CAN RLVR MODELS THINK MORE CREATIVELY? A STUDY OF OVERCONFIDENCE AND HIGH ENTROPY SEGMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in Large Language Models (LLMs) are often attributed to improved reasoning abilities through Supervised Fine-Tuning (SFT) and Reinforcement Learning with Verifiable Rewards (RLVR). Although RLVR models perform strongly on mathematical benchmarks, their ability to generate creative solutions remains an open question. In this work, we compare RLVR models and math-pretrained models from the DeepSeek and Qwen families on mathematical problem solving with supplied reference solutions. RLVR models yield fewer responses judged both correct and novel, while their self-scoring high lower-tail token likelihood (HL) rates exceed those of the Math models by 15.85 and 74.70 percentage points, respectively. HL operationalizes token-level Overconfidence (OC) in this study, rather than calibration error. Motivated by the hypothesis that SFT and policy optimization reshape token probabilities, we analyze High Entropy Segments (HES), defined as windows whose mean token entropy exceeds the response mean. Cross-scoring reveals scorer-dependent HES coverage and overlap, while the association between HL and novelty depends on the outcome denominator and likelihood threshold. Furthermore, RLVR models have lower off-argmax ratios under self-scoring, accompanied by a higher fraction of positions with top-token probability at least 0.80. These observational findings characterize the tested RLVR models and suggest directions for improving creative generation without isolating causal training effects.

1 INTRODUCTION

Can models trained with Reinforcement Learning with Verifiable Rewards (RLVR) produce more creative mathematical solutions than their math-pretrained counterparts? Improved answer accuracy alone does not answer this question: a model must also find an approach that differs from the solutions already available to it. We study this question in the DeepSeek and Qwen families, comparing Math and RLVR models on CreativeMath and examining their token-probability profiles. Figure 1 shows the model families and training stages.

We operationalize creative problem solving as *reference-relative novelty*: a response passes two LLM judges' correctness checks and at least one judge identifies a substantive difference from the supplied solutions. CreativeMath already distinguishes correctness from novelty (Ye et al., 2025). Our focus is the Math–RLVR comparison and the hypothesis that concentrated token probabilities accompany reduced creative generation. We evaluate four generators—DeepSeek-Math, DeepSeek-RLVR, Qwen-Math, and Qwen-RLVR—and score the same responses with several models from each family. RLVR denotes the models listed in Table 1; their training pipelines also include SFT, so this comparison evaluates the resulting models rather than an isolated RL update.

Both RLVR generators produce fewer responses judged correct and novel than their Math counterparts, despite higher correctness-acceptance rates. We examine this behavioral gap alongside token likelihood, entropy-window structure, and observed-prefix token choices. Rescoring the same responses distinguishes properties of the generated response collection from the probability profile assigned by a particular model. Three observations organize our analysis:

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.

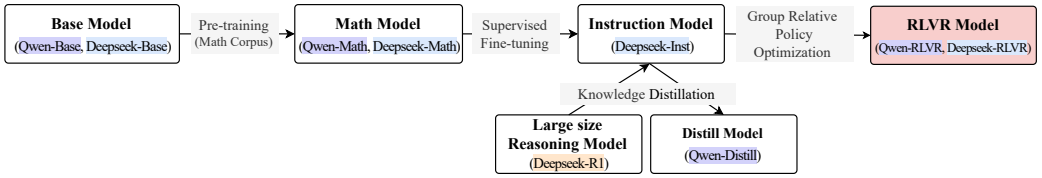


Figure 1: Schematic of the model families and training stages: mathematical pre-training, supervised fine-tuning, knowledge distillation, and reinforcement learning with group relative policy optimization.

1. **Higher judge acceptance coexists with lower accepted-and-novel yield.** Under our reference-conditioned evaluation protocol, both RLVR generators have higher correctness acceptance but lower accepted-and-novel yield than their Math counterparts (Table 2). These rates describe the retained response sets.

