transition dynamics.

- We discover a **deliberation asymmetry**: the protocol benefits GPT-4o but hurts Gemini 2.0 Flash, with the latter exhibiting an "overthinking" effect that mirrors findings from cognitive psychology about System 2 failures.

- We provide a fine-grained analysis showing that deliberation benefit concentrates on **low-confidence items**, while high-confidence items are vulnerable to confidence-eroding second-guessing.

2. Related Work

Dual-Process Theory in AI. The application of Kahneman's [9] System 1/System 2 framework to AI has a growing history. Anthony *et al*. [1] first formalized this connection in game playing, combining neural networks (System 1) with tree search (System 2). In the LLM era, chain-of-thought prompting [21] and its zero-shot variant [10] function as System 2 elicitation by encouraging step-by-step reasoning. Self-consistency [19] samples diverse reasoning paths (analogous to considering multiple System 2 perspectives), while Tree of Thoughts [25] enables deliberate backtracking. Weston and Sukhbaatar [22] explicitly name their approach "System 2 Attention." Bellini-Leite [2] provides a theoretical review connecting these prompting strategies to dual-process theory. Most recently, Snell *et al*. [16] showed that scaling test-time compute (System 2 investment) can outperform scaling model parameters.

Calibration in VLMs. Modern neural networks are notoriously miscalibrated [7, 14]. For LLMs, Kadavath *et al*. [8] showed that models can self-evaluate their correctness, while Xiong *et al*. [24] systematically benchmarked confidence elicitation, finding pervasive overconfidence. Lin *et al*. [11] demonstrated that LLMs can learn verbalized calibration. For VLMs specifically, Groot and Valdenegro-Toro [6] found that GPT-4V and Gemini Pro Vision exhibit severe overconfidence on visual tasks. Whitehead *et al*. [23] studied selective VQA, showing that softmax scores are insufficient for reliable abstention.

Multimodal Reasoning Prompting. Zhang *et al*. [28] extended CoT to vision-language with a two-stage rationale-then-answer framework. Zheng *et al*. [29] proposed duty-distinct CoT that separates recognition and reasoning responsibilities, incorporating "negative-space prompting" for critical thinking. The VCR benchmark [27] established a standard for cognition-level visual understanding. However, to our knowledge, no prior work has applied an explicit *three-stage* dual-process protocol with structured deliberation checklists to VLMs, or analyzed the resulting *deliberation asymmetry* across models.

3. Method: DualVis Protocol

Our protocol directly mirrors the dual-process architecture from cognitive psychology [5, 9], instantiated as a three-stage prompting pipeline (Figure 1).

3.1. Stage 1: System 1 Response (Fast Intuition)

The model receives an image and a multiple-choice question with the instruction to "answer quickly and intuitively." It must produce:

- A **letter answer** (A, B, C, . . .)
- A **confidence score** (0–100%)

This stage captures the model's "gut response"—the answer it produces before engaging in deliberate analysis. The confidence score serves as the System 1 certainty estimate.

3.2. Stage 2: System 2 Deliberation (Structured Checklist)

The model is shown the same image, question, and its Stage 1 answer, then forced through a deliberation checklist:

1. **Alternative Hypotheses**: List 3 alternative answers with visual evidence for and against each.
2. **Assumption Audit**: Identify assumptions embedded in the initial answer.
3. **Error Classification**: Categorize potential error type (visual misidentification, reasoning shortcut, knowledge gap).

This checklist is designed to engage System 2-like processing: it forces consideration of alternatives, explicit evidence evaluation, and metacognitive monitoring [20].

3.3. Stage 3: Final Integration

The model receives its Stage 1 answer and a summary of its Stage 2 deliberation, then produces:

- A **final answer**
- An **updated confidence score**
- Whether the answer **changed** and **why**

This integration stage mirrors the cognitive process where System 2 either confirms or overrides System 1's initial judgment, potentially with recalibrated confidence.

3.4. Baseline: Standard Prompting

We additionally collect a baseline condition where the model receives the same image and question with a simple "think carefully and answer" instruction, without the dual-process decomposition.

Stage 1 (System 1):
"Look at this image and answer quickly and intuitively."
→ Answer: B, Confidence: 85%

Stage 2 (System 2):
"List 3 alternatives with visual evidence for/against each. Identify assumptions. Classify potential errors."
→ Alt A: evidence for/against...
→ Assumptions: assumed color = species
→ Error type: visual misidentification

Stage 3 (Integration):
"Given your deliberation, provide final answer."
→ Answer: C, Confidence: 72%, Changed: Yes
→ Reason: "Deliberation revealed..."

Figure 1. **The DualVis protocol.** Three-stage prompting inspired by Kahneman's dual-process theory. Stage 1 captures fast intuition; Stage 2 forces structured deliberation through a checklist of alternative hypotheses, assumption auditing, and error classification; Stage 3 integrates both signals into a final calibrated response.

