BEYOND CORRECTNESS: BENCHMARKING AND ALIGNING RESPONSE BEHAVIORS IN HYBRID-THINKING MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid-thinking multimodal large language models (MLLMs) support thinking and non-thinking modes, but answer correctness alone does not capture the quality of their final responses. We introduce **PatternEval**, a failure-enriched diagnostic benchmark of 2,415 multimodal prompts across nine task categories. It evaluates correctness separately from four response-pattern failures: chain-of-thought leakage, response repetition, logical contradiction, and performative reasoning. Across 17 evaluated model configurations, non-thinking response-pattern failure rates exceed those in thinking mode by 2.86–48.64 percentage points, motivating direct evaluation of the response delivered through each interface. We further develop **PatternRM**, a reward model that supplies pattern-specific penalties to augment verifier-based group relative policy optimization (GRPO). Incorporating PatternRM into verifier-based reinforcement learning reduces non-thinking response-pattern failures while largely preserving accuracy on PatternEval. Comparison with an explicit length penalty reveals a trade-off: length regularization achieves lower aggregate failure rates, whereas GRPO with PatternRM achieves higher downstream scores while reducing failures relative to the verifier-based baseline. Together, our findings highlight the value of response-pattern metrics for evaluating reinforcement learning beyond task accuracy and demonstrate that integrating pattern-aware reward modeling with verifier-based training can improve both response quality and downstream task performance.

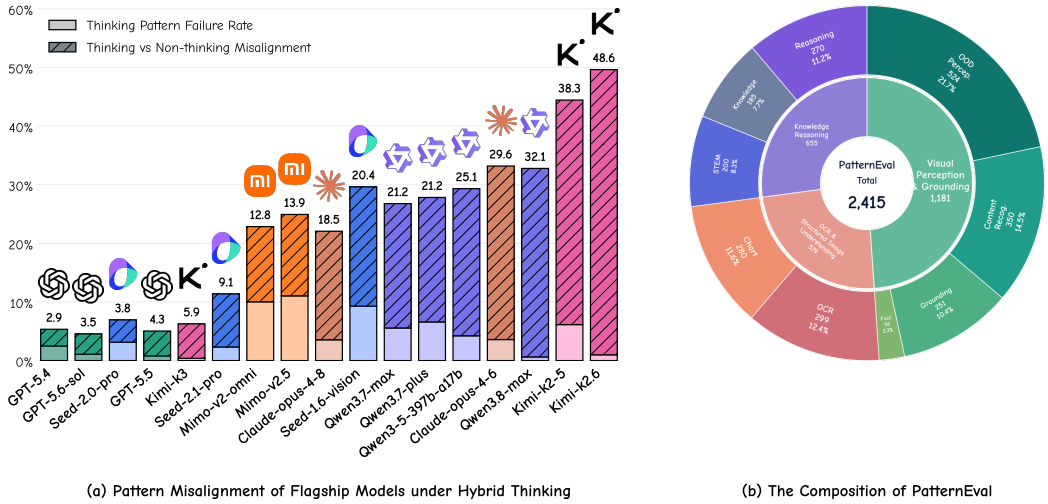
1 INTRODUCTION

A model can reach the right answer and still deliver a poor response: it may repeat claims, contradict itself, or fill a direct answer with unfinished reasoning. Users consume the full response, so correctness captures only part of its quality (Liang et al., 2022; Zhou et al., 2023). This distinction matters for hybrid-thinking systems, which expose thinking and non-thinking modes (Qwen Team, 2025; GLM-4.5 Team, 2025). Either mode may serve the same image and question; although the appropriate explanation can differ, both should deliver coherent responses without unnecessary repetition.

Recent work identifies residual deliberation in non-thinking outputs and explores data design, length control, and architectural separation to improve mode control (Wang et al., 2025c; Gan et al., 2026; Wang et al., 2026). This motivates our question: *how can we measure failures in final multimodal responses and use those measurements to guide post-training?* We study four **response-pattern failures**: chain-of-thought leakage, response repetition, logical contradiction, and performative reasoning (Figure 2). These labels concern observable behavior in the delivered answer; useful explanations and detailed reasoning remain appropriate.

We introduce **PatternEval**, a failure-enriched diagnostic benchmark of 2,415 multimodal prompts across nine task categories. It records correctness and the four failure labels separately under matched thinking and non-thinking interfaces. Across multiple evaluated model configurations, the rate of responses with at least one failure is higher in non-thinking mode by 2.86–48.64 percentage points. This gap motivates evaluating each interface's final responses directly.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Pattern Misalignment of Flagship Models under Hybrid Thinking

(b) The Composition of PatternEval

Figure 1: Panel (a) compares response-pattern failure rates between thinking and non-thinking modes across 17 model configurations, and panel (b) shows the composition of PatternEval.

To translate this diagnosis into training feedback, we develop **PatternRM**, a response-level reward model that evaluates the four failure categories. We incorporate its pattern-specific penalties into verifier-based reinforcement learning, retaining answer correctness as the primary reward while providing a complementary signal for improving response behavior. We also compare against explicit length penalization to distinguish the benefits of targeting undesirable response patterns from those of simply encouraging shorter outputs.

Experiments on Qwen3-VL-4B and 8B show that incorporating PatternRM into verifier-based reinforcement learning reduces non-thinking response-pattern failure rates by 11.93 and 15.49 percentage points relative to GRPO using only the verifier reward, respectively, while leaving PatternEval accuracy nearly unchanged. Average scores across 11 downstream benchmarks improve by 9.01 and 4.76 points, respectively. Although explicit length penalization achieves lower aggregate failure rates, GRPO with PatternRM yields higher downstream scores, showing different trade-offs between failure suppression and task performance. Together, these results connect response-pattern diagnosis to reward-model-based training, highlighting the value of behavioral metrics beyond task accuracy and showing that pattern-aware rewards can complement verifier-based optimization to improve both response quality and downstream performance relative to GRPO using only the verifier reward.

2 Related Work

Reasoning effort and mode control. Chain-of-thought prompting and self-consistency elicit or sample intermediate reasoning (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023b); reasoning-specialized models extend this use of test-time computation (DeepSeek-AI, 2025; He et al., 2025; Xu et al., 2025). Hybrid approaches learn when to reason (Fang et al., 2025; Wang et al., 2025a) or regulate reasoning budgets (Wen et al., 2025; Luo et al., 2026; Zhang et al., 2025a); direct answering can remain competitive under tight budgets (Ma et al., 2025). Closer to mode control, Demystifying Hybrid Thinking studies data recipes (Wang et al., 2025c), TNT uses query-specific length limits (Gan et al., 2026), and Path-Lock Expert separates mode-specific computation (Wang et al., 2026). We connect a taxonomy of final-response failures across multimodal interfaces to an explicit post-training signal.

Evaluating final multimodal responses. Multimodal benchmarks measure perception and reasoning (Fu et al., 2025; Liu et al., 2023b; Yue et al., 2023), object hallucination and visual consistency (Li et al., 2023; Guan et al., 2023; Wang et al., 2023a), and broader trustworthiness (Zhang et al., 2025b). Text-based evaluations cover truthfulness (Lin et al., 2021) and safety (Röttger et al., 2024), while consistency-based methods detect hallucinated content (Manakul et al., 2023). Pat-

ternEval complements these evaluations by recording repetition, contradiction, and inappropriate reasoning presentation separately from correctness, and comparing these failures across thinking interfaces.

Reward models and behavioral feedback. RewardBench (Lambert et al., 2025) and RewardBench 2 (Malik et al., 2025) assess language-model reward judgments; Multimodal RewardBench (Yasunaga et al., 2025) evaluates vision–language reward models. PatternRM predicts PatternEval's explicit failure labels and supplies auxiliary feedback for verifier-based training. We compare GRPO with PatternRM against GRPO using only the verifier reward and GRPO with a length penalty, evaluating both response behavior and downstream performance.

3 PATTERNEVAL: BENCHMARKING BEYOND ANSWER CORRECTNESS

3.1 OVERVIEW

PatternEval comprises 2,415 image–question pairs drawn from multimodal task pools covering visual perception and grounding, OCR and structured-image understanding, and knowledge reasoning. We construct the diagnostic set by inspecting rollout failures, expanding related tasks and response forms, and selecting examples across nine categories while capping high-volume sources. The three task families contain 1,181, 579, and 655 questions, respectively. This failure-enriched design targets recurring response problems across diverse inputs. The category counts underlying the composition overview in Figure 1 are provided in Appendix B.

We annotate four observable response-pattern failures, illustrated in Figure 2: **CoT leakage**, or unintended process narration in the final answer; **response repetition**, or unnecessary repeated content; **logical contradiction**, or unresolved incompatible claims; and **performative reasoning**, or reasoning-like content without supporting evidence. Studies of text degeneration and repetition motivate explicitly measuring such failures (Finlayson et al., 2024; Hiraoka & Inui, 2025). Polished explanations and useful restatements are allowed. The labels describe the delivered response; they do not establish faithfulness to the model's internal computation, a separate concern in reasoning evaluation (Chen et al., 2025). Detailed definitions and label-attribution rules appear in Appendix B.1; qualitative response comparisons appear in Appendix I.

3.2 JUDGE PROTOCOL

For each image–question pair, we evaluate the final user-visible response under matched thinking and non-thinking interfaces. Any separately returned reasoning channel is excluded. Model-based evaluation provides a scalable way to assess responses (Zheng et al., 2023); prompted criteria and explicit rubrics make the target properties more specific (Liu et al., 2023a; Fan et al., 2024). We use two judges with distinct roles: **Qwen3-Max** assesses answer correctness using the question, response, reference answer, and source-task criteria; **Seed-2.0-Pro** assesses response-pattern failures using the image, question, and response. The pattern judge returns four binary labels with brief justifications under a shared rubric, reproduced in Appendix J.

Let c_i indicate answer correctness and b_{ik} indicate pattern k. We report

$$\text{Acc} = \frac{1}{|V|}\sum_{i\in V} c_i, \qquad \text{Trigger} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\max_k b_{ik} = 1], \tag{1}$$

where V contains examples with valid correctness references. Accuracy measures answer correctness; Trigger measures whether any failure occurs. Labels may co-occur, so their rates are not summed. Denominators and attribution details are given in Appendix B.2.

3.3 META-JUDGE EVALUATION

Automatic judgments can be sensitive to response length (Dubois et al., 2024), evaluator self-preference (Panickssery et al., 2024), and reasoning-related judging biases (Wang et al., 2025b). We therefore evaluate candidate pattern judges against human reference labels on 2,500 fixed responses drawn from the same source tasks. The set combines three-judge consensus and disputed



Figure 2: The four columns contrast thinking and non-thinking responses for CoT leakage, repetition, contradiction, and performative reasoning. Highlighted passages illustrate the response behaviors targeted by PatternEval.

Table 1: Meta-judge performance is measured by precision, recall, and F1 (%). Overall scores are macro-averages across four bad pattern labels. Bold and underlined values indicate the best and second-best scores, respectively.

Judge	Input	CoT			Rep			Con			PR			Overall		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT-5.5	Image + Text	98.5	**92.4**	**95.3**	83.5	92.9	87.9	46.3	85.5	60.1	54.4	**79.3**	**64.5**	70.7	**87.5**	77.0
	Text	98.4	92.2	95.2	83.2	**93.8**	**88.2**	47.5	**88.9**	61.9	55.2	75.0	63.6	71.1	**87.5**	**77.2**
Seed-2.0-Pro	Image + Text	98.6	84.8	91.2	83.5	91.5	87.3	53.7	74.8	**62.5**	50.4	74.1	60.0	71.6	81.3	75.3
	Text	98.6	84.7	91.1	83.4	91.6	87.3	55.6	67.5	61.0	44.8	68.1	54.0	70.6	78.0	73.4
Hy3	Text	**98.7**	83.0	90.2	84.2	80.7	82.4	48.8	69.8	57.4	49.5	57.1	53.0	70.3	72.7	70.8
Kimi-K2.6	Image + Text	97.7	89.1	93.2	**87.1**	76.8	81.6	**55.8**	70.1	62.1	58.4	40.7	48.0	**74.8**	69.2	71.2
	Text	97.7	87.7	92.4	86.5	79.2	82.7	49.4	65.0	56.1	**64.1**	36.0	46.1	74.4	67.0	69.3
Qwen3.5-397B	Image + Text	97.6	87.3	92.2	86.4	84.6	85.5	44.7	75.2	56.1	36.2	61.6	45.6	66.2	77.2	69.9
	Text	97.5	88.5	92.8	83.2	88.6	85.8	52.0	77.8	62.3	55.6	33.5	41.8	72.1	72.1	70.7

responses at approximately 7:3, with a predicted positive rate near 70%. Sampling and annotation details appear in Appendix D.

We report precision, recall, and F1 for five candidate judges across nine image–text and text-only configurations in Table 1. CoT leakage and repetition yield higher F1 scores than contradiction and performative reasoning. GPT-5.5 achieves the highest macro-F1, while Seed-2.0-Pro with image access achieves the highest contradiction F1 (62.5%) and a macro-F1 of 75.3%. We use the latter as the operational pattern judge, retaining access to the visual evidence. Additional analysis appears in Appendix D.

4 Response Quality Across Inference Modes

4.1 Experimental Setup

We evaluate models from six families—Qwen, Kimi, Seed, Claude, MiMo, and GPT—on the same 2,415 PatternEval prompts in thinking and non-thinking modes. Model-specific sampling parameters and reasoning controls are listed in Table 7 in Appendix B.3; these settings may differ between modes. Hosted systems are evaluated through the controls available in their APIs. Only the final user-visible response is scored, excluding any separately returned reasoning channel, following Section 3.2. The comparison therefore characterizes the deployed interfaces, including their separation of reasoning from the final answer.

We report accuracy, Trigger, individual pattern rates, and task-family Trigger in Table 2. The mode gaps are defined as thinking minus non-thinking accuracy and non-thinking minus thinking Trigger, both in percentage points. The Qwen3.5 scaling and response-length analyses use their original evaluation cohorts; category-wise results for the earlier 25-configuration cohort appear in Appendix C.