2. **The self-scoring probability profile alone does not explain the yield gap.** RLVR responses have higher self-scoring HL frequencies and lower OAR (Tables 3 and 7). Yet the HL–yield association differs between the two RLVR generators, and the same responses receive different probability profiles under other scorers.

3. **Scorer, outcome denominator, and likelihood threshold determine the comparison.** Qwen HESR increases with either scorer held fixed but decreases under self-scoring (Table 6). For QW-RLVR, the HL–novelty contrast reverses with the outcome denominator and with the tested likelihood thresholds (Tables 3 and 4).

The behavioral comparison establishes the yield gap; the probability diagnostics identify which interpretations depend on evaluation choices. Together, these observations motivate reporting accepted-and-novel yield alongside scorer-specific diagnostics with explicit denominators and thresholds. The analysis characterizes the tested checkpoints and their responses rather than isolating a causal training effect.

2 Related Work

Reference-relative mathematical novelty. Ye et al. (2025) introduced CreativeMath and evaluated whether generated solutions differ from supplied references. That work already identified limitations of math-specialized models on this task. Our additional focus is the within-family checkpoint comparison and rescoring of fixed responses, rather than the general observation that correctness and novelty differ. Literature results are listed separately in Appendix D.2; their generation and judging protocols are not interchangeable with ours.

Calibration and token likelihood. Calibration compares predicted confidence with empirical correctness; expected calibration error aggregates their discrepancies within bins (Guo et al., 2017). Token likelihood has also been investigated as a signal for detecting unreliable generation (Lee, 2023; Quevedo et al., 2025). Here, a lower quantile of observed-token likelihood is a scoring statistic. It is neither a probability of solution correctness nor a calibration-error estimate. We distinguish token likelihood, verbalized confidence, and calibration.

RLVR and entropy. DeepSeekMath and Qwen2.5-Math describe mathematical pre-training, SFT, and GRPO-based reinforcement learning (Shao et al., 2024; Yang et al., 2024). Studies of reasoning models have examined inference-time sampling and the role of high-entropy token positions (Yue et al., 2025; Wang et al., 2025). Our entropy-window statistic instead selects windows relative to each response's mean entropy. It does not identify semantic branching tokens, quantify sudden entropy changes, or establish the effect of a training intervention.

Table 1: Checkpoint identities and roles. G: generator; S: scorer. All generators are scored by the four checkpoints in their own family. DS-Inst and QW-Distill are not evaluated as generators.

Alias	Checkpoint	Role
DS-Base	DeepSeek-Coder-Base-v1.5-7B	S
DS-Math	DeepSeek-Math-7B-Base	G, S
DS-Inst	DeepSeek-Math-7B-Instruct	S
DS-RLVR	DeepSeek-Math-7B-RL	G, S
QW-Base	Qwen2.5-7B	S
QW-Math	Qwen2.5-Math-7B	G, S
QW-RLVR	Qwen2.5-Math-7B-Instruct	G, S
QW-Distill	DeepSeek-R1-Distill-Qwen-7B	S

3 Experimental Setup

3.1 Generators, Scorers, and Judges

We distinguish three roles. A *generator* produces a response; a *scorer* assigns token probabilities to that fixed response; a *judge* assigns correctness and novelty labels. Four checkpoints serve as both generators and scorers, while four additional checkpoints serve only as scorers.

DeepSeek-Math continues DeepSeek-Coder-Base pre-training; DS-Inst adds SFT, and DS-RLVR applies GRPO with a learned reward model to DS-Inst (Shao et al., 2024). Qwen2.5-Math-Instruct combines SFT and GRPO, with both verifier and learned reward signals described in its technical report (Yang et al., 2024). No intermediate Qwen SFT-only checkpoint is included in this study. QW-Distill supplies an additional scoring reference (Guo et al., 2025). These roles do not support a conclusion about its novel-solution generation ability.