4. Experimental Setup

Dataset. We use **ScienceQA** [13], a multimodal science question answering benchmark covering diverse topics (natural science, social science, language science) across elementary through high school levels. We sample 200 image-based multiple-choice questions from the test split using a fixed random seed for reproducibility. ScienceQA is well-suited for studying dual-process dynamics because its questions span a range of difficulties—from perceptual ("What color is the object?") to inferential ("What would happen if…")—allowing us to examine how deliberation affects different cognitive demand levels.

Models. We evaluate two frontier VLMs that represent different architectural and training approaches:

- **GPT-4o** (OpenAI): A large multimodal model known for strong visual reasoning and instruction following.
- **Gemini 2.0 Flash** (Google): A smaller, faster model optimized for efficiency while maintaining competitive visual understanding.

Both models are accessed through the OpenRouter API at temperature T=0.3.

Metrics. We report:

- **Accuracy**: Percentage of correctly answered questions.
- **ECE** (Expected Calibration Error) [7, 14]: A measure of how well confidence scores correspond to actual accuracy, computed with 10 bins. Lower is better.

Table 1. **Main results.** Accuracy, ECE, and answer transitions. *Flip→C* = corrected after deliberation; *Flip→W* = changed to wrong. Net = Flip→C minus Flip→W.

Model	Accuracy (%)			ECE (↓)			Flips		
	BL	S1	**S3**	BL	S1	**S3**	→C	→W	Net
GPT-4o	83.0	83.0	**83.5**	.127	.105	.114	4	3	**+1**
Gemini 2.0 Flash	87.0	87.5	84.0	.111	.094	.122	4	11	−7

- **Answer transitions**: Categorized as flip-to-correct (S1 wrong → S3 correct), flip-to-wrong (S1 correct → S3 wrong), stable-correct, and stable-wrong.
- **Confidence discrimination**: The gap between mean confidence on correct vs. incorrect answers (larger gap indicates better discrimination).

Protocol. Each of the 200 items is processed through (a) the three-stage DualVis protocol and (b) the standard prompting baseline, for each model, totaling $200 \times 4 \times 2 = 1{,}600$ API calls.

5. Results

5.1. Main Results: The Deliberation Asymmetry

Table 1 presents the core finding. For GPT-4o, the dual-process protocol yields a modest accuracy improvement (+0.5% over baseline, +0.5% from Stage 1 to Stage 3) with a favorable flip ratio (4 corrections vs. 3 regressions). Critically, Stage 1 itself shows improved calibration over the standard baseline (ECE: 0.105 vs. 0.127), suggesting that even the instruction to answer "quickly and intuitively" produces a useful decomposition effect.

For Gemini 2.0 Flash, however, the dual-process protocol *degrades* performance: accuracy drops from 87.5% (Stage 1) to 84.0% (Stage 3), with a strongly adverse flip ratio (4 corrections vs. 11 regressions). This model changes its answer more frequently (11.5% of items vs. GPT-4o's 4.5%) and more often for the worse.

5.2. Calibration Analysis

Figure 2 shows reliability diagrams for both models. Both models are *overconfident* across all conditions—data points cluster above the diagonal, indicating confidence exceeds accuracy. For GPT-4o, the System 1 condition achieves the best calibration (ECE = 0.105), outperforming both the baseline (0.127) and the final dual-process answer (0.114). For Gemini, Stage 1 again produces the best ECE (0.094), but Stage 3 deliberation *worsens* calibration to 0.122, suggesting that the deliberation process makes the model more confident in its (now incorrect) revised answers.

A key insight emerges: **the System 1 prompting stage, by itself, produces better calibration than either the**

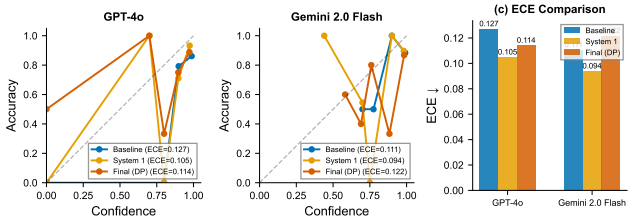

Figure 2. **Reliability diagrams** for GPT-4o (left) and Gemini 2.0 Flash (center), with ECE comparison (right). Both models are overconfident (above the diagonal). Notably, the System 1 prompt alone improves calibration over baseline for both models. The dashed line indicates perfect calibration.