4.2 Results and Analysis

Non-thinking interfaces exhibit higher response-pattern failure rates. Table 2 provides the detailed results behind the cross-mode overview in Figure 1(a). Every model pair has higher Trigger in non-thinking mode, with gaps ranging from 2.86 to 48.64 percentage points. The magnitude varies across interfaces: some keep failures rare in both modes, whereas others show gaps exceeding 20 points. Thus, the overview reflects a consistent direction of difference across the surveyed interfaces, with substantial variation in its size.

Table 2: Thinking and non-thinking modes are compared on PatternEval. Scores are percentages, and gaps are percentage-point differences. Blue and orange indicate the best and worst results within each mode, respectively.

Model	Mode	Mode Gap		Overall Results		Pattern Failures				Task-family Trigger		
		Δ_{acc}	Δ_{pat}	Acc↑	Trigger↓	CoT↓	Rep↓	Con↓	PR↓	VG↓	OS↓	KR↓
Qwen3.5-397B-A17B	Non-thk	6.97	25.09	61.96	29.32	25.71	15.65	0.66	1.12	30.51	9.48	44.73
	Thk			68.93	4.23	2.98	1.12	0.17	0.83	4.75	0.69	6.42
Qwen3.7-Max	Non-thk	5.23	21.21	67.07	26.77	22.28	16.79	0.46	0.58	29.69	13.26	33.33
	Thk			72.30	5.56	3.98	2.70	0.17	0.75	6.54	2.25	6.73
Qwen3.7-Plus	Non-thk	4.69	21.23	66.28	27.82	23.13	17.50	0.58	0.66	29.24	13.99	37.52
	Thk			70.97	6.59	**4.85**	**3.90**	0.21	0.70	7.63	1.73	9.04
Qwen3.8-Max	Non-thk	6.29	32.15	**71.96**	32.77	28.83	**24.57**	**0.12**	**0.33**	40.08	5.17	44.04
	Thk			78.25	0.62	**0.00**	0.25	0.21	0.21	0.25	0.69	1.22
Kimi-K2.5	Non-thk	5.73	38.29	61.33	44.43	42.07	11.59	0.37	0.91	44.24	24.14	62.75
	Thk			67.06	6.14	3.55	1.00	1.09	1.34	6.69	2.76	8.19
Kimi-K2.6	Non-thk	6.72	48.64	66.47	**49.63**	**47.39**	16.74	0.54	0.50	**47.92**	**28.45**	**71.45**
	Thk			73.19	0.99	0.13	0.13	0.27	0.49	1.41	0.34	0.85
Kimi-K3	Non-thk	10.99	5.88	65.47	6.26	3.15	1.57	1.66	1.62	6.69	1.90	9.34
	Thk			76.46	**0.38**	0.04	0.17	0.13	**0.09**	**0.09**	1.03	**0.31**
Seed-1.6-Vision	Non-thk	2.24	20.35	59.49	29.63	24.28	20.77	0.71	3.34	41.38	10.52	25.31
	Thk			61.73	9.28	2.84	3.55	1.63	4.05	14.10	**3.10**	5.97
Seed-2.0-Pro	Non-thk	6.13	3.81	66.85	6.94	4.51	0.75	1.34	1.04	7.31	3.62	9.28
	Thk			72.98	3.13	2.09	0.25	0.21	0.67	3.91	0.86	3.77
Seed-2.1-Pro	Non-thk	13.10	9.06	70.44	11.37	8.90	3.05	1.25	0.54	12.83	3.45	15.88
	Thk			**83.54**	2.31	1.72	0.34	0.05	0.39	2.14	2.24	2.67
Claude-Opus-4.6	Non-thk	8.18	29.59	59.45	33.18	30.35	8.79	1.23	1.27	39.48	5.37	47.12
	Thk			67.63	3.59	1.45	0.23	0.56	1.49	4.83	1.22	3.84
Claude-Opus-4.8	Non-thk	6.58	18.51	61.12	22.06	18.93	5.11	1.18	0.89	29.01	4.68	25.16
	Thk			67.70	3.55	1.93	0.48	0.44	1.01	4.24	1.56	4.18
Mimo-v2-Omni	Non-thk	0.31	12.83	57.21	22.85	16.37	3.74	3.66	4.03	27.23	14.19	22.63
	Thk			57.52	10.02	3.50	1.30	1.94	5.10	13.99	2.71	9.48
Mimo-v2.5	Non-thk	0.42	13.88	55.03	24.92	16.47	3.87	5.03	4.83	34.22	8.65	22.63
	Thk			55.45	11.04	3.41	2.34	**2.13**	6.52	15.70	2.29	**10.47**
GPT-5.4	Non-thk	8.88	2.86	60.66	5.34	1.20	0.91	0.79	2.82	**5.68**	1.21	8.40
	Thk			69.54	2.48	**0.00**	0.62	**0.04**	1.82	3.56	0.86	1.98
GPT-5.5	Non-thk	15.70	4.26	59.20	5.04	0.29	0.42	0.75	4.00	7.32	1.21	4.31
	Thk			74.90	0.78	**0.00**	0.18	0.14	0.46	1.17	0.52	0.34
GPT-5.6-Sol	Non-thk	13.49	3.47	62.25	**4.55**	**0.12**	**0.04**	0.75	3.73	6.86	**0.69**	**3.82**
	Thk			75.74	1.08	**0.00**	**0.08**	0.08	0.91	1.86	**0.17**	0.46

High accuracy does not ensure low Trigger. Qwen3.8-Max reaches 71.96% non-thinking accuracy while triggering at least one pattern on 32.77% of responses; its thinking interface has 0.62% Trigger. The pattern breakdown further shows that failures differ in composition: CoT leakage and repetition dominate this model's non-thinking responses, whereas performative reasoning is the largest component for GPT-5.5 and GPT-5.6-Sol. These results show that evaluating final-response behavior adds information beyond task accuracy.

Within-family scaling leaves a persistent non-thinking gap. Figure 3 compares seven Qwen3.5 variants. Accuracy generally increases with scale, while Trigger initially decreases in both modes. From 27B onward, however, thinking-mode Trigger stays below 7%, whereas non-thinking Trigger remains between 28.5% and 33.1%. Non-thinking accuracy rises from 59.1% at 27B to 62.0%

at 397B-A17B, yet Trigger is 29.3% for both. Stronger task performance within the same family therefore does not ensure fewer response-pattern failures in direct answers.

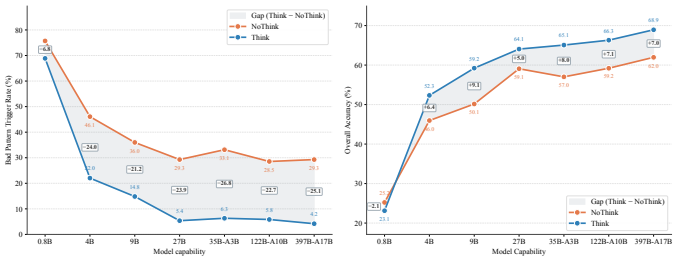


Figure 3: Trigger and accuracy vary across the Qwen3.5 series. Shaded annotations show thinking–non-thinking gaps in percentage points.

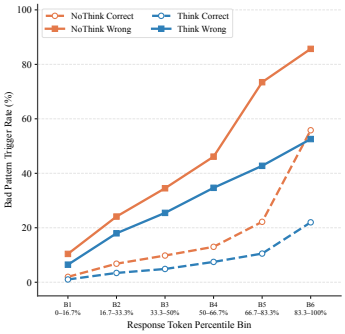


Figure 4: Trigger increases across response-length in both modes and correctness groups.

Longer responses exhibit more failures even when answers are correct. In Figure 4, Trigger rises across response-length sextiles for both correct and incorrect answers in both modes. The trend among correct answers shows that the length association extends beyond incorrect responses. Model-level length correlations provide a complementary view in Appendix C.2. These descriptive associations motivate the explicit length-penalty comparison in Section 5, which examines whether targeting response patterns offers benefits beyond encouraging shorter outputs.

5 Pattern-Aware Reinforcement Learning

This section evaluates whether PatternRM improves non-thinking responses when incorporated into verifier-based reinforcement learning. The comparison with an explicit length penalty examines the trade-off between response-pattern suppression and downstream task performance.

5.1 Learning a Behavioral Reward

Our pattern-aware GRPO recipe combines a task verifier with PatternRM, a reward model trained to detect the four response-pattern failures. AI feedback has been used to construct supervision for multimodal alignment (Li et al., 2024); here, the targets are the explicit failure labels defined by PatternEval. We annotate a heterogeneous pool of 90K responses with three judges and retain complete four-label agreement, yielding 52,344 unique examples. Oversampling 5,234 selected examples gives 57,578 supervised instances. PatternRM is initialized from Qwen3.5-27B and trained to predict the labels from response text. Its NoThink mode reaches 71.3% macro-F1 on the human-annotated meta-judge evaluation set, compared with 70.2% in Think mode, and avoids generating an additional reasoning trace. We use NoThink mode for rollout scoring; full validation and corpus details appear in Table 3 and Appendix E.1.

Reward design. For a correctness reward s_{ver}, PatternRM predictions $\hat{b}_k$, and an invocation indicator $z \sim \text{Bernoulli}(0.6)$, we use

$$P = \min\left(0.1, \sum_k w_k \hat{b}_k\right), \qquad r = \text{clip}(s_{\text{ver}} - zP, 0, 1). \tag{2}$$

The weights are 0.05 for CoT and repetition and 0.02 for contradiction and performative reasoning. Under the binary verifier formulation, this becomes $r = s_{\text{ver}}(1 - zP)$: correct responses receive 0.9 to 1 and incorrect responses receive zero. The objective then distinguishes cleaner answers within the verifier-correct set while retaining correctness as the primary reward ordering. The cap bounds the raw reward adjustment; group-normalized advantages determine the policy update. Verifier implementation details are discussed in Appendix G.

Table 3: PatternRM and reference judges are evaluated using precision, recall, and F1 (%). Overall scores are macro-averages across four labels. Bold and underlined values indicate the best and second-best scores, excluding Seed reference rows.

Judge	Input	CoT			Rep			Con			PR			Overall		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Seed-2.0-Pro	Image + Text	98.6	84.8	91.2	83.5	91.5	87.3	53.7	74.8	62.5	50.4	74.1	60.0	71.6	81.3	75.3
	Text	98.6	84.7	91.1	83.4	91.6	87.3	55.6	67.5	61.0	44.8	68.1	54.0	70.6	78.0	73.4
Qwen3.5-27B-Think	Image + Text	97.0	88.3	92.4	81.3	88.1	**84.6**	45.7	**75.5**	56.9	48.0	48.0	48.0	68.0	75.0	70.5
	Text	97.4	88.6	92.8	79.2	88.2	83.5	53.8	66.7	**59.5**	**63.2**	29.3	40.0	73.4	68.2	69.0
Qwen3.5-27B-NoThink	Image + Text	97.8	86.4	91.7	75.8	79.6	77.7	33.9	48.7	40.0	29.4	72.6	41.8	59.2	71.8	62.8
	Text	**98.0**	87.6	92.5	57.8	**92.3**	71.1	36.7	62.4	46.2	30.2	**82.3**	44.2	55.7	**81.2**	63.5
PatternRM-Think	Text	96.8	88.2	92.1	**89.5**	78.1	83.4	55.0	61.5	58.1	59.0	40.5	48.0	75.1	67.1	70.2
PatternRM-NoThink	Text	97.6	**89.2**	**93.3**	88.0	80.2	83.9	**60.0**	51.0	55.1	58.0	48.5	**52.8**	**75.9**	67.2	**71.3**

Table 4: PatternEval results compare the base models and three verifier-based GRPO variants. Trained policies use NoThink mode. Rates are percentages, and bold values mark the best RL result within each model scale.

Model	Method	Overall Results		Pattern Failures				Task-family Trigger		
		Acc↑	Trigger↓	CoT↓	Rep↓	Con↓	PR↓	VG↓	OS↓	KR↓
Qwen3-VL-4B	Base (Think)	45.20	21.78	9.98	11.68	5.38	13.62	26.95	6.72	25.80
	Base (NoThink)	45.71	44.22	25.80	33.00	12.96	20.12	49.32	20.34	56.18
	GRPO	**47.34**	51.43	32.17	42.69	13.58	22.61	56.61	27.07	63.66
	GRPO with PatternRM	47.21	39.50	21.82	22.77	13.13	17.06	43.64	18.62	50.53
	GRPO with Length Penalty	46.12	**32.46**	**18.92**	**22.65**	**8.57**	**14.41**	**38.39**	**14.66**	**37.56**
Qwen3-VL-8B	Base (Think)	47.55	15.98	6.21	8.16	4.06	10.06	21.86	4.14	15.88
	Base (NoThink)	49.37	36.77	23.44	25.34	9.65	15.90	41.19	15.52	47.63
	GRPO	50.44	46.38	33.83	37.18	10.64	17.39	51.36	20.52	60.31
	GRPO with PatternRM	**50.57**	30.89	17.39	**14.53**	10.97	13.29	37.12	12.07	36.34
	GRPO with Length Penalty	48.64	**27.91**	**15.03**	15.49	**10.19**	**12.92**	**33.64**	**9.66**	**33.74**

5.2 Experimental Setup

Models and training. We train Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct with GRPO (Shao et al., 2024) on 44,200 multimodal prompts covering math, logic, and document understanding. All recipes use the same prompt mixture, 8 rollouts per prompt, learning rate, and update batch size. Training sources, hyperparameters, and checkpoint details appear in Appendix G and Table 10.

Training variants. All three variants use verifier-based GRPO. **GRPO** uses the verifier reward alone, judged by GPT-OSS-120B; **GRPO with PatternRM** adds the response-pattern penalties in Equation (2); and **GRPO with Length Penalty** penalizes longer positively scored answers relative to the shortest such answer in each rollout group. PatternRM denotes the reward model, while GRPO with PatternRM denotes the policy trained with its feedback. The two auxiliary rewards differ in strength and application scope, as detailed in Appendix G.

Evaluation. Trained policies are evaluated in NoThink mode. PatternEval measures accuracy and response-pattern failures, while an 11-benchmark suite with 27,180 examples measures downstream performance. The suite uses a common 8,192-token output cap and reports the unweighted mean of task scores. Rates and normalized downstream scores are expressed as percentages; the downstream mean combines task-specific metrics and is not a shared accuracy measure. Table 4 additionally includes the instruction-tuned base models in Think and NoThink modes as references. Scoring rules, answer extraction, and dataset-overlap limitations appear in Appendix G.