3.2 Task and Generation Settings

We use a 400-problem subset of CreativeMath, with 50 problems from each of eight competitions (Ye et al., 2025). Prompts supply up to two human-written reference solutions and request a different approach. We sample responses at temperature 0.7, top-$p = 1.0$, top-$k = 50$, and a maximum of 1024 output tokens. Responses reaching the output-token cap or containing attempted API calls are regenerated with at most 20 retries. A temperature-1.4 comparison is reported separately.

The analysis unit is a retained response to a reference-conditioned prompt. We evaluate 1510 DS-Math, 1585 DS-RLVR, 1527 QW-Math, and 1531 QW-RLVR responses. Table 8 gives their class counts. Outcome rates describe these retained response sets rather than pass@k or an equal-budget comparison. Multiple responses can share a problem, so N counts responses, not independent problems; the comparisons below are descriptive response-level rates.

3.3 Judge Labels and Outcome Denominators

The two judges are Gemini-1.5-Pro (`gemini-1.5-pro-002`) and o4-mini (`o4-mini-2025-04-16`). A response receives correctness acceptance only if both judges answer yes to the reference-result agreement prompt. Among those responses, at least one positive novelty judgment assigns the label Crea; the remainder receive Typi. Fail denotes failure to obtain agreement on correctness. The correctness prompt asks whether the response reaches the same result as the supplied references.

For a retained set of $N = n_{\text{Fail}} + n_{\text{Crea}} + n_{\text{Typi}}$ responses, the metrics are

$$\text{Cor} = \frac{n_{\text{Crea}} + n_{\text{Typi}}}{N}, \quad \text{Cre} = \frac{n_{\text{Crea}}}{N}, \quad \text{Cre/Cor} = \frac{n_{\text{Crea}}}{n_{\text{Crea}} + n_{\text{Typi}}}. \tag{1}$$

The same denominators apply within a likelihood group. Cre/Cor conditions on each generator's own judge-accepted responses, so its comparison reflects both acceptance and the novelty composition of the accepted set.

Table 2: Response-level outcomes at temperature 0.7. Cor is acceptance by both correctness judges; Cre additionally requires reference-relative novelty; Cre/Cor is the fraction labeled novel among accepted responses. N is the number of retained responses (Table 8).

Generator	N	Cor (%)	Cre (%)	Cre/Cor (%)
DS-Math	1510	35.70	16.75	46.94
DS-RLVR	1585	44.04	11.42	25.93
QW-Math	1527	56.65	22.92	40.46
QW-RLVR	1531	58.07	15.81	27.22

3.4 Reporting Conventions

We compute Cor and Cre over retained responses and Cre/Cor over correctness-accepted responses. For each generator, every scorer partitions the same response set into HL and non-HL. We calculate rates and between-model differences before rounding and express differences between percentages in percentage points. Conditional rates are undefined for empty groups.

Scorers evaluate fixed responses and judge labels, but assign different probabilities and HL groups. Entropy, HESR, and Jaccard report response-level means and standard deviations. OAR and top-token concentration pool scored positions, weighting responses by length (Equation 8). Values are rounded; standard deviations are not confidence intervals.

4 Behavioral Comparison

Table 2 shows that correctness acceptance and reference-relative novelty rank the checkpoints differently within each family. In DeepSeek, Cor rises from 35.70% for Math to 44.04% for RLVR, while Cre falls from 16.75% to 11.42% and Cre/Cor falls from 46.94% to 25.93%. Qwen exhibits the same directions: Cor rises from 56.65% to 58.07%, while Cre falls from 22.92% to 15.81% and Cre/Cor falls from 40.46% to 27.22%.

The effect sizes differ across families. DeepSeek's correctness-acceptance gap is 8.34 percentage points, compared with 1.42 for Qwen; the novelty-rate decreases are 5.34 and 7.11 points, respectively, calculated from the class counts before rounding. This establishes Observation 1: higher judge acceptance coexists with lower accepted-and-novel yield. The diagnostics examine how the self-scoring profile relates to this gap (Observation 2) and how scorer, denominator, and threshold choices shape its interpretation (Observation 3). They investigate the scope of probability-based explanations rather than identify the gap's causal origin.