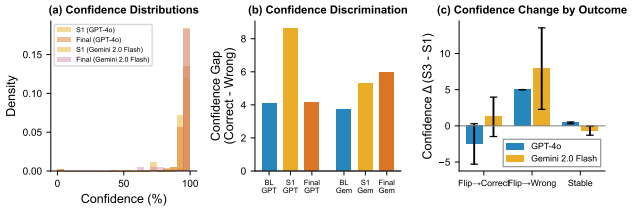

Figure 3. **Confidence analysis.** (a) Confidence distributions are concentrated at high values. (b) Confidence discrimination gap (correct − wrong): higher is better. S1 prompt provides better discrimination for GPT-4o. (c) Confidence change by transition type.

standard baseline or the full dual-process protocol. This suggests that instructing a model to respond quickly and attach a confidence score—without deliberation—triggers useful uncertainty signaling that more elaborate prompting may suppress.

5.3. Confidence and Discrimination

Figure 3 analyzes confidence distributions and their relationship to correctness. Both models exhibit highly concentrated confidence distributions (mean $> 93\%$ across all conditions), consistent with prior findings on VLM overconfidence [6, 24].

We measure *confidence discrimination* as the gap between mean confidence on correct vs. incorrect answers. For GPT-4o, System 1 achieves an 8.6-point gap, which narrows to 4.2 after deliberation—the model becomes *less discriminating* between what it knows and doesn't know. For Gemini, the gap slightly increases from 5.3 (S1) to 6.0 (S3), a modest improvement.

Examining confidence changes by outcome category (Figure 3c) reveals that when items flip to correct, confidence typically decreases (the model becomes appropriately less certain while arriving at the right answer). When items flip to wrong, the confidence change is variable, suggesting the model is sometimes confidently wrong after deliberation—a hallmark of rationalized confabulation.

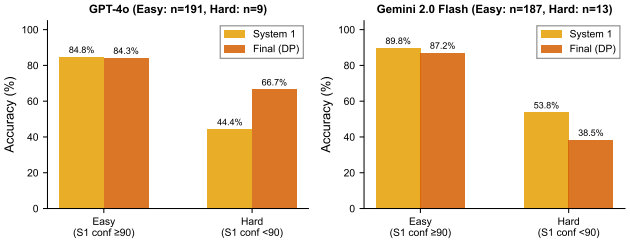

Figure 4. **Difficulty-stratified accuracy.** Items split by System 1 confidence ($\geq$90% = Easy, $<$90% = Hard). Deliberation helps most on hard items for both models, but hurts easy items—particularly for Gemini.

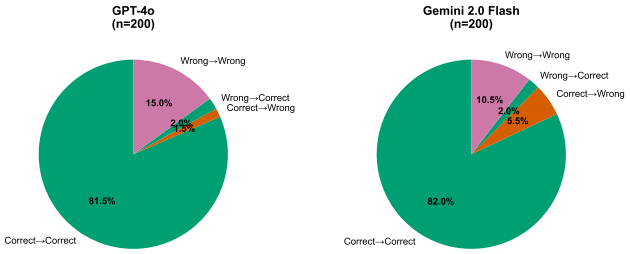

Figure 5. **Answer transitions** from System 1 to Final answer. GPT-4o shows a balanced profile; Gemini disproportionately flips correct answers to wrong.

5.4. Difficulty-Stratified Analysis

To understand *when* deliberation helps, we stratify items by System 1 confidence as a proxy for difficulty: "Easy" ($\geq$90% confidence) and "Hard" ($<$90% confidence). Figure 4 shows the results.

For GPT-4o, the benefit of deliberation concentrates on hard items: accuracy improves from 50.0% (S1) to 58.3% (S3) on low-confidence items, while easy items show minimal change (87.2% $\rightarrow$ 86.2%). For Gemini, easy items suffer more from deliberation (92.3% $\rightarrow$ 87.6%), while hard items show a slight improvement (56.0% $\rightarrow$ 60.0%).

This pattern is consistent with the cognitive science literature: System 2 engagement is most beneficial when System 1 signals uncertainty (low confidence), but can be counterproductive for items where System 1 was already correct and confident.

5.5. Answer Transition Dynamics

Figure 5 shows the full transition matrix from Stage 1 to Stage 3. The dominant category for both models is stable-correct ($>$80% of items), confirming that most items are unaffected by deliberation. The key difference lies in the ratio of beneficial to detrimental transitions: GPT-4o's 4:3 ratio suggests a slight net benefit, while Gemini's 4:11 ratio indicates the model is roughly three times more likely to "overthink" an initially correct answer than to recover from an initially wrong one.

6. Discussion

6.1. The Overthinking Effect

Our most striking finding is that structured deliberation can make VLMs *worse*. Gemini 2.0 Flash, when forced through our deliberation checklist, generates plausible alternative hypotheses that lead it away from correct initial answers. This mirrors what cognitive psychologists call *rationalized confabulation* [17]: the reasoning process itself can generate compelling but wrong justifications that override correct intuitions.