5.3 Main Results

Accuracy gains alone do not ensure better response behavior. Table 4 shows that verifier-only GRPO improves PatternEval accuracy over the NoThink base at both model scales, but also increases

Table 5: Downstream performance is evaluated on 11 benchmarks. Avg. and Overall denote domain and 11-task means, respectively. Bold and underlined values indicate the best and second-best scores within each model scale.

Method	Math Reasoning						Logic Reasoning			Document Understanding					Overall
	Math Vista	Math Vision	Math Verse	Dyna Math	WeMath	**Avg.**	Logic Vista	Visu Logic	**Avg.**	AI2D	Chart QA	Doc VQA	Info VQA	**Avg.**	
Qwen3-VL-4B															
GRPO	72.20	25.33	52.69	59.80	39.14	49.83	45.09	20.60	32.84	74.55	79.88	<u>73.04</u>	<u>73.75</u>	<u>75.31</u>	<u>56.01</u>
GRPO with PatternRM	**78.40**	**37.50**	**61.88**	**66.87**	**54.86**	**59.90**	**60.27**	<u>26.40</u>	**43.34**	**83.91**	<u>80.76</u>	**83.92**	**80.51**	**82.28**	**65.02**
GRPO with Length Penalty	<u>74.50</u>	<u>32.57</u>	<u>58.38</u>	<u>63.35</u>	<u>52.57</u>	<u>56.27</u>	<u>54.69</u>	**26.60**	<u>40.64</u>	<u>81.64</u>	**81.60**	37.64	51.76	63.16	55.94
Qwen3-VL-8B															
GRPO	<u>75.50</u>	32.57	<u>54.77</u>	62.59	42.10	53.51	48.66	<u>25.10</u>	36.88	79.15	80.56	90.40	<u>82.29</u>	83.10	61.24
GRPO with PatternRM	**78.50**	**41.45**	**57.31**	**66.39**	**57.71**	**60.27**	**57.37**	**25.90**	**41.63**	<u>82.77</u>	**81.52**	**92.66**	**84.41**	**85.34**	**66.00**
GRPO with Length Penalty	75.40	<u>36.18</u>	50.51	<u>65.65</u>	<u>55.33</u>	<u>56.61</u>	<u>55.13</u>	24.20	<u>39.66</u>	**82.80**	<u>81.20</u>	<u>90.49</u>	81.90	<u>84.10</u>	<u>63.53</u>

the frequency of response-pattern failures. This divergence indicates that correctness rewards alone leave undesirable response patterns insufficiently constrained, making behavioral quality a distinct post-training objective.

PatternRM improves response behavior while retaining task performance. Adding PatternRM to GRPO reduces Trigger by 11.93 and 15.49 percentage points at 4B and 8B, respectively, while PatternEval accuracy remains close to the verifier-only baseline. Relative to the NoThink base, the resulting policies achieve both higher accuracy and fewer pattern failures. On the downstream suite reported in Table 5, they outperform verifier-only GRPO on all 11 benchmarks at both scales, spanning math, logic, and document understanding. The consistency across task families suggests that behavioral feedback can complement verifier rewards without requiring a trade-off against downstream performance.

Aggregate failure suppression and task performance diverge. Length penalization achieves the lowest aggregate Trigger among the three training variants. As shown in Tables 4 and 5, PatternRM attains higher PatternEval accuracy and higher downstream means, with an advantage on most tasks at both scales. This contrast indicates that aggregate failure rate alone is insufficient for selecting a training objective; behavioral suppression must be assessed alongside task performance.

5.4 Analysis of Response Changes

Length penalization relies more heavily on shortening. Length penalization reduces median response length to less than half that of verifier-only GRPO at both scales as shown in Table 11. The length decomposition in Table 13 attributes its aggregate Trigger reduction to the shift toward shorter responses, while the within-bin component is positive. Together with lower PatternEval accuracy, this pattern suggests a risk of overcompression at the expense of answer quality. PatternRM also shortens responses, but its within-bin component favors improvement at both scales, with stronger evidence at 8B: Trigger is lower than GRPO in every bin at or above 200 characters, as shown in Figure 5. The contrast indicates that fewer failures overall need not mean better behavior within comparable length ranges. Appendix H provides the full comparison and uncertainty estimates.

Repetition shows the most consistent improvement. PatternRM lowers all marginal pattern rates at 4B and all except contradiction at 8B. Table 14 shows that, after standardizing to the GRPO response-length distribution, repetition remains lower at both scales, whereas CoT leakage improves only at 8B and contradiction and performative reasoning increase at both scales. Repetition therefore provides the clearest improvement across scales and length adjustments. These differences motivate pattern-specific reward calibration, particularly for contradiction and performative reasoning.

Answer extraction explains part of the downstream advantage. The extraction results in Table 12 show fewer empty extracted answers for policies trained with PatternRM than for either comparison variant. Rescoring the same DocVQA and InfoVQA outputs without relying on answer

tags narrows the downstream mean gaps, but PatternRM remains ahead of both alternatives at both scales. This indicates that answer formatting contributes to measured downstream scores, while the remaining advantage extends beyond the tested answer-tag convention. Together with the length-conditioned results, this supports a more robust quality–performance balance for PatternRM across the evaluated scales and tasks.

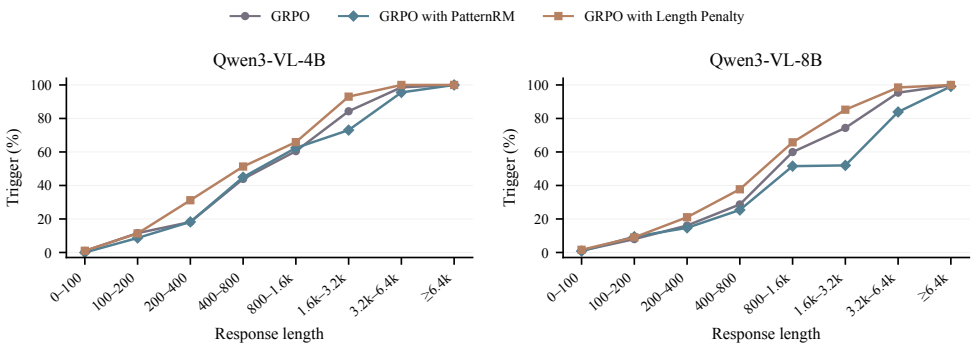


Figure 5: Curves show within-bin Trigger across response lengths for Qwen3-VL-4B and 8B on PatternEval.

5.5 Discussion

Limits of alignment during reinforcement learning. GRPO with PatternRM reduces response-pattern failures but leaves Trigger at 39.50% and 30.89% for 4B and 8B (Table 4). Earlier training data may contribute to these residual behaviors, although their origin and amenability to further RL remain untested. DeepSeek-R1 uses curated cold-start supervision and multi-stage training to address readability and language-mixing issues observed with pure RL (DeepSeek-AI, 2025). This motivates testing pattern-aware data curation and supervision during midtraining or supervised fine-tuning, followed by RL that reinforces both correctness and response quality.

Model scale and reward balance. PatternEval accuracy remains close to GRPO at both scales, while downstream scores improve. Reward weights may nevertheless need to reflect model scale and task difficulty. Recent controlled comparisons of RL-trained policies find that accuracy can peak at intermediate output lengths, with optimal length-penalty strength varying across methods and models (Nohara et al., 2026). Models relying more on extended exploration or self-correction might be more sensitive to behavioral penalties; this hypothesis requires controlled evaluation. Studies of overoptimization in RLHF and direct alignment show that stronger optimization need not improve response quality (Gao et al., 2022; Rafailov et al., 2024); matched-strength comparisons are needed to determine the appropriate balance here.

6 Conclusion

PatternEval shows that answer correctness alone does not capture the quality of final responses, with non-thinking interfaces exhibiting more response-pattern failures across the evaluated models. PatternRM turns this diagnosis into explicit behavioral feedback for verifier-based reinforcement learning, improving response behavior while largely preserving PatternEval accuracy and strengthening downstream task performance relative to verifier-only training. The comparison with length penalization highlights a distinction between suppressing visible failures through shorter responses and improving the quality of the answers delivered. Aggressive shortening can accompany reduced task accuracy, pointing to a risk of overcompression. PatternRM offers a more robust balance between behavioral quality and task performance across the evaluated settings. Together, these findings support treating response behavior as an explicit post-training objective and assessing progress jointly through answer correctness and the quality of the final response.

AI USE STATEMENT

In this work, we used multiple large language models, including **Seed-2.0-Pro** and **Kimi-K2.6**, to generate Bad Pattern data for supervised fine-tuning (SFT) to generate the PatternRM. During manuscript preparation, AI assistants were also used for language polishing and formatting corrections. The authors take full responsibility for the accuracy, originality, and integrity of the research and manuscript, including all AI-assisted content.

REFERENCES

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025. URL https://arxiv.org/abs/2505.05410.

Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 16259–16273. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.962. URL https://aclanthology.org/2024.findings-acl.962/.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. In *Conference on Language Modeling*, 2024.

Zhiyuan Fan, Weinong Wang, Xing W, and Debing Zhang. SedarEval: Automated evaluation using self-adaptive rubrics. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16916–16930, 2024. doi: 10.18653/v1/2024.findings-emnlp.984. URL https://aclanthology.org/2024.findings-emnlp.984/.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Thinkless: Llm learns when to think. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. Closing the curious case of neural text degeneration. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/34899013589ef41aea4d7b2f0ef310c1-Abstract-Conference.html.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Siyuan Gan, Jiaheng Liu, Boyan Wang, Tianpei Yang, Runqing Miao, Yuyao Zhang, Fanyu Meng, Junlan Feng, Linjian Meng, Jing Huo, and Yang Gao. Thinking-based non-thinking: Solving the reward hacking problem in training hybrid reasoning models via reinforcement learning. *arXiv preprint arXiv:2601.04805*, 2026. doi: 10.48550/arXiv.2601.04805. URL https://arxiv.org/abs/2601.04805.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*, 2022.

Deepanway Ghosal, Vernon Toh, Yew Ken Chia, and Soujanya Poria. Algopuzzlevqa: Diagnosing multimodal reasoning challenges of language models with algorithmic multimodal puzzles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9615–9632. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.486. URL https://aclanthology.org/2025.naacl-long.486/.

GLM-4.5 Team. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

Tatsuya Hiraoka and Kentaro Inui. Repetition neurons: How do language models produce repetitions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 483–495. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-short.41. URL https://aclanthology.org/2025.naacl-short.41/.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4999–5007, July 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Kembhavi_Are_You_Smarter_CVPR_2017_paper.html.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025. doi: 10.18653/v1/2025.findings-naacl.96. URL https://aclanthology.org/2025.findings-naacl.96/.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6227–6246, 2024. doi: 10.18653/v1/2024.emnlp-main.358. URL https://aclanthology.org/2024.emnlp-main.358/.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023a. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023b.

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Hoang Anh Duy Le, Shaochen Zhong, Hongyi Liu, Jiayi Yuan, Yang Sui, Vladimir Braverman, Vipin Chaudhary, and Xia Hu. Autol2s: Auto long-short reasoning for efficient large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 16836–16858, 2026. doi: 10.18653/v1/2026.findings-acl.831. URL https://aclanthology.org/2026.findings-acl.831/.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*, 2025.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*, 2025.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Viraj Bagal, Rubén Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022. URL https://openaccess.thecvf.com/content/WACV2022/html/Mathew_InfographicVQA_WACV_2022_paper.html.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. URL https://arxiv.org/abs/2503.07365.

Daisuke Nohara, Taishi Nakamura, and Rio Yokota. On the optimal reasoning length for RL-trained language models. *arXiv preprint arXiv:2602.09591*, 2026. URL https://arxiv.org/abs/2602.09591v3.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, Yifan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.983. URL https://aclanthology.org/2025.acl-long.983/.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e45caa3d5273d105b8d045e748636957-Abstract-Conference.html.

Paul Röttger, Fabio Pernisi, Bertie Vidgen, and Dirk Hovy. Safetyprompts: A systematic review of open datasets for evaluating and improving large language model safety. *arXiv preprint arXiv:2404.05399*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jiaqi Wang, Kevin Qinghong Lin, James Cheng, and Mike Zheng Shou. Think or not? selective reasoning via reinforcement learning for vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023a.

Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. Assessing judging bias in large reasoning models: An empirical study. *arXiv preprint arXiv:2504.09946*, 2025b.

Shouren Wang, Wang Yang, Xianxuan Long, Qifan Wang, Vipin Chaudhary, and Xiaotian Han. Demystifying hybrid thinking: Can llms truly switch between think and no-think? *arXiv preprint arXiv:2510.12680*, 2025c.

Shouren Wang, Wang Yang, Chuang Ma, Debargha Ganguly, Vikash Singh, Chaoda Song, Xinpeng Li, Xianxuan Long, Vipin Chaudhary, and Xiaotian Han. Path-lock expert: Separating reasoning mode in hybrid thinking via architecture-level separation. *arXiv preprint arXiv:2604.27201*, 2026.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. In *Advances in Neural Information Processing Systems*, volume 38, 2025d. URL https://papers.nips.cc/paper_files/paper/2025/hash/ac3cea0be817ebac21299b77fd114ddf-Abstract-Conference.html.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*, 2023b.

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

Hao Wen, Xinrui Wu, Yi Sun, Feifei Zhang, Liye Chen, Jie Wang, Yunxin Liu, Yunhao Liu, Ya-Qin Zhang, and Yuanchun Li. Budgetthinker: Empowering budget-aware llm reasoning with control tokens. *arXiv preprint arXiv:2508.17196*, 2025.

Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, et al. Redstar: does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*, 2025.

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal reward-bench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

Kaichen Zhang, Keming Wu, Zuhao Yang, Bo Li, Kairui Hu, Bin Wang, Xingxuan Li, and Lidong Bing. Openmmreasoner: Pushing the frontiers in multimodal reasoning with an open and general recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19276–19286, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_OpenMMReasoner_Pushing_the_Frontiers_in_Multimodal_Reasoning_with_an_Open_CVPR_2026_paper.html.

Ruiqi Zhang, Changyi Xiao, and Yixin Cao. Long or short cot? investigating instance-level switch of large reasoning models. *arXiv preprint arXiv:2506.04182*, 2025a.