5 Token-Probability Diagnostics

Let $y = (y_1, \ldots, y_T)$ be a fixed generated response and let e index a scorer in the generator's family. We write $p_{e,t}(v)$ for that scorer's next-token distribution at the observed prefix. A change of scorer holds the response text fixed; a change of generator changes the response collection. This distinction is maintained in the comparisons below.

5.1 Entropy and Lower-Tail Token Likelihood

At each scored position, token entropy is

$$H_{e,t} = -\sum_{v \in V_e} p_{e,t}(v) \log p_{e,t}(v). \tag{2}$$

This is entropy of the vocabulary distribution, not the surprisal of the observed token. The latter token's probability enters the lower-tail likelihood statistic

$$L_{e,q}(y) = \exp\big(\mathrm{Quantile}_q\{\log p_{e,t}(y_t)\}_{t=1}^{T}\big), \qquad \mathrm{HL}_e(y) = \mathbf{1}\{L_{e,q}(y) > \theta\}. \tag{3}$$

HL denotes *high lower-tail token likelihood*; non-HL denotes its complement. This is the operational statistic for the token-level Overconfidence (OC) terminology used in the title and abstract. We

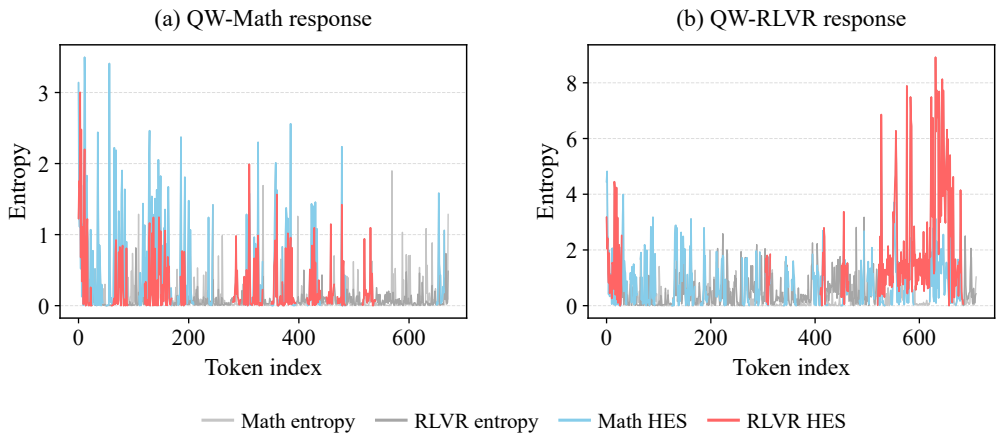


Figure 2: Entropy profiles and HES for two different responses. Each panel compares Math and RLVR scorers on the same response; cross-scorer comparisons are within panels. Colored curve segments mark HES relative to each scorer's response mean. The panels use separate entropy scales.

use $q = 5\%$ and $\theta = 0.35$ for the main summaries and report the existing threshold variants in Appendix C. Unlike calibration overconfidence, HL does not compare confidence with correctness.

5.2 Relative High-Entropy Windows

For a width-W window starting at i, let $\mu_e^{(i)} = W^{-1} \sum_{t=i}^{i+W-1} H_{e,t}$ and let $\mu_e = T^{-1} \sum_{t=1}^{T} H_{e,t}$. With width $W = 10$ and stride $S = 5$, a High Entropy Segment (HES) is a selected window:

$$\mathrm{HES}_e(i) = \mathbf{1}\{\mu_e^{(i)} > \mu_e\}. \tag{4}$$

The mask $M_e(t)$ is one if position t belongs to at least one selected window, and zero otherwise. Window coverage and mask overlap are

$$\mathrm{HESR}_e = \frac{1}{T} \sum_{t=1}^{T} M_e(t), \tag{5}$$

$$J(g, e) = \frac{|\{t : M_g(t) = 1\} \cap \{t : M_e(t) = 1\}|}{|\{t : M_g(t) = 1\} \cup \{t : M_e(t) = 1\}|}. \tag{6}$$