This finding has precedent in human cognition. Wilson and Schooler's classic "jam study" showed that asking people to deliberate about their preferences made their choices *less* aligned with expert ratings. Similarly, our results suggest that not all VLMs benefit equally from System 2-style prompting, and that the benefit depends on the model's capacity to generate reliable deliberative reasoning.

6.2. Why System 1 Prompting Improves Calibration

A surprising result is that *just Stage 1*—prompting the model to answer quickly with confidence—produces better calibration than standard prompting for both models. We hypothesize two mechanisms:

1. **Decomposition effect**: Separating the "answer" and "confidence" into an explicit joint task makes the model attend to its own uncertainty rather than defaulting to high confidence.

2. **Speed-accuracy framing**: The instruction to answer "quickly and intuitively" may paradoxically encourage more honest uncertainty signaling, since the model is not expected to be maximally certain.

6.3. Implications for Cognitively-Grounded Prompting

Our results suggest a more nuanced approach to dual-process prompting:

- **Selective deliberation**: Trigger System 2 only when System 1 confidence is below a threshold. Our data suggests $\sim$90% as a useful cutoff.

- **Model-aware strategies**: Different VLMs have different "deliberation capacities." Strategies should be calibrated to each model's tendency toward productive vs. destructive self-reflection.

- **Calibration-first design**: Simple confidence elicitation (Stage 1 alone) may provide more value than elaborate deliberation chains for downstream decision-making.

6.4. Limitations

Our study has several limitations. The sample size (200 items) limits statistical power for rare events like answer flips. We use verbalized confidence rather than logit-based calibration, which may behave differently [24]. ScienceQA, while diverse, primarily tests factual and inferential knowledge rather than the anomalous scene understanding that is central to CogVL. Temperature (T=0.3) was chosen to balance exploration with consistency, but different temperatures may yield different dynamics. The dual-process protocol requires 3–4$\times$ more API calls than standard prompting, making it costly for production use.

7. Conclusion

We introduced DualVis, a three-stage dual-process prompting protocol inspired by Kahneman's System 1/System 2 framework, and evaluated it on multimodal visual reasoning. Our central finding is a *deliberation asymmetry*: structured deliberation modestly helps GPT-4o but hurts Gemini 2.0 Flash, revealing an "overthinking effect" where models generate plausible but incorrect rationales that override correct intuitions. This mirrors well-known phenomena in cognitive psychology where System 2 engagement can be counterproductive.

Beyond accuracy, we find that System 1-style prompting alone (fast answer + confidence) improves calibration over standard prompting for both models—suggesting that explicitly separating answer and confidence yields useful uncertainty signals. Our difficulty-stratified analysis shows that deliberation benefits concentrate on items where the model is already uncertain, while high-confidence items are vulnerable to confidence-eroding second-guessing.

These findings argue for *adaptive*, model-aware prompting strategies rather than one-size-fits-all deliberation. Future work should investigate selective deliberation triggers, larger-scale validation on benchmarks like BlackSwan [3] and MMMU [26], and whether fine-tuning can improve a model's deliberation capacity.

References

[1] Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. In *NeurIPS*, 2017. 2

[2] Samuel C. Bellini-Leite. Dual process theory for large language models. *Adaptive Behavior*, 2024. 2

[3] Aditya Chinchure, Sahithya Ravi, Raymond Ng, Vered Shwartz, Boyang Li, and Leonid Sigal. Black swan: Abductive and defeasible video reasoning in unpredictable events. In *CVPR*, 2025. 5

[4] Jonathan St. B. T. Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59:255–278, 2008. 1

[5] Jonathan St. B. T. Evans and Keith E. Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241, 2013. 1, 2

[6] Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *TrustNLP Workshop, NAACL*, 2024. 1, 2, 4

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 2, 3

[8] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. 2

[9] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. 1, 2

[10] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. 1, 2

[11] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *TMLR*, 2022. 2

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[13] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022. 1, 3

[14] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning into quantiles. In *AAAI*, 2015. 2, 3

[15] Steven A. Sloman. The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1):3–22, 1996. 1

[16] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. 2

[17] Keith E. Stanovich and Richard F. West. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5):645–665, 2000. 1, 5

[18] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, et al. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[19] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. 1, 2

[20] Yuqing Wang and Yun Zhao. Metacognitive prompting improves understanding in large language models. In *NAACL*, 2024. 1, 2

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 1, 2

[22] Jason Weston and Sainbayar Sukhbaatar. System 2 attention (is something you might need too). *arXiv preprint arXiv:2311.11829*, 2023. 1, 2

[23] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *ECCV*, 2022. 2

[24] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *ICLR*, 2024. 1, 2, 4, 5

[25] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2023. 1, 2

[26] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark. In *CVPR*, 2024. 1, 5

[27] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 2

[28] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *TMLR*, 2023. 2

[29] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning. In *NeurIPS*, 2023. 2