Yichi Zhang, Yao Huang, Yifan Wang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, and Jun Zhu. Unveiling trust in multimodal large language models: Evaluation, analysis, and mitigation. *arXiv preprint arXiv:2508.15370*, 2025b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

APPENDIX A SUPPLEMENTARY MATERIAL

The appendices provide the benchmark composition and evaluation protocol (Appendix B), category-wise and response-length results (Appendix C), and the human-reference evaluation of candidate judges (Appendix D). PatternRM supervision is described in Appendix E, followed by the training configuration in Table 10. Appendix G details the RL protocol, length penalty, and additional diagnostics; Appendix H gives the length-bin analysis and counts. Qualitative comparisons and the complete judge prompt appear in Appendix I and Appendix J, respectively.

APPENDIX B BENCHMARK COMPOSITION AND EVALUATION DETAILS

Table 6: Example counts describe the composition of PatternEval by task family and category.

Constituent category	**Number**
VG Visual perception and grounding	
OOD perception	524
Content recognition	350
Grounding	251
Factuality	56
– *Subtotal*	*1,181*
OS OCR and structured-image understanding	
OCR	299
Chart understanding	280
– *Subtotal*	*579*
KR Multimodal knowledge reasoning	
STEM	200
Knowledge	185
Reasoning	270
– *Subtotal*	*655*
Total	**2,415**

APPENDIX B.1 RESPONSE-PATTERN DEFINITIONS

CoT leakage includes draft analysis, thinking markers, unresolved self-correction, and process narration in the final response; polished explanations are excluded. Repetition refers to unnecessary repeated sentences, semantic content, or conclusions, excluding useful emphasis. Contradiction requires incompatible claims about the same object under the same conditions without resolution. Performative reasoning refers to unsupported reasoning wrappers or evidence that does not support the conclusion.

APPENDIX B.2 LABEL ATTRIBUTION AND SCORING

Labels may co-occur when supported by distinct evidence. Under the rubric's CoT-priority rule, a passage already attributed to leaked deliberation is not counted again as contradiction or performative reasoning. The complete operational prompt appears in Appendix J, and qualitative response comparisons appear in Appendix I.

The post-training comparison uses all 2,415 responses for Trigger and the same 2,387 valid references for accuracy across methods. The remaining 28 examples lack a valid correctness reference. Per-label rates describe the incidence of each pattern; Trigger counts each response at most once.

APPENDIX B.3 MODEL AND INFERENCE CONFIGURATIONS

Table 7 lists the model-specific reference configurations for the cross-mode evaluation. Think and NoThink settings can differ in temperature and nucleus-sampling probability as well as reasoning controls. The reported comparisons therefore reflect the complete interface configurations.

Table 7: Model-specific settings define the Think and NoThink configurations. T / P denotes `temperature` / `top_p`; a dash indicates no additional setting or use of the default. Qwen3.5-series covers the survey and within-family variants.

Model	**Think**		**NoThink**
	T / P	Reasoning effort	T / P
Qwen3.5-series	0.6 / 0.95	–	0.7 / 0.8
Qwen3.7-Max	0.6 / 0.95	–	0.7 / 0.8
Qwen3.7-Plus	0.6 / 0.95	–	0.7 / 0.8
Qwen3.8-Max	0.6 / 0.95	`xhigh`	0.7 / 0.8
Kimi-K2.5	1.0 / 0.95	–	0.6 / 0.95
Kimi-K2.6	1.0 / 0.95	–	0.6 / 0.95
Kimi-K3	1.0 / 0.95	`max` (default)	–
Seed-1.6-Vision	1.0 / 0.7	–	1.0 / 0.7
Seed-2.0-Pro	1.0 / 0.95	–	1.0 / 0.95
Seed-2.1-Pro	1.0 / 0.7	–	1.0 / 0.7
Claude-Opus-4.6	–	`high` (default)	–
Claude-Opus-4.8	–	`high`	–
Mimo-v2-Omni	–	–	–
Mimo-v2.5	1.0 / 0.95	–	1.0 / 0.95
GPT-5.4	–	`high`	–
GPT-5.5	–	`medium` (default)	–
GPT-5.6-Sol	–	`medium` (default)	–

APPENDIX C ADDITIONAL RESULTS

APPENDIX C.1 CATEGORY-WISE RESULTS

Table 8 shows higher non-thinking Trigger across several task categories. Gaps are often large in OOD perception and knowledge-intensive reasoning, and smaller in content recognition and chart understanding. Scaling also affects categories differently: thinking-mode Trigger declines within several model families, while non-thinking rates remain elevated or vary non-monotonically in OOD perception, STEM, and general reasoning. The category breakdown therefore reveals variation hidden by aggregate rates, without identifying the causes of the mode differences.

Table 8: Category-wise Trigger rates (%) are compared between inference modes; lower values indicate fewer response-pattern failures.

Model	Mode	VG				OS		KR		
		OOD↓	Rec.↓	Grd.↓	Fact.↓	OCR↓	Chart↓	STEM↓	Know.↓	Reas.↓
Qwen3.5-0.8B	Non-thk	98.10	81.14	63.20	69.09	58.33	41.79	94.50	79.46	78.15
	Thk	86.86	74.86	52.00	69.09	58.33	48.57	91.50	74.05	55.19
Qwen3.5-4B	Non-thk	83.05	28.86	21.20	14.55	17.00	21.79	77.50	43.78	64.81
	Thk	44.38	16.00	4.80	0.00	6.33	10.71	25.50	25.41	32.59
Qwen3.5-9B	Non-thk	72.76	12.86	14.00	21.82	16.67	10.00	59.50	32.43	52.96
	Thk	27.81	11.14	4.80	5.45	2.67	6.07	21.50	13.04	24.44
Qwen3.5-27B	Non-thk	54.29	6.86	18.80	16.36	10.00	4.64	65.50	16.22	51.48
	Thk	9.14	2.29	2.80	3.64	1.67	0.72	8.50	1.63	13.38
Qwen3.5-35B-A3B	Non-thk	65.14	6.86	20.00	16.36	13.67	11.43	60.00	17.84	54.07
	Thk	15.25	2.29	2.00	1.82	1.67	1.43	8.50	3.78	10.23
Qwen3.5-122B-A10B	Non-thk	55.62	5.43	16.40	21.82	10.00	7.14	58.50	15.14	47.41
	Thk	9.63	1.43	2.00	1.82	0.67	1.07	11.50	4.35	16.67
Qwen3.5-397B-A17B	Non-thk	58.67	2.00	15.60	10.91	11.67	7.14	61.50	9.19	56.67
	Thk	8.57	1.43	2.40	0.00	1.00	0.36	4.00	2.70	10.78
Qwen3.6-27B	Non-thk	65.14	8.31	16.00	14.55	13.33	7.86	63.00	16.76	57.04
	Thk	12.26	2.00	2.80	3.64	1.67	1.43	8.00	3.24	11.85
Qwen3.6-35B-A3B	Non-thk	69.52	5.43	17.20	20.00	12.00	8.57	59.00	14.05	55.56
	Thk	10.56	2.29	3.20	1.82	0.67	1.79	10.50	5.41	12.12
Qwen3.7-Max	Non-thk	50.29	6.59	18.80	29.09	13.33	13.19	29.50	8.15	53.33
	Thk	9.94	1.43	7.20	3.64	3.00	1.43	6.50	4.35	8.52
Qwen3.7-Plus	Non-thk	52.19	3.71	15.60	34.55	14.33	13.62	31.50	15.22	57.25
	Thk	14.48	0.86	3.60	3.64	1.33	2.15	10.50	3.26	11.90
Qwen3.8-Max	Non-thk	84.19	2.00	8.80	3.64	6.33	3.93	55.50	7.07	60.74
	Thk	0.38	0.29	0.00	0.00	1.00	0.36	2.00	0.54	1.11
Qwen3-VL-30B-A3B	Non-thk	69.20	12.29	31.60	52.73	30.33	18.57	76.65	25.95	60.67
	Thk	30.19	4.00	1.60	5.45	4.67	1.07	9.00	10.81	8.89
Qwen3-VL-32B	Non-thk	70.75	10.57	24.40	29.09	15.00	10.36	59.00	16.85	51.11
	Thk	12.83	2.86	0.00	7.27	0.67	2.14	3.89	8.11	4.85
Kimi-K2.5	Non-thk	82.86	2.86	23.20	34.55	35.67	11.79	82.00	27.03	72.96
	Thk	11.89	1.72	4.00	1.85	4.00	1.43	4.10	1.63	15.67
Kimi-K2.6	Non-thk	83.05	3.15	30.80	74.55	37.67	18.57	85.50	32.43	87.78
	Thk	2.44	0.29	1.20	1.82	0.67	0.00	0.00	1.08	1.47
Kimi-K3	Non-thk	13.14	1.14	0.00	10.91	2.33	1.43	11.00	2.72	12.64
	Thk	0.22	0.00	0.00	0.00	1.33	0.71	0.50	0.00	0.38
Seed-1.6-Vision	Non-thk	76.57	6.63	20.00	21.82	11.67	9.29	33.50	15.76	25.79
	Thk	21.33	6.34	10.80	9.09	2.67	3.57	7.50	7.07	3.97
Seed-2.0-Pro	Non-thk	10.86	1.15	8.80	5.45	2.67	4.64	10.50	8.15	9.13
	Thk	3.62	1.73	8.40	0.00	1.67	0.00	5.50	1.63	3.97
Seed-2.1-Pro	Non-thk	19.62	3.17	13.60	5.45	2.67	4.29	19.50	10.33	17.06
	Thk	1.39	0.29	5.65	1.89	3.33	1.07	4.52	1.66	1.80
Claude-Opus-4.6	Non-thk	75.92	6.12	15.32	12.73	7.41	3.21	59.80	11.24	62.75
	Thk	8.62	1.17	2.82	12.73	1.35	1.07	0.56	1.70	8.26
Claude-Opus-4.8	Non-thk	57.61	2.04	10.48	10.91	7.41	1.79	30.65	3.37	36.44
	Thk	7.74	0.87	2.42	5.45	0.67	2.50	1.52	0.56	8.91
Mimo-v2-Omni	Non-thk	39.85	15.19	16.06	34.55	9.67	19.06	24.00	9.19	30.86
	Thk	24.04	2.22	8.29	9.09	3.33	1.98	7.00	4.32	14.87
Mimo-v2.5	Non-thk	44.38	20.63	30.45	40.00	3.67	14.03	31.50	11.89	23.42
	Thk	30.10	3.15	3.79	3.64	2.67	1.87	3.50	3.91	20.31
GPT-5-Nano	Non-thk	77.90	12.64	10.00	7.27	30.77	14.64	55.00	25.41	32.71
	Thk	42.29	0.86	2.00	5.45	2.01	2.86	7.00	5.95	11.15
GPT-5.4	Non-thk	9.71	3.43	0.80	3.64	1.67	0.71	12.00	6.49	7.04
	Thk	4.95	3.71	0.80	1.82	1.33	0.36	3.00	1.08	1.85
GPT-5.5	Non-thk	14.86	1.74	0.40	1.82	1.34	1.07	4.50	3.83	4.49
	Thk	1.61	1.16	0.40	1.82	1.01	0.00	0.51	0.55	0.00
GPT-5.6-Sol	Non-thk	14.67	0.86	0.40	0.00	1.00	0.36	4.00	3.24	4.07
	Thk	4.00	0.29	0.00	0.00	0.33	0.00	0.50	0.00	0.74

APPENDIX C.2 MODEL-LEVEL RESPONSE LENGTH AND PATTERN FAILURES

Figure 6 complements the response-length sextiles in Figure 4 by comparing model-level averages in the original survey cohort. Correct and incorrect responses are analyzed separately in each inference mode. Longer mean responses tend to accompany higher Trigger, consistent with the length-binned analysis. These correlations describe associations across models and do not establish that shortening an individual response removes its pattern failures.

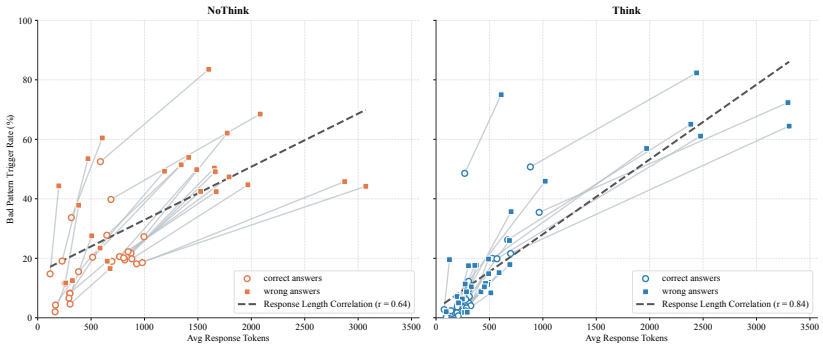


Figure 6: Points show model-level mean response length and Trigger for correct and incorrect answers. Dashed lines show mode-specific correlations.

APPENDIX D META-JUDGE EVALUATION

APPENDIX D.1 EVALUATION-SET CONSTRUCTION

We collect responses from policies at different post-training stages on examples drawn from the same source tasks as PatternEval. Seed-2.0-Pro, Kimi-K2.6, and Qwen3.5-397B independently annotate each response using the four-label rubric. Complete agreement defines the consensus stratum; disagreement defines the disputed stratum. We sample the two strata at approximately 7:3 and control the predicted positive rate near 70%, yielding 2,500 fixed responses. Human annotators then assign reference labels under the same rubric. All candidate judges are evaluated on this common response set, allowing their predictions to be compared against the same human references.

APPENDIX D.2 EVALUATION PROTOCOL AND RESULTS

Table 9 reports precision, recall, and F1 for all candidate judges and input configurations. Macro-averages weight the four patterns equally under the same rubric.

CoT leakage F1 ranges from 90.2% to 95.3%, and repetition F1 ranges from 81.6% to 88.2%. Contradiction and performative reasoning are more difficult, with F1 ranges of 56.1–62.5% and 41.8–64.5%, respectively. Image access improves the macro-F1 of Seed-2.0-Pro and Kimi-K2.6. GPT-5.5 has the highest macro-F1 overall; Seed-2.0-Pro with image access has the highest contradiction F1 and serves as the operational pattern judge.