Here M_g is the mask obtained by rescoring the same response with its generator. HESR measures coverage by a union of windows; it is not the proportion of individual tokens whose entropy exceeds the mean. The criterion detects relative level, not a derivative or an abrupt change. For fixed windows it is invariant to $H_{e,t} \mapsto aH_{e,t} + b$ with $a > 0$, but this property does not imply independence from length or from other response characteristics. It does not identify semantic fork tokens.

5.3 Departures from the Scorer's Tokenwise Argmax

$$\mathrm{OAR}_e(y) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left\{ y_t \neq \arg\max_{v \in V_e} p_{e,t}(v) \right\}. \tag{7}$$

The off-argmax ratio (OAR) compares the observed token with the scorer's preferred token at the *observed* prefix. It is not a distance from a separately generated greedy sequence, whose prefix can change after the first departure. For a response collection $\mathcal{D}$, the tabled OAR pools token positions:

$$\mathrm{OAR}_e(\mathcal{D}) = \frac{\sum_{y \in \mathcal{D}} T_y \, \mathrm{OAR}_e(y)}{\sum_{y \in \mathcal{D}} T_y}. \tag{8}$$

The same convention applies separately to HL and non-HL, and to the fraction of scored positions with $\max_v p_{e,t}(v) \geq 0.80$. Thus the overall OAR combines group OARs using token counts, whereas Cor and Cre combine groups using response counts. Probability concentration and argmax departures describe related aspects of token choice; neither directly measures reference-relative novelty.

Table 3: High lower-tail likelihood groups ($q = 5\%$, $\theta = 0.35$). Ratio is the percentage of retained responses assigned to each group by the scorer. Cor and Cre are computed within the group; Cre/Cor conditions on correctness acceptance. All values are percentages. **Bold** marks the highest HL ratio for each generator.

		HL				non-HL			
Generator	**Scorer**	**Ratio**	**Cor**	**Cre**	**Cre/Cor**	**Ratio**	**Cor**	**Cre**	**Cre/Cor**
DS-Math	Base	9.87	65.10	22.15	34.02	90.13	32.48	16.16	49.77
	Math	**13.11**	60.61	21.72	35.83	86.89	31.94	16.01	50.12
	Inst	9.14	63.77	21.74	34.09	90.86	32.87	16.25	49.45
	RLVR	7.95	65.83	22.50	34.18	92.05	33.09	16.26	49.13
DS-RLVR	Base	0.63	70.00	0.00	0.00	99.37	43.87	11.49	26.19
	Math	0.76	58.33	0.00	0.00	99.24	43.93	11.51	26.19
	Inst	7.57	45.00	7.50	16.67	92.43	43.96	11.74	26.71
	RLVR	**28.96**	46.19	8.71	18.87	71.04	43.16	12.52	29.01
QW-Math	Base	0.59	55.56	0.00	0.00	99.41	56.65	23.06	40.70
	Math	**8.97**	72.99	26.28	36.00	91.03	55.04	22.59	41.05
	RLVR	3.14	75.00	22.92	30.56	96.86	56.05	22.92	40.89
	Distill	0.20	100.00	0.00	0.00	99.80	56.56	22.97	40.60
QW-RLVR	Base	3.59	85.45	25.45	29.79	96.41	57.05	15.45	27.08
	Math	8.75	80.60	26.87	33.33	91.25	55.91	14.75	26.38
	RLVR	**83.67**	63.70	17.02	26.72	16.33	29.20	9.60	32.88
	Distill	1.50	69.57	30.43	43.75	98.50	57.89	15.58	26.92

6 Diagnostic Results

The diagnostics develop Observations 2 and 3. Tables 4 and 6 provide the central comparisons: threshold-sensitive conditional novelty and scorer-dependent HESR. Tables 3 and 5 supply the full generator–scorer results, and Table 7 completes the probability profile with observed-prefix token choices.