Table 9: Precision, recall, and F1 (%) are reported for each meta-judge configuration. Overall scores are macro-averages across four labels. Bold and underlined values indicate the best and second-best scores, respectively.

Judge	Input	CoT			Rep			Con			PR			Overall		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT-5.5	Image+text	98.5	**92.4**	**95.3**	83.5	92.9	87.9	46.3	85.5	60.1	54.4	**79.3**	**64.5**	70.7	**87.5**	77.0
	Text only	98.4	92.2	95.2	83.2	**93.8**	**88.2**	47.5	**88.9**	61.9	55.2	75.0	63.6	71.1	**87.5**	**77.2**
Seed-2.0-Pro	Image+text	98.6	84.8	91.2	83.5	91.5	87.3	53.7	74.8	**62.5**	50.4	74.1	60.0	71.6	81.3	75.3
	Text only	98.6	84.7	91.1	83.4	91.6	87.3	55.6	67.5	61.0	44.8	68.1	54.0	70.6	78.0	73.4
Hy3	Text only	**98.7**	83.0	90.2	84.2	80.7	82.4	48.8	69.8	57.4	49.5	57.1	53.0	70.3	72.7	70.8
Kimi-K2.6	Image+text	97.7	89.1	93.2	**87.1**	76.8	81.6	**55.8**	70.1	62.1	58.4	40.7	48.0	**74.8**	69.2	71.2
	Text only	97.7	87.7	92.4	86.5	79.2	82.7	49.4	65.0	56.1	**64.1**	36.0	46.1	74.4	67.0	69.3
Qwen3.5-397B	Image+text	97.6	87.3	92.2	86.4	84.6	85.5	44.7	75.2	56.1	36.2	61.6	45.6	66.2	77.2	69.9
	Text only	97.5	88.5	92.8	83.2	88.6	85.8	52.0	77.8	62.3	55.6	33.5	41.8	72.1	72.1	70.7

APPENDIX E PATTERNRM SUPERVISION CORPUS

APPENDIX E.1 SUPERVISION CORPUS

PatternRM is initialized from Qwen3.5-27B. Its initial corpus draws 15K examples from each of four completed rollout collections. Within each collection, trajectories are partitioned into three step-count buckets and sampled at a 2:3:5 ratio with question-level deduplication. Quality filtering excludes the NED-only subset, retains half of the overlong cases, removes responses shorter than five characters, and removes entirely Chinese responses to English prompts. New responses to the filtered prompts are then generated by Qwen3-VL-32B-Instruct, Qwen3.5-4B-Think, and Kimi-K2.6-NoThink, contributing 10K examples per policy. The resulting pool contains 90K responses.

Kimi-K2.6, Seed-2.0-Pro, and Qwen3.5-397B independently assign the four PatternEval labels, and only unanimous four-label vectors are retained. For the retained targets, thinking-format examples use Kimi-K2.6's thinking content as the rationale field, whereas non-thinking-format examples use the reasoning in its answer field. Consensus filtering yields 52,344 unique examples. Oversampling 5,234 instances produces 57,578 SFT instances in total, comprising 34,547 thinking-format and 23,031 non-thinking-format targets. The incomplete A20B PatternRM-fusion trajectory collection is excluded from these counts.

APPENDIX F TRAINING CONFIGURATION

Table 10 records the reference configuration for GRPO with PatternRM. The evaluated checkpoints and settings for the main comparison are described in Appendix G.

Table 10: The table lists the reference training configuration for GRPO with PatternRM.

Setting	Value
Optimization algorithm	GRPO
Optimizer	Adam
Learning rate	1×10^{-6}
Clip ratio	0.2
KL coefficient	0
Entropy coefficient	0
Training batch size	128
PPO mini-batch size	128
Micro-batch size per GPU	1
Max prompt length	2,048
Max response length	16,384
Rollouts per prompt	8
Temperature	1.0
Top-p	1.0
Training epochs	5
Training steps	600
Actor parallelism	$\mathrm{TP} = 2, \mathrm{PP} = 1$
Rollout parallelism	$\mathrm{TP} = 2$
Random seed	42

APPENDIX G RL PROTOCOL AND ADDITIONAL DIAGNOSTICS

Scope and denominators. The September 18 evaluation compares GRPO, GRPO with PatternRM, and GRPO with Length Penalty on Qwen3-VL-4B and 8B. All variants are evaluated at step 510 for 4B and step 450 for 8B. Pattern rates use 2,415 responses per variant, while accuracy uses 2,387 after excluding the same 28 examples without valid reference answers. Trigger is the percentage of responses with at least one failure; individual labels can overlap. Each configuration has a single training run, so small accuracy differences should not be interpreted as evidence of equivalence. Appendix H describes the exploratory within-bin intervals.

Training data and configuration. The 44,200-prompt mixture draws on OpenMMReasoner (Zhang et al., 2026) and supplementary resources: We-Math, MM-Eureka, and ThinkLite for visual reasoning (Qiao et al., 2025; Meng et al., 2025; Wang et al., 2025d); PuzzleVQA and AlgoPuzzleVQA for abstract and algorithmic puzzles (Chia et al., 2024; Ghosal et al., 2025); and textbook, chart, and infographic question answering (Kembhavi et al., 2017; Masry et al., 2022; Mathew et al., 2022). The runs use eight rollouts per prompt, a constant learning rate of 10^{-6}, training batch size 128, and PPO mini-batch size 128. The 4B run using PatternRM resumes from an earlier checkpoint. PatternEval response lengths are measured in characters; downstream response lengths are measured in tokens. Additional training parameters appear in Table 10.

Dataset separation. The training mixture and downstream evaluation share several dataset families. The available records do not establish item-level separation between RL training and evaluation, or between PatternRM supervision and the judge-evaluation and benchmark sets.

Suite scoring and decoding. The downstream suite contains 11 tasks and 27,180 questions. MathVista, MathVision, MathVerse, LogicVista, and AI2D use `acc_score`; DynaMath uses its average score; WeMath uses strict accuracy; VisuLogic uses its task accuracy; ChartQA uses relaxed overall accuracy; DocVQA and InfoVQA use ANLS. Avg is an unweighted task mean, not pooled accuracy. All six evaluations use the vLLM offline backend, tensor parallelism 8, batch size 128, maximum model length 32,768, maximum new tokens 8,192, and the full task splits. Suite token means and PatternEval character lengths measure different evaluation sets and units.

Extraction sensitivity. Standard answer extraction first looks for an `<answer>` block, then a boxed answer, and finally a trailing numeric answer; failure to extract an answer returns an empty string and a score of zero. The fallback mean replaces only the DocVQA and InfoVQA scores

Table 11: Accuracy, response-pattern rates, and response lengths are reported for all six training runs on PatternEval. Rates are percentages, and Mean and Median summarize response length in characters.

Model	Method	Acc	Trigger	CoT	Rep	PR	Con	Clean	Mean	Median
Qwen3-VL-4B	GRPO	47.34	51.43	32.17	42.69	22.61	13.58	48.57	3903	623
	GRPO with PatternRM	47.21	39.50	21.82	22.77	17.06	13.13	60.50	1319	455
	GRPO with Length Penalty	46.12	32.46	18.92	22.65	14.41	8.57	67.54	1516	214
Qwen3-VL-8B	GRPO	50.44	46.38	33.83	37.18	17.39	10.64	53.62	3708	653
	GRPO with PatternRM	50.57	30.89	17.39	14.53	13.29	10.97	69.11	1360	479
	GRPO with Length Penalty	48.64	27.91	15.03	15.49	12.92	10.19	72.09	814	259

Table 12: Answer-format and extraction diagnostics are reported on a 0–100 scale. Raw and Fallback denote document ANLS before and after rescoring, respectively. Overall is the 11-task mean after fallback rescoring.

		Tag↑	Empty↓	DocVQA		InfoVQA		Overall
Model	Method			Raw	Fallback	Raw	Fallback	
Qwen3-VL-4B	GRPO	73.18	13.14	73.04	90.83	73.75	80.99	58.28
	GRPO with PatternRM	97.57	1.45	83.92	92.68	80.51	84.17	66.15
	GRPO with Length Penalty	81.40	10.71	37.64	83.86	51.76	81.42	62.83
Qwen3-VL-8B	GRPO	80.77	7.17	90.40	93.56	82.29	84.97	61.77
	GRPO with PatternRM	95.82	1.11	92.66	94.14	84.41	85.93	66.27
	GRPO with Length Penalty	94.82	1.88	90.49	93.82	81.90	84.04	64.02

with tag-independent ANLS rescoring of the same responses. This isolates sensitivity to answer extraction without generating new outputs. The complete fallback algorithm is not documented in the available records, which limits reproducibility of this check. Tag and Empty report the mean answer-tag rate and extraction-empty rate over ten tasks, excluding VisuLogic, which requests a bare option letter. DocVQA tag rates for GRPO/GRPO with PatternRM/GRPO with Length Penalty are 71.43/86.48/20.10% at 4B and 96.26/98.26/96.30% at 8B; InfoVQA tag rates are 83.33/94.79/56.19% and 93.72/98.14/97.18%, respectively.

Reward interventions. The reward formulation assumes a binary correctness verifier. The run summary identifies a local grader and a remote fallback, but their complete score-domain implementation requires verification against the original reward logs or code. GRPO with PatternRM adds a non-positive auxiliary reward with floor -0.1 to the correctness reward and clips their sum to $[0, 1]$. The category weights are 0.05 for Rep and CoT, and 0.02 for PR and Con; the reward model is invoked with probability 0.6 per rollout.

Length-penalty baseline. GRPO with Length Penalty penalizes relatively long, positively scored answers within each rollout group. The following description is based on the code-inspection record for `compute_length_rewards_v2`, its GRPO caller, and the accompanying configuration; the original run logs were unavailable for verification.

For a prompt group, let s_i be the outcome score before the length adjustment and let $\mathcal{C} = \{i : s_i > 0\}$ denote the positively scored rollouts. If $\mathcal{C}$ is empty, no penalty is applied. Otherwise, $\ell_{\min}$ and $\ell_{\max}$ are the minimum and maximum answer-block lengths among these rollouts. For an eligible topic, the penalty is

$$p_i = \begin{cases} -0.6 \min\left(\dfrac{\ell_i - \ell_{\min}}{\ell_{\min} + \epsilon},\, 0.5\right), & i \in \mathcal{C} \text{ and } \ell_{\max} - \ell_{\min} > 20, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where ϵ stabilizes the denominator. The adjustment lies in $[-0.3, 0]$; the shortest positively scored answer receives no penalty. The 20-token dead zone suppresses adjustments when correct-answer lengths are similar.

Table 13: Trigger differences from GRPO are decomposed into length-composition and within-bin components in percentage points. Std. denotes length-standardized Trigger (%).

Model	Method	Raw Δ	Composition	Within-bin	Std.
Qwen3-VL-4B	GRPO with PatternRM	-11.93	-11.25	-0.67	50.92
	GRPO with Length Penalty	-18.96	-23.61	+4.65	56.43
Qwen3-VL-8B	GRPO with PatternRM	-15.49	-11.30	-4.18	42.64
	GRPO with Length Penalty	-18.47	-22.88	+4.41	51.06

Answer-block extraction and application. If the opening and closing answer delimiters each occur exactly once, ℓ_i counts tokens between them; otherwise, it counts the entire response. Model-specific single-token placeholders represent these delimiters when available, with literal `<answer>` and `</answer>` tags as the fallback. There is no separate extraction branch for text following `</think>`. The adjustment is added to the scalar outcome score used for advantage computation, $s'_i = s_i + p_i$; the token-level reward tensor is not modified.

Scope and strength. The mechanism is enabled for every group, with eligibility determined by each sample's reward-function topic. Eligible topics are math, inference, and knowledge; document understanding is excluded. In contrast, PatternRM covers all topics and is invoked with probability 0.6 per rollout, with an auxiliary penalty capped at 0.1. The reported mean adjustments are -0.0547 and -0.0482 for GRPO with Length Penalty at 4B and 8B, versus -0.0071 and -0.0062 for GRPO with PatternRM. These comparisons therefore differ in topic coverage, invocation frequency, and penalty magnitude, as well as the feedback signal.

Interpretation. Answer-block formatting affects the penalty through the length rule above and can separately affect answer extraction during evaluation. For 4B, GRPO with Length Penalty has a DocVQA answer-tag rate of 20.10%, versus 71.43% for GRPO, and an ANLS score of 37.64 versus 73.04. The extraction check in Table 12 examines this sensitivity. It does not identify the cause of document-task degradation or separate reward targeting from penalty strength.

Length standardization and decomposition. Responses are grouped by character length using boundaries at 0, 100, 200, 400, 800, 1,600, 3,200, 6,400, and ∞. Every variant has at least 43 responses in each bin, so no empty-bin extrapolation is needed. Let w_a^k be the proportion in bin k and r_a^k its Trigger rate for arm a, with b denoting the same-scale GRPO. Standardization uses $\sum_k w_b^k r_a^k$. The symmetric Kitagawa identity is

$$R_a - R_b = \underbrace{\sum_k (w_a^k - w_b^k)\frac{r_a^k + r_b^k}{2}}_{\text{composition component}} + \underbrace{\sum_k (r_a^k - r_b^k)\frac{w_a^k + w_b^k}{2}}_{\text{within-bin component}}. \tag{4}$$

This identity separates differences in the length distribution from differences in failure rates within bins. Because training changes response length, the bins can contain different prompts and levels of task difficulty across variants. Neither the within-bin component nor the comparison of category weights identifies a causal effect. Values are transcribed at reported precision, so sums can differ by 0.01 percentage point. In particular, displayed 4B GRPO with Length Penalty Trigger values imply -18.97 points, while the reported decomposition uses -18.96.

The standardized GRPO with PatternRM rates improve for Rep at both scales and CoT at 8B, but increase for CoT at 4B and for PR/Con at both scales. With a pooled reference length distribution, the report gives Trigger rates of 39.98/39.24/45.09% for GRPO/GRPO with PatternRM/GRPO with Length Penalty at 4B and 35.22/31.61/40.01% at 8B. The ordering is unchanged under this alternative reference distribution, subject to the same limitation on causal interpretation.

Table 14: Label rates (%) are standardized to the GRPO response-length distribution at the same model scale.

	4B			8B		
Label	GRPO	GRPO with PatternRM	GRPO with Length Penalty	GRPO	GRPO with PatternRM	GRPO with Length Penalty
Trigger	51.43	50.92	56.43	46.38	42.64	51.06
Rep	42.69	35.69	45.78	37.18	27.81	38.10
CoT	32.17	33.51	39.55	33.83	29.86	35.93
PR	22.61	27.18	23.53	17.39	21.90	24.62
Con	13.58	18.17	14.17	10.64	17.46	18.75

APPENDIX H RESPONSE PATTERNS WITHIN LENGTH BINS

This analysis compares failure rates within fixed response-length ranges to distinguish aggregate improvements from within-bin changes. The boundaries are 0, 100, 200, 400, 800, 1,600, 3,200, 6,400, and infinity, in characters. Intervals are left-closed and right-open. Each variant has 2,415 responses, with at least 43 in every bin. The longest bin includes responses of exactly 6,400 characters. The figures are computed directly from the integer counts in Table 15.

Interval construction. For k labeled responses among n in a bin, we compute a Wilson score interval $[L, U]$ for $p = k/n$ with $z = 1.96$. For an intervention a and baseline b, the Newcombe interval for $d = p_a - p_b$ is

$$\left[d - \sqrt{(p_a - L_a)^2 + (U_b - p_b)^2},\quad d + \sqrt{(U_a - p_a)^2 + (p_b - L_b)^2}\right].$$

All variants were evaluated on the same prompts, but the available summaries do not contain joint prompt-level outcomes and lengths. The intervals therefore treat bin samples as independent and omit cross-variant dependence. The 32 comparisons across variants, scales, and bins are exploratory, with no multiplicity adjustment. Since generated length also changes bin membership, these comparisons do not identify a causal effect at fixed length.

Fixed-bin results. GRPO with PatternRM has four interval-excluding reductions across the two scales and no interval-excluding increases; GRPO with Length Penalty has four increases and no reductions. For 8B GRPO with PatternRM, the reductions are -8.4 points in $[800, 1600)$, -22.4 in $[1600, 3200)$, and -11.5 in $[3200, 6400)$. Its point estimates are lower in all six bins from 200 characters onward. At 4B, the only interval-excluding reduction is -11.2 points in $[1600, 3200)$, while several other bins have small positive point estimates. Thus, the evidence is stronger at 8B and does not support uniform improvement in every bin at both scales.

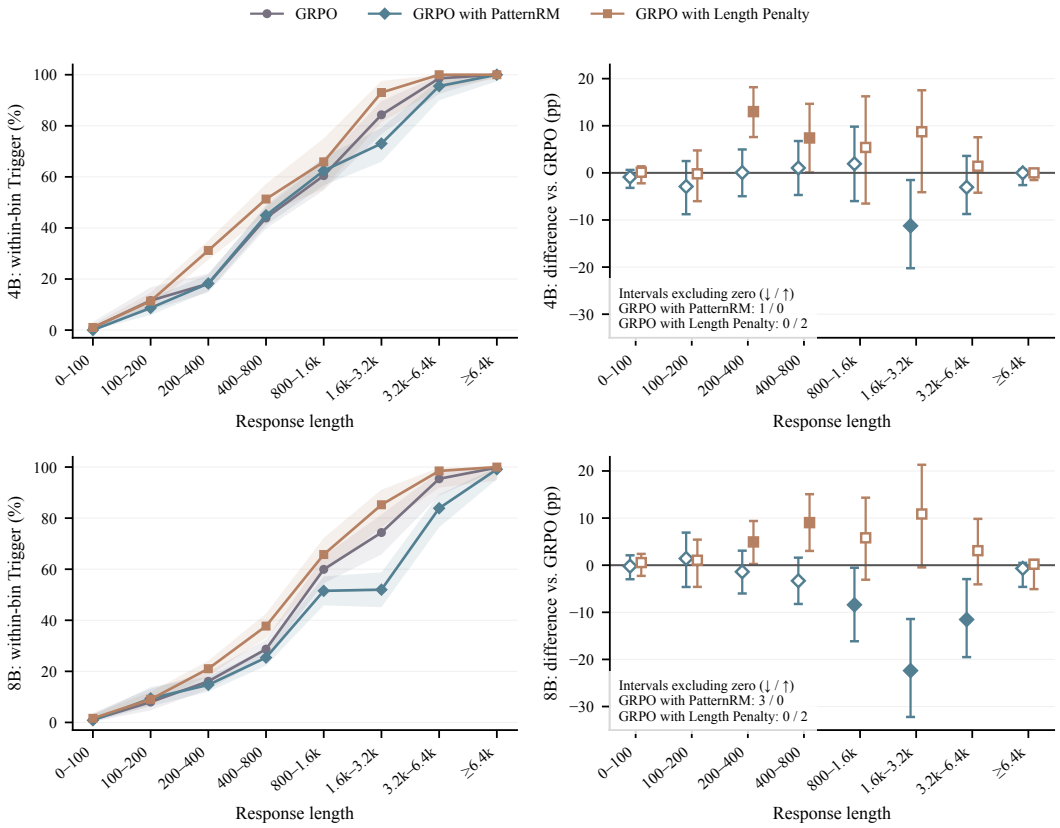


Figure 7: The top and bottom rows show within-bin response-pattern failures for 4B at step 510 and 8B at step 450, respectively. Left panels show Trigger with Wilson 95% intervals; right panels show differences from GRPO with Newcombe independent-proportion 95% intervals. Filled markers indicate intervals excluding zero, and negative differences favor the intervention. These pointwise, unadjusted intervals provide exploratory approximations rather than paired-prompt or across-seed uncertainty estimates.

Table 15: Each entry gives the number of triggered responses followed by the total number of responses in the length bin. PatternRM and Length Penalty denote the corresponding auxiliary rewards added to GRPO.

	4B			8B		
Characters	GRPO	PatternRM	Length Penalty	GRPO	PatternRM	Length Penalty
$[0, 100)$	2/225	0/276	7/683	2/190	2/237	7/435
$[100, 200)$	23/199	24/277	51/449	13/163	22/234	43/476
$[200, 400)$	76/418	96/526	179/574	71/441	81/551	150/713
$[400, 800)$	234/533	280/623	137/267	171/596	167/658	146/387
$[800, 1600)$	173/286	181/290	56/85	190/317	152/295	119/181
$[1600, 3200)$	102/121	122/167	40/43	87/117	105/202	75/88
$[3200, 6400)$	70/71	107/112	66/66	83/87	104/124	65/66
$[6400, \infty)$	562/562	144/144	248/248	503/504	113/114	69/69
Total	1242/2415	954/2415	784/2415	1120/2415	746/2415	674/2415

APPENDIX I QUALITATIVE RESPONSE COMPARISONS

Each case compares responses to the same PatternEval prompt, alongside its input images. The excerpts are translated from Chinese, with omissions marked by ellipses and failure evidence highlighted in dark red. Answer correctness and pattern labels follow the evaluation record. Each pair contrasts a focal failure with a response from GRPO with PatternRM in which no pattern was detected. The cases illustrate how the rubric distinguishes response behaviors; they are not a representative sample for estimating failure prevalence or causal training effects.

APPENDIX I.1 COT LEAKAGE DESPITE A CORRECT FINAL ANSWER

Case 1: OCR and chart understanding
Prompt. During 2010–2020, in which two consecutive years did China's government education expenditure have the same share of GDP?
Reference answer. 2012 and 2013.



Pattern detected

Qwen3-VL-4B
GRPO with Length Penalty
Answer correctness: correct

Response excerpt
"2013 and 2014, both 4.13%? **No**, 2013 is 4.13% and 2014 is 4.11%; they are not the same.

Looking more carefully:
2012: 4.13%
2013: 4.13% → same as 2012
2014: 4.11% → different

. . . Correct answer: 2012 and 2013."

Failure location. The opening correction ("No) and subsequent reconsideration ("Looking more carefully) expose intermediate deliberation in the final response.

No pattern detected

Qwen3-VL-4B
GRPO with PatternRM
Answer correctness: correct

Response excerpt
"According to the chart . . ., the values for each year are:
. . .
2012: 4.13%
2013: 4.13% ← same as 2012
2014: 4.11%
. . .
The values for 2012 and 2013 are both 4.13%.

Answer: 2012 and 2013."

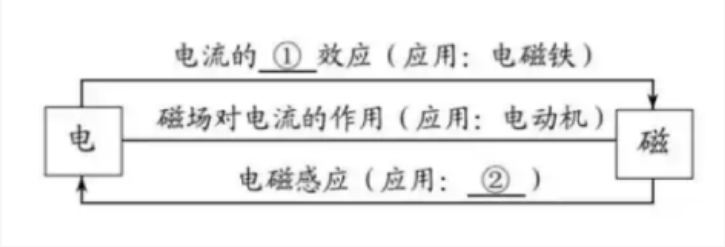
Recorded judgment. The data listing provides a finished explanation without exposing an internal correction process.

Diagnostic insight. Both answers are correct. The distinction concerns exposed deliberation, rather than explanation length: the source reports 239 characters for GRPO with Length Penalty and 298 for GRPO with PatternRM.

Figure 8: Two correct answers differ in whether they expose intermediate self-correction. Highlighted text identifies the leaked deliberation in the translated excerpt.

APPENDIX I.2 REPEATED ANSWER BLOCKS AND ACCEPTABLE SUMMARIES

Case 2: School science
Prompt. Fill in the two missing entries in a concept map about electricity and magnetism.
Reference answer. (1) Magnetism; (2) generator.



Pattern detected

Qwen3-VL-4B
GRPO
Answer correctness: correct

Response excerpt
". . . The magnetic effect of electric current . . . is used in electromagnets.
. . . Electromagnetic induction . . . is used in generators.

Therefore:
(1) Magnetism
(2) Generator

Answer:
(1) Magnetism; (2) generator."

Failure location. The consecutive closing blocks headed "Therefore and "Answer repeat the same two entries without adding information.

No pattern detected

Qwen3-VL-4B
GRPO with PatternRM
Answer correctness: correct

Response excerpt
"(1) The effect of electric current . . . is its magnetic effect . . ., so the entry is magnetism.

(2) Electromagnetic induction . . . is used in a generator . . ., so the entry is generator.

Answer:
(1) Magnetism
(2) Generator."

Recorded judgment. A single final summary after the explanation is treated as a normal answer structure.

Diagnostic insight. Both responses explain the concepts and restate the answer. The recorded distinction is between two consecutive answer blocks and one closing summary, illustrating a rubric-sensitive boundary rather than a prohibition on all restatement.

Figure 9: Two correct science answers illustrate the recorded distinction between redundant answer blocks and an acceptable closing summary. Highlighted text marks the duplicated blocks.

APPENDIX I.3 AN INTERNAL CONTRADICTION WITH THE SAME FINAL ORDERING

Case 3: Visual action ordering
Prompt. Analyze seven cocktail-making photographs labeled a–g and return their chronological ordering.
Reference answer. No reference is provided; answer correctness is not assessed in this case.



Pattern detected

Qwen3-VL-4B
GRPO
Answer correctness: not assessed

Response excerpt
"... Image c uses a measuring cup ... and should precede pouring directly into the cup.

... Image c is more precise and **usually comes after image b**, because a measuring cup gives a precise quantity.

... So the order should be:
e → **c** → **b** → a → d → f → g.

Chronological order:
`ecbadfg`"

Failure location. The explanation places c after b, whereas the concluding sequence places c before b.

No pattern detected

Qwen3-VL-4B
GRPO with PatternRM
Answer correctness: not assessed

Response excerpt
"... Image c measures the liquid precisely and should occur before the liquid is poured into the mixing cup.

... First, prepare the ingredients (e). Then measure and pour the liquid (c). Next, add more liquid (b), possibly another spirit or syrup.

... The reasonable order is:
e → c → b → a → d → f → g.

Chronological order:
`ecbadfg`"

Recorded judgment. The explanation and final sequence give consistent ordering relations.

Diagnostic insight. The final answer strings are identical. Internal consistency separates the responses even when final-answer agreement cannot reveal the contradiction; this comparison does not establish which ordering is correct.

Figure 10: Responses with the same final ordering differ in internal consistency. Highlighted spans identify the conflicting relations between b and c.

APPENDIX I.4 PATTERN JUDGMENTS AND ANSWER CORRECTNESS CAN DIVERGE

Case 4: Historical knowledge
Prompt. Which dynasty does the pictured artwork come from?
Reference answer. The Ming dynasty.



Pattern detected

Qwen3-VL-8B-Instruct
Original instruction-tuned model
Answer correctness: correct

Response excerpt
"This painting comes from the Ming dynasty.

. . . The figures, clothing, outlines, and coloring **all fit the characteristics of Ming court or literati painting**.

. . . Although the writing is somewhat blurred, its layout, script . . ., and content **all fit Ming literati painting conventions**.

. . . In conclusion, this painting should be a Ming work."

Failure location. The style and inscription justifications both assert that features "all fit the predicted dynasty without identifying specific visible evidence.

No pattern detected

Qwen3-VL-8B
GRPO with PatternRM
Answer correctness: incorrect

Response excerpt
"This painting comes from the Qing dynasty.

. . . The vertical text on the left quotes the Analects: 'In Chen their provisions ran out; the followers fell ill and could not rise. . . .'

. . . This juxtaposition of a historical story and canonical text is a common arrangement in Qing illustrated editions of the Analects or biographies of sages.

. . . Conclusion: this is a Qing painting."

Recorded judgment. The answer supplies concrete supporting dimensions, including textual content and the image–text arrangement, rather than only restating the conclusion.

Diagnostic insight. The response with no recorded pattern failure gives the wrong dynasty. The pattern labels characterize response behavior; they do not validate the cited inscription or historical claims and cannot replace answer-correctness evaluation. This pair uses an instruction-tuned reference rather than GRPO.

Figure 11: The recorded performative-reasoning judgment differs from answer correctness in this pair. Highlighted text identifies generic justifications in the correct answer, while the response with no detected pattern predicts the wrong dynasty.