6.1 Likelihood Frequency and Conditional Outcomes

The self-scoring HL frequencies are 28.96% for DS-RLVR and 83.67% for QW-RLVR (Table 3), compared with 13.11% for DS-Math and 8.97% for QW-Math. On the RLVR response collections, other scorers assign lower HL frequencies: 0.63–7.57% for DeepSeek and 1.50–8.75% for Qwen. Thus the same RLVR responses receive markedly different HL assignments depending on the scorer.

The link between HL and yield differs across the two RLVR generators. Self-scored DS-RLVR has Cre of 8.71% in HL and 12.52% in non-HL; its Cre/Cor values are 18.87% and 29.01%. For QW-RLVR, the same novel-response counts produce opposite comparisons under the two denominators:

$$\begin{aligned} \text{Cre}: &\quad \frac{218}{1281} = 17.02\% > \frac{24}{250} = 9.60\%, \\ \text{Cre/Cor}: &\quad \frac{218}{816} = 26.72\% < \frac{24}{73} = 32.88\%. \end{aligned} \tag{9}$$

In each comparison the HL group is on the left. Cre measures the yield of accepted, novel responses over the entire group; Cre/Cor measures the novel fraction among accepted responses. Thus the higher HL yield and lower accepted-response novelty fraction describe different outcomes. Writing $C_G = \text{Cor}_G$ and $R_G = (\text{Cre/Cor})_G$ as proportions, with H and N denoting HL and non-HL, gives

$$\begin{aligned} \text{Cre}_G &= C_G R_G, \qquad G \in \{H, N\}, \\ \text{Cre}_H - \text{Cre}_N &= (C_H - C_N) R_N + C_H (R_H - R_N). \end{aligned} \tag{10}$$

For QW-RLVR, correctness acceptance is 63.70% in HL and 29.20% in non-HL. This acceptance difference offsets the lower novel fraction among accepted HL responses, producing the higher overall HL yield.

Table 4: QW-RLVR self-scoring: the conditional-novelty contrast changes sign across settings. N counts responses and A counts judge-accepted responses within each group; $R = \mathrm{Cre}/\mathrm{Cor}$ (%). $\Delta R = R_{\mathrm{HL}} - R_{\text{non-HL}}$ in percentage points. The default is $q = 5\%$, $\theta = 0.35$. All rows partition the same 1,531 responses, including 889 accepted responses.

			HL			non-HL			
q (%)	θ	HL (%)	N	A	R	N	A	R	ΔR
5	0.30	90.99	1393	854	27.52	138	35	20.00	+7.52
5	0.35	83.67	1281	816	26.72	250	73	32.88	-6.16
5	0.40	75.44	1155	770	26.23	376	119	33.61	-7.38
3	0.35	55.58	851	635	26.14	680	254	29.92	-3.78
7	0.35	94.45	1446	866	27.60	85	23	13.04	+14.55

Table 5: Same-response token diagnostics across generators and scorers. **Length** is the total scored-token count (Table 14) divided by the retained response count, rounded to one decimal place. **Entropy**, **HESR**, and **Jaccard** are reported as $\mu \pm \sigma$ over responses. Jaccard compares each scorer's HES mask with the generator's mask for the same response.

Generator	Scorer	Length	Entropy	HESR	Jaccard
DS-Math	Base	532.8	0.73 ± 0.38	0.43 ± 0.12	0.90 ± 0.07
	Math	532.8	0.69 ± 0.37	0.43 ± 0.12	1.00 ± 0.00
	Inst	532.8	0.47 ± 0.25	0.44 ± 0.13	0.85 ± 0.09
	RLVR	532.8	0.43 ± 0.23	0.44 ± 0.13	0.83 ± 0.09
DS-RLVR	Base	426.9	0.77 ± 0.26	0.52 ± 0.08	0.79 ± 0.09
	Math	426.9	0.68 ± 0.25	0.50 ± 0.08	0.81 ± 0.10
	Inst	426.9	0.37 ± 0.15	0.50 ± 0.08	0.86 ± 0.08
	RLVR	426.9	0.32 ± 0.13	0.50 ± 0.08	1.00 ± 0.00
QW-Math	Base	651.7	0.62 ± 0.21	0.48 ± 0.07	0.72 ± 0.10
	Math	651.7	0.46 ± 0.19	0.44 ± 0.07	1.00 ± 0.00
	RLVR	651.7	1.00 ± 0.61	0.36 ± 0.10	0.56 ± 0.14
	Distill	651.7	0.46 ± 0.17	0.48 ± 0.06	0.73 ± 0.09
QW-RLVR	Base	642.0	0.44 ± 0.12	0.52 ± 0.06	0.56 ± 0.12
	Math	642.0	0.36 ± 0.11	0.49 ± 0.06	0.58 ± 0.12
	RLVR	642.0	0.40 ± 0.30	0.43 ± 0.08	1.00 ± 0.00
	Distill	642.0	0.30 ± 0.11	0.50 ± 0.06	0.58 ± 0.11

Table 4 gives the first central comparison across all five QW-RLVR self-scoring settings. The HL-minus-non-HL difference in Cre/Cor is positive at $q = 5\%$, $\theta = 0.30$ and at $q = 7\%$, $\theta = 0.35$, but negative in the other three settings. The positive comparison at $q = 7\%$ uses only 23 accepted non-HL responses, of which three are labeled novel. This small denominator makes the reversal sensitive to individual labels. Reusing the same responses across settings isolates threshold sensitivity within this collection. High self-scoring HL frequency is therefore more stable than the sign of its conditional association with novelty.

6.2 Entropy with Generator and Scorer Held Explicit

Table 5 provides full token diagnostics supporting Table 6. Holding the scorer fixed, QW-RLVR responses have lower mean entropy than QW-Math responses for each of the four Qwen scorers. The same direction holds for three DeepSeek scorers; DS-Base is an exception, assigning 0.77 to DS-RLVR responses and 0.73 to DS-Math responses. The entropy contrast therefore depends on the scorer.

Holding the response collection fixed instead, QW-RLVR assigns mean entropy 1.00 to QW-Math responses, compared with 0.46 under the QW-Math scorer. On QW-RLVR responses the corresponding means are 0.40 and 0.36. The entropy contrast consequently depends on which axis is held fixed. Mean response lengths are 532.8 versus 426.9 tokens for DS-Math and DS-RLVR, and 651.7 versus 642.0 for QW-Math and QW-RLVR.

Table 6: Qwen HESR by generator and scorer, extracted from Table 5. Columns hold the scorer fixed; bold entries form the self-scoring comparison. The fixed-scorer increases and self-scoring decrease concern different comparisons.

Generator	QW-Math scorer	QW-RLVR scorer
QW-Math	**0.44**	0.36
QW-RLVR	0.49	**0.43**

6.3 Coverage and Alignment of Entropy Windows

Holding each scorer fixed, HESR is higher on RLVR-generated than Math-generated responses in both families. The second central comparison, Table 6, isolates the two axes for Qwen: with the Math scorer, coverage rises from 0.44 to 0.49; with the RLVR scorer, it rises from 0.36 to 0.43. Comparing the self-scoring diagonal instead gives a decrease from 0.44 to 0.43. Self-scoring changes both the generated response collection and the scorer, whereas a fixed-scorer comparison changes only the collection. The scorer change offsets the fixed-scorer increase in Qwen, reversing the direction of the comparison. For DeepSeek, self-scored HESR rises from 0.43 to 0.50.