APPENDIX J PATTERNEVAL JUDGE PROMPT

The pattern judge receives the question, image, and response and assigns the four labels defined in Section 3.1. Images are supplied separately as multimodal inputs, not serialized into the text placeholders below. The English translation precedes the original Chinese prompt. Only typesetting has changed: the decision rules, priority policy, examples, output order, and placeholders are preserved. Reproducing the full evaluation also requires the API serialization, image preprocessing, and model-version metadata in the release manifest.

APPENDIX J.1 ENGLISH VERSION

Response-Pattern Meta-Judge Prompt (English Translation)

You are a strict analyzer of undesirable patterns in model responses. Given the user question, any relevant image, and the model response, determine whether the response exhibits each of the following four undesirable patterns. For every pattern, output a binary decision (1 or 0) and a one-sentence justification.

I. Decision Principles

1. Base the judgment only on the question, any relevant image, and the model response. Every decision must be grounded in observable text in the model response.
2. Use the image to verify image-related descriptions, counts, localization, recognition, and spatial relations in the response. However, do not mark a pattern merely because the answer is incorrect.
3. Factual errors, visual-recognition errors, calculation errors, and incomplete answers are not undesirable response patterns by themselves. Mark 1 only when the error manifests as one of the named patterns below.
4. Judge the four labels independently; multiple labels may be positive. Mark 1 only when there is explicit textual evidence; otherwise mark 0.
5. Length, many steps, or detailed explanation alone do not constitute an undesirable pattern. Do not mark 1 merely because a response is long or contains extensive reasoning.

II. Decision Order (Important)

Apply the following fixed order:

1. First judge **chain-of-thought leakage**: read the full response and check whether it exposes internal deliberation that should remain hidden, including self-correction, tentative wavering, or process-oriented narration that lays out the solution process from the beginning, according to Category 1.
2. Next judge **response repetition** independently of whether leakage is present, according to Category 2.
3. Finally judge **logical contradiction** and **performative reasoning**. First determine whether content that appears contradictory or superficially reasoned is actually part of exposed internal deliberation, such as self-correction, repeated trial, process-level wavering, or draft-like reasoning. If so, attribute that content to chain-of-thought leakage and do not count it again as contradiction or performative reasoning; mark those categories 0 for that content. An exception applies when the final delivered answer contains a separate hard contradiction or an independently unsupported shell of reasoning. Logical contradiction concerns only unresolved, mutually incompatible claims retained in the final delivered content. Performative reasoning concerns analysis that adds no valid support for the conclusion, or staged analysis on a low-reasoning task, when chain-of-thought leakage is not the better explanation.

III. Four Undesirable Patterns

1. Chain-of-Thought Leakage

Definition. The model directly exposes intermediate reasoning, internal self-talk, tentative analysis, task decomposition, self-correction, or draft-like thinking that should remain hidden, thereby reducing the naturalness of the response or violating the expected output style.

Mark 1 only when there is explicit evidence of exposed internal reasoning in at least one of the following forms:

- **First-person process narration:** phrases such as “let me look first,” “I will analyze this first,” “let me think,” “let me check,” “I need to judge this,” or “I need to confirm this first”; or reader-guiding phrases such as “let us analyze these one by one,” “let us analyze step by step,” “let us first look at,” and similar uses of “let me,” “let us,” or “we first.” These expressions narrate how the model is looking or thinking and lead the reader through the model’s working process. Once such process narration appears, it counts as leakage even if followed by bullet points or section headings.
- **Self-correction, reversal, wavering, or repeated trial:** traces such as “no, let me look again,” “what I said earlier was wrong; it should be,” “wait, let me reconsider,” or “try A first; no, then try B.”
- **Exposed internal decisions or drafts:** phrases such as “end of thinking,” “now confirmed,” “my preliminary judgment is,” or “let me run through it mentally,” as well as internal decisions, execution strategies, task decomposition, draft reasoning, or residual markers such as `think`.

Mark 0 in the following cases, provided that none of the above traces is present:

- User-facing structured solution or analysis: sections such as "Analysis," "Step 1," and "Step 2," or objective statements such as "to determine the conclusion, first analyze . . . ," are not leakage when the response is a polished explanation directed to the user, without first-person process narration, self-correction, wavering, or exposed internal decisions. If such analysis is unsupported, classify it as performative reasoning rather than leakage.
- The user explicitly requests reasoning, steps, a draft, or a method, and the response provides a polished, user-facing derivation, tutorial, procedure, or solution.
- A direct conclusion with a brief justification or concise supporting points.
- A short elimination of alternatives, comparison, or user-facing explanation of multiple perspectives or conditional branches.

Positive example. The question asks for an object's orientation. The response says, "Let me look at this image first. It seems to face left—no, looking again, it should face right," exposing repeated tentative inspection.

Negative example. The question asks for the orientation of a piece of furniture. The response contains an "Analysis" section and objectively derives the orientation in Step 1 and Step 2 without first-person self-talk or self-correction. This is not leakage; if the derivation provides no valid image evidence, it is performative reasoning.

2. Response Repetition

Definition. The response unnecessarily repeats the same sentence, meaning, structure, or conclusion to a degree that clearly harms information efficiency, readability, or answer quality.

Mark 1 in the following cases:

- The same conclusion, recommendation, or disclaimer is repeated consecutively or frequently.
- The same information or conclusion is restated across multiple sections—for example, analysis, verification, summary, and final answer—without adding information, even if the wording changes.
- Paragraphs are highly homogeneous and merely paraphrase one another.
- The reasoning circles back, repeatedly lists the same candidates or clues, or repeatedly rejects and recomputes the same result beyond what normal emphasis requires, producing a mechanical or cyclic pattern.

Mark 0 in the following cases:

- Limited repetition used to emphasize a key point.
- Necessary local repetition in an enumeration or stepwise structure when the items differ substantively.
- A single useful self-correction or one concluding restatement that naturally connects the reasoning to the final answer without creating obvious redundancy.

Positive example. When asked for a painting's name, the response gives the title and artist in a heading, then immediately repeats the same title, artist, and year in a "Basic Information" list without adding information.

Negative example. When asked for the answer to a problem, the response provides a derivation and closes with one sentence restating the final answer.

3. Logical Contradiction

Definition. In the final delivered content, the response gives two claims about the same object, value, or conclusion that cannot both be true under the same conditions. Both claims remain presented as valid information: the response does not choose between them, invalidate either one, or distinguish their conditions. A positive decision requires locating two specific, mutually incompatible passages, A and B.

A hard contradiction must satisfy all of the following conditions: A and B concern the same object, quantity, or conclusion; they use the same conditions, time, and measurement convention; they cannot both be true; the model does not state that one has been withdrawn or is merely an alternative; and both remain active claims in the answer.

Mark 1 for any of the following four forms:

1. **Mutually exclusive final conclusions.** The same question, object, or answer position receives two incompatible final conclusions without a choice. For example, the response first identifies an artifact as a bronze *zun* but ends with "Final answer: bronze *ding*," or first states that A is correct and B violates the conditions but ultimately selects B.
2. **Incompatible attributes of the same object.** The response assigns mutually exclusive values of the same attribute to one object and retains both. For example, it calls block b horizontal in a Klotski puzzle and later instructs the user to move the vertical block b; or it says a function is odd and symmetric about the origin, then says it is symmetric about the y-axis and therefore even.
3. **Incompatible values of the same quantity.** The response retains two conflicting values for the same number, date, coordinate, count, angle, area, or temperature. For example, it uses two different base areas without a correction; states that an image contains six circles but itemizes three on the left and four on the right for a total of seven; or gives point P first as $(2, 3)$ and later as $(3, 2)$.
4. **An unretracted derivation conflicts with the final answer.** The reasoning explicitly derives and retains X, but the final answer gives incompatible Y without abandoning the earlier result. For example, the steps

identify 8 and 9 as mandatory waypoints but the final waypoint list omits them; or the derivation obtains $x = 3$, verifies it by substitution, and then gives $x = 5$ as the final answer.

Mark 0 in the following cases:

- **Corrected, abandoned, or resolved alternatives.** An earlier statement is explicitly rejected, corrected, or abandoned, or one candidate is selected from several and only one final conclusion remains. Process-level wavering of this kind belongs to chain-of-thought leakage.
- **Hedging or uncertainty.** Expressions such as "possibly," "probably," "more likely," "tends to be," or "cannot rule out" indicate alternatives or confidence differences rather than two definite incompatible conclusions. For example, "possibly late Shang, though early Western Zhou cannot be ruled out."
- **Different conditions, times, viewpoints, or conventions are distinguished.** For example, the result is X under Assumption 1 and Y under Assumption 2; or an object is on the left from the observer's viewpoint but on the right relative to its own facing direction.
- **A single wrong answer, visual error, or calculation error without an internal contradiction.** This category does not evaluate correctness. Mark 0 if the answer is wrong but does not retain two mutually incompatible claims about the same object, quantity, or conclusion.
- **Imprecision, equivalent restatement, unit conversion, repetition, rough reasoning, or performative reasoning without a locatable pair of incompatible claims.** Examples include calling an altocumulus cloud a "mackerel sky" or stating that one meter equals 100 centimeters.

Key distinction. A hard contradiction retains two propositions about the same object attribute, quantity, or conclusion that cannot both be true. Process-level wavering, hedging, visual or factual errors, distinguished conditions, and resolved alternatives do not qualify.

Positive example. When asked which position from the left contains the white-haired person wearing glasses, the response identifies the target as item 4 in its list but concludes "therefore, the fifth." It retains two incompatible answers, 4 and 5, without resolving them.

Negative example. When asked which animal appears in an image, the response consistently answers "a golden retriever" and adds supporting details such as long fur and golden coloring.

4. Performative Reasoning

Definition. The response presents the appearance of analysis, derivation, argument, or evidential support, but that material adds no valid information that bears on the final conclusion. It performs being well-supported rather than providing support.

Valid information gain means an observation, comparison, count, elimination, disambiguation, calculation, or verification that can support the conclusion. For image tasks, valid support must refer to clearly observable content, such as specific objects, text, colors, quantities, positions, relations, chart values, or paths. It may compare size, distance, occlusion, direction, shape, or state; reason from visible paths, walls, nodes, arrows, table rows or columns, and flow branches; deduplicate or match people and objects across images; or read text, legends, coordinates, labels, and values. Generic phrases such as "as can be seen from the image," "considering the overall layout," "based on visual characteristics," or "after comprehensive judgment" usually constitute performative reasoning when they identify no concrete visible content.

Mark 1 only when the response adopts an analytical, inferential, argumentative, or evidence-bearing posture and exhibits one of the following forms:

1. **Unsupported reasoning wrapper.** The response uses the form of reasoning or evidence without substantive support. Typical signs include "first," "second," "therefore," or "in summary" connecting only generic assertions; merely restating the question or conclusion; claiming to reason "from the image" without naming any concrete element, position, count, text, color, direction, occlusion, or relation; or invoking frameworks such as semantics, structure, and context without connecting them to specific evidence. For example, when asked to identify a plant, the response says that leaf shape, growth habit, and overall visual characteristics establish the species but identifies none of those visible characteristics.
2. **Evidence that does not support the conclusion.** The response offers apparently relevant analysis or observations that do not entail the conclusion, or merely packages an answer that could be read directly from the image or was already supplied by the user. Typical cases include analyzing A but concluding B; providing background that cannot establish the answer; identifying something from generic impressions such as "looks similar," "the style fits," or "the shape is close" rather than discriminative evidence; adding analysis to a low-reasoning task without any new observation, comparison, count, elimination, disambiguation, or verification; restating a user-provided route without comparing position, roads, distance, reachability, or obstacles; or calling maze nodes "key" or "mandatory" without checking wall and path connectivity or excluding alternatives. This also includes identification answers padded with biography, work lists, general stylistic history, typical roles, or other background that does not explain how the image establishes the identity through a face, inscription, signature, emblem, or other discriminative detail.

Mark 0 in the following cases:

- A direct answer or brief explanation that does not stage an analysis. Merely saying "based on the image" without naming an image element is treated as a direct answer.

- Image analysis that cites concrete visible evidence supporting the conclusion, even if brief, approximate, verbose, or imperfect. Examples include "the lower-right button reads 'Submit,' so it is the submit button" and comparing the water levels and vessel shapes of two cups. The evidence must be sufficiently discriminative, such as visible text, an inscription or signature, an exact count or position, or a comparable state. Generic terms such as "similar shape," "matching style," "technological feel," "flame-like core," or "similar colors" do not qualify as valid evidence.
- A genuine image-connectivity analysis that reaches the wrong result because of a visual error. Mark 0 only when the response actually traces whether adjacent cells connect through walls or passages and uses that connectivity to rule out alternatives. Merely naming cells or nodes as "key" or "mandatory," or listing their locations without proving connectivity or excluding alternatives, remains performative reasoning.
- The user explicitly requests justification, analysis, stepwise reasoning, or evidence, and the analysis is not entirely empty; for example, an elimination argument explaining why option C is selected.
- A visual, calculation, or factual error when the cited evidence genuinely bears on the conclusion; for example, counting real objects in the image but missing or miscounting one.
- A cautious statement that the image is unclear and the conclusion cannot be confirmed, followed by a tentative answer or refusal to assert one.

Key distinction. Direct answers, concrete image evidence, valid reasoning, and chain-of-thought leakage receive 0 for this category. Analysis that only restates the question or conclusion, stacks transitions and generic assertions, or claims image evidence without identifying a concrete visible detail receives 1. Evidence may be mistaken yet still be genuine support; do not mark 1 merely because the conclusion is wrong.
Positive example. When asked which button submits a form, the response says that the interface layout, button semantics, and operation path must be considered together and therefore chooses the lower-right button, without naming its text, color, or icon.
Negative example. When asked how many apples appear in an image, the response says that one is on the left and two are on the plate on the right, for a total of three.

IV. Required Output
Follow the exact order and format below. Do not output unrelated content and do not omit any category. For each category, first output 1 or 0, then give a one-sentence justification. When a category is positive, identify the key evidence; for logical contradiction, identify both mutually incompatible passages A and B.

```
1. Chain-of-thought leakage: 1 / 0
Reason:
2. Response repetition: 1 / 0
Reason:
3. Logical contradiction: 1 / 0
Reason:
4. Performative reasoning: 1 / 0
Reason:
```

V. Input
Question:
`{question}`

Model response:
`{response}`

APPENDIX J.2 CHINESE VERSION

Response-Pattern Meta-Judge Prompt (Chinese Original)

你是一个严格的模型回复坏模式分析器。请结合用户问题、相关图片（若有）和模型回答，判断模型回答是否存在以下四类坏模式，并为每一类给出1 或0 的判定和一句话理由。

一、判定原则

1. 只依据问题、相关图片（若有）和模型回答来判断，判定要落在模型回答的可观察文本上。
2. 要结合图片核验模型回答里和图像有关的描述、计数、定位、识别、空间关系是否成立；但不要因为答得对不对本身就判坏模式。
3. 事实错误、看图错误、算错、答得不全，这些本身不算坏模式，只有当它们表现为下面某一类可命名的坏模式时才判1。
4. 四个标签各自独立判断，允许同时成立；只有存在明确文本证据时才判1，否则判0。
5. 回答篇幅长、步骤多、解释详细，本身都不是坏模式，不能只因为篇幅长或推理多就判1。

二、判定顺序（重要）

请按下面的固定顺序判断：

1. 先判思维链外溢：通读回答，看是否暴露了本该隐藏的内部思考、自我纠错、试探摇摆，或把解题过程从头摊开的过程腔（按第1 类标准）。
2. 再判病态重复：独立判断，和是否外溢无关（按第2 类标准）。
3. 再判逻辑性冲突和伪装推理：先看那些看起来像矛盾或像空泛推理的内容，是否其实来自暴露出来的内部思考过程（自我纠错、反复试探、过程摇摆、草稿式推演）。若是，则这部分归思维链外溢、不再计入逻辑性冲突或伪装推理（这两项就该内容判0），除非回答在最终交付内容里另有独立的硬矛盾或独立的空壳推理。逻辑性冲突只针对最终交付内容里成对并列、无法同时成立、且未取舍的硬矛盾；伪装推理只针对推理对结论没有有效信息增量、或在低推理负担任务上表演式分析，且该问题不能被思维链外溢更好地解释时。

三、四类坏模式

1. 思维链外溢

定义：模型把本该隐藏的中间推理过程、内部自我对话、试探性分析、任务分解、自我纠错或草稿式思考，直接暴露在最终回复里，影响回答的自然度或不符合输出预期。

判1 必须有明确的内部思考外露痕迹，落在下面任一类上：

- 第一人称过程腔：让我先看看、我先分析一下、让我想想、我来看看、这里我需要判断一下、我得先确认一下，以及我们来逐一分析、我们来逐步分析、我们先看、我们可以先等以让我、我们来、我们先等开头、把怎么看怎么想讲出来、带着读者一步步做题的表述。这类一旦出现，即使后面分点、带小标题，也算外溢。
- 自我纠错、推翻、摇摆、反复试探：不对再看、我前面说错了应该是、等一下重新看、先试A 不对再试B 等中途否定再改或来回试探的痕迹。
- 暴露内部决策或草稿：结束思考、现在确认、我的初步判断是、先在脑子里过一遍等内部决策、执行策略、任务分解的过程，或残留think 这类思考、推理标签。

判0 的情形（没有上面这些痕迹就判0）：

- 面向用户的结构化解题或分析：设了分析过程、步骤一步骤二等板块、把推导分点展开，或以要确定某结论需先分析某某这类非第一人称的客观陈述开头，只要整体是面向用户、成品式地讲题或讲分析（解题步骤、逐项说明、逐图说明），没有让我、我们来这类过程腔、没有自我纠错摇摆、没有暴露内部决策或草稿，就不算外溢。这类回答若分析空转无有效依据，应归伪装推理，不归外溢。
- 用户明确要求展示推理、步骤、草稿，且回答是面向用户的成品式步骤、推导或方法说明（教程、操作流程、解题步骤）。
- 直接交付结论加简短依据或要点。
- 简短的候选排除、比较，或面向用户的多角度、条件分支说明。

示例（判1）：问某物体朝向。回答写让我先看看这张图，嗯好像朝左，不对再看一下应该是朝右，把自己来回试探的过程写了出来。

示例（判0）：问推断某家具朝向。回答设了分析过程板块，按步骤一步骤二客观地把朝向推导讲给用户，没有第一人称自言自语或自我纠错，则不算外溢（若推导无有效据图依据，归伪装推理）。

2. 病态重复

定义：模型在回复里不必要地重复同一句话、同一语义、同一结构或同一结论，这种重复已经明显影响信息效率、阅读体验或回答质量。

判1 的情形：

- 同一结论、同一建议、同一免责声明被连续或高频重复，反复表达同一个意思。
- 同一个信息或结论在多个段落（比如分析、验证、总结、最终答案）反复重申，即使分属不同小节、措辞略有变化，只要没有新增信息，也算重复。
- 段落之间内容高度同质，只是换少量措辞重新说一遍。
- 推理兜圈、候选和线索反复罗列、同一计算反复否定重来等，超出正常强调需要，呈机械化、循环化倾向。

判0 的情形：
- 为了强调重点做的有限重复。
- 枚举或分步结构里必要的局部重复，但各项信息有实质差异。
- 一次性的有效自我纠错、首尾呼应式的一次复述等自然承接，没有造成明显的冗余堆叠。

示例（判1）：问某画的名字。回答先在标题里给出画名和作者，紧接着又用基本信息列表把同样的画名、作者、年代再列一遍，没有新增信息。
示例（判0）：问某题答案。回答给出推导后，结尾用一句话重申最终答案作为收尾，只重复了一次。

3. 逻辑性冲突
定义：模型回答在最终交付内容里，针对同一对象的同一属性、同一个量的取值、或同一个结论，给出两个无法同时为真的说法，且两者都被当作有效信息并列保留，没有择一、没有声明其中一个作废、也没有用条件区分。这是硬矛盾，判1 时必须能成对定位互斥的两处具体文字A 和B。
构成硬矛盾要同时满足：A、B 针对同一对象或同一量或同一结论；处在同一条件、同一时间、同一口径下；逻辑上无法同时为真；模型没有说明其中一个已作废或只是备选；A、B 都在回答里作为有效信息并列保留。
判1 的四类情形（要能指出A、B 两处具体文字）：
1. 同一最终结论并列互斥：对同一问题、同一对象或同一答案位置给出两个无法同时成立的最终结论，且没有择一。例如问藏品名称，先说是青铜尊，最后又写最终答案青铜鼎；又如先说A 正确、B 不符合条件，最终答案却选B。
2. 同一对象属性互斥：对同一对象的同一属性给出无法同时成立的判断，且两处都保留使用。例如华容道里先说b 是横向木块，移动步骤里又写移动竖向木块b；又如先说某函数是奇函数关于原点对称，又说它关于y 轴对称所以是偶函数。
3. 同一量取值互斥：对同一个数值、日期、坐标、数量、角度、面积、温度给出两个无法同时成立的取值，且没有说明修正关系。例如同一底面积先写成一个值后又写成另一个值且都在用；又如说图中共6 个圆，分项却给出左3 右4 共7 个，总数口径互斥；又如点P 坐标先写成(2, 3) 后又写成(3, 2)。
4. 推导结论与最终答案互斥且推导未被放弃：推导中明确得到并保留X，最终答案却给出与X 互斥的Y，且没有声明前文作废。例如步骤里说8、9 是必经点，最终必经点列表却不含8、9；又如推导得x 等于3 并代入验证成立，最终答案却写x 等于5。

判0 的情形：
- 已修正、已放弃或已择一：先出现一种说法，随后明确否定、修正、放弃，或在多个候选中明确选定，最终只保留唯一结论（这类过程摇摆归思维链外溢）。
- 模糊、推测、对冲表达：用可能、大概、更像、倾向于、不排除等表示候选或置信度差异，没有形成两个确定互斥的结论。例如说可能是商代晚期，也不排除西周早期。
- 条件、时间、视角、口径不同且已区分：例如按假设一结果为X、按假设二结果为Y；又如从观察者视角看在左侧、按物体自身朝向看在右侧。
- 答案错误、看图错误、计算错误，但内部没有成对互斥的并列说法：本类不判断答案对错，看错图、数错数、算错值，只要内部没有对同一对象或同一量或同一结论并列保留两个互斥说法，就判0。
- 表达不严谨、同义改写、单位换算、重复说明、推理粗糙、伪装推理，但指不出两处并列保留且无法同时成立的具体说法。例如说是高积云即鱼鳞云（分类层级或俗称），又如说1 米也就是100 厘米（等价换算）。

关键区分：硬矛盾是最终交付内容里对同一对象属性、同一量取值或同一结论并列保留两个无法同时为真的命题，判1；过程摇摆、模糊措辞、看错算错事实错、条件区分或已择一，都判0。
示例（判1）：问从左到右数白头发戴眼镜的人是第几个。回答列表里写第4 项是目标人物，结尾却写所以是第5 个，对同一答案并列保留4 和5 两个互斥结论且没有取舍。
示例（判0）：问图中是什么动物。回答从头到尾只说是金毛犬，并补充长毛、金黄色等依据，没有互相否定的说法。

4. 伪装推理
定义：模型表面上呈现出分析、推导、论证或证据支撑的样子，但这些内容没有为最终结论提供有效信息增量，只是在制造结果有理有据的表象。
有效信息增量指能支持结论的观察、比较、计数、排除、消歧、计算或核验。在图片任务中，有效增量必须落在图中明确可见的内容上，例如：指出具体物体、文字、颜色、数量、位置；比较多个对象的大小、距离、遮挡、方向、形状、状态；依据图中可见路径、墙体、节点、箭头、表格行列、流程分支判断；对多图人物或物体做去重、匹配、排除；读取图中文字、图例、坐标、标签、数值并用于支撑结论。如果只是套用从图中可以看出、结合整体布局、根据视觉特征、综合判断等话术，却没有指出任何具体可见内容，通常属于伪装推理。
判1 的两类情形（仅当回答已经摆出分析、推导、论证、证据支撑的姿态时）：
1. 一是无依据的推理包装：表面用了推理或证据支撑的表达，却没有提供任何能支持结论的实质依据。典型表现：用首先、其次、因此、综上、可见等连接词但中间只是空泛表态；只是重复题干、改写结论、复述用户已给信息；图片任务里声称根据图中某某、从视觉上判断、结合空间布局，却不指出具体元素、位置、数量、文字、颜色、方向、遮挡或关系；堆砌从语义结构上下文三方面看等分析框架却不连接到题面或图中具体信息。例如问图中植物是什么，回答说结合叶片形态、生长习性和整体视觉特征可判断是某植物，却不指出任何具体可见特征。

2. 二是依据不支撑结论：给出一些看似相关的分析或观察，但推不出最终结论，或只是在包装一个本可直接读取、直接观察或用户已给出的答案。典型表现：分析的是A 结论却落到B；前文只是背景介绍推不出答案；识别题靠看起来像、风格符合、造型接近等泛泛印象，而不是能确认身份或对象的真实依据；低推理负担任务里答案本可直接观察读取定位，额外分析没有新增观察、比较、计数、排除、消歧或核验；路线规划只复述用户已给顺序，没有比较位置、道路、距离、可达性或障碍；迷宫或路径题只说某点是关键节点、必经节点，却没有用墙体、通道连通性或替代路线排除来证明。例如认人题只凭穿着正式、气质沉稳像某类演员就下身份结论；又如计数题用小狗有四条腿毛茸茸来包装共4 只这个数。这里还包括识别鉴定类的背景包装：给出名称或身份后主要堆砌生平、作品列表、风格通论、常演角色、历史介绍等与如何从图中确认无关的背景，却不指出图中能锁定结论的可辨识特征（面部、铭文、款识、标志性细节）。

判0 的情形：

- 直接作答或只做简短说明，没有展开分析来包装结论。注意只说结合图片而不指出图中具体要素，按直接作答处理。
- 图片分析有有效可见依据：指出了图中具体可见内容并用来支撑结论，即使简短、概括、啰嗦、不完美。例如说右下角按钮写着提交所以是提交按钮；又如比较两杯水位和容器形状判断哪杯更多。注意：这里要求的是能唯一锁定结论的具体可辨识细节，例如读出文字、铭文、款识，给出确切的数量、位置、可比较的状态；如果只给造型接近、风格符合、科技感、火焰核心、画风色彩相近、看着像这类泛泛特征词，不算有效依据，应按判1 处理。
- 据图做了真实连通性推进但看错（看图错误，判0）：仅当模型确实逐格交代了墙体或通道是否连通、相邻格能否通行，并据此排除替代路线（即给出了能支撑路径结论的连通性核验），只是把某段墙体或通道看错时，才算看图错误判0。如果只是泛泛点名某些格子、节点为关键点或必经点、或只复述各数字所在位置，而没有用连通性或排除来证明，仍按上面的依据不支撑结论判1。
- 用户明确要求说明理由、分析过程、逐步判断或给出依据，且分析不是完全空转。例如用排除法说明为什么选C。
- 看错、算错、事实错，但依据和结论之间存在真实支撑关系。例如基于图中真实对象计数，只是漏数或数错。
- 看不清图片时的谨慎说明：明确表示图片模糊、无法确认，并给出谨慎结论或拒绝断言。

关键区分：直接答题、有效图像依据、有效推理、思维链外溢都判0；看似在分析实则重复题干、复述结论、堆连接词、泛泛表态，或声称结合图中具体特征却不给任何具体可见细节，判1；依据可能错但确实在尝试基于图片或题面支撑结论的，判0。不要因为回答长、步骤多、语气像推理就判1，也不要因为结论错就判1。

示例（判1）：问图中哪个按钮是提交按钮。回答说需要结合界面布局、按钮语义和操作路径综合判断，因此是右下角那个，没有指出按钮文字、颜色或图标等具体可见依据。

示例（判0）：问图中有几个苹果。回答说左侧1 个、右侧盘子里2 个、一共3 个，给出了具体的图像观察和计数依据。

四、输出要求

请严格按下面顺序和格式输出，不要输出无关内容，不要省略任何类别。每一类先给1 或0 的判定，再用一句话说明理由（判1 时点出关键依据，逻辑性冲突要点出互斥的A、B 两处）。

```
1. 思维链外溢：1 / 0
理由：
2. 病态重复：1 / 0
理由：
3. 逻辑性冲突：1 / 0
理由：
4. 伪装推理：1 / 0
理由：
```

五、输入

问题：

```
{question}
```

模型回答：

```
{response}
```