Let $F(g, e)$ denote mean HESR for generator g and scorer e, and abbreviate QW-Math and QW-RLVR as M and R. The self-scoring difference decomposes into two changes:

$$\begin{aligned} F(R,R) - F(M,M) &= [F(R,M) - F(M,M)] + [F(R,R) - F(R,M)] \\ &= (0.49 - 0.44) + (0.43 - 0.49) = -0.01. \end{aligned} \tag{11}$$

Using the rounded table means, the generator contrast under the Math scorer contributes $+0.05$, while changing the scorer on RLVR responses contributes -0.06. This separates the two axes underlying the self-scoring comparison.

Jaccard values compare each scorer's mask with the generator's mask on the same response. For QW-Math responses, overlap is 0.72 with QW-Base and 0.73 with QW-Distill, but 0.56 with QW-RLVR. For QW-RLVR responses, the three other scorers give 0.56, 0.58, and 0.58. DeepSeek shows a less uniform pattern: overlap with DS-Base falls from 0.90 on DS-Math responses to 0.79 on DS-RLVR responses, while overlap with DS-Inst is 0.85 and 0.86, respectively. The window alignment thus depends on both the generator and the scorer; QW-Distill contributes a scoring comparison only.

We also compare the distillation checkpoint with the RLVR scorer. On QW-RLVR responses, QW-Distill assigns mean entropy 0.30, below 0.40 under the RLVR scorer, but labels only 1.50% of those responses HL, compared with 83.67% under self-scoring. Low vocabulary entropy and high likelihood of the observed tokens therefore need not move together: a scorer can concentrate probability on tokens other than the observed continuation. Its HES overlap with the QW-RLVR generator is 0.58, compared with 0.73 on QW-Math responses. The distillation scorer therefore supplies a further example of how entropy, observed-token likelihood, and HES alignment capture distinct aspects of a probability profile.

6.4 Observed-Prefix Token Choices

Each generator has its lowest OAR under its own scorer. Self-scoring OAR is 11.10% for DS-Math versus 7.32% for DS-RLVR, and 9.10% for QW-Math versus 4.02% for QW-RLVR (Table 7). These token-pooled values measure local choices at observed prefixes. Holding the responses fixed, DS-RLVR has OAR 7.32% under their own scorer but 13.99% under DS-Math, while QW-RLVR responses have OAR 4.02% under their own scorer but 8.80% under QW-Math.

The HL/non-HL split further localizes the off-argmax difference. Under self-scoring, DS-Math has OAR 2.86% in HL and 12.59% in non-HL, compared with 4.77% and 8.74% for DS-RLVR. The corresponding Qwen values are 4.96% and 9.56% for Math, and 3.39% and 6.78% for RLVR. Thus all four generators depart from their scorer's preferred token less often in HL. However, DS-RLVR's HL OAR exceeds DS-Math's even though its overall OAR is lower. The overall Math–RLVR contrast combines within-group token choices with different likelihood-group frequencies and token totals.

Holding DS-RLVR responses fixed, the fraction of positions with top-token probability at least 0.80 is 68.68% under DS-Math and 81.11% under DS-RLVR. On QW-RLVR responses it is 79.74% under

Table 7: Observed-prefix token choices and probability concentration. OAR compares each observed token with the scorer's argmax and pools token positions within each column's response set (Equation 8). HL uses $q = 5\%$, $\theta = 0.35$. Underlines mark Math/RLVR scorers in the Top-1 column. Top-1 is the percentage of scored positions with maximum next-token probability at least 0.80.

		OAR (%)			Top-1 (%)
Generator	Scorer	All	HL	non-HL	$\geq \mathbf{0.80}$
DS-Math	Base	13.01	2.78	14.33	71.31
DS-Math	Math	11.10	2.86	12.59	<u>72.53</u>
DS-Math	Inst	14.17	3.16	15.51	77.49
DS-Math	RLVR	15.01	3.07	16.24	<u>78.38</u>
DS-RLVR	Base	16.04	3.40	16.15	65.46
DS-RLVR	Math	13.99	3.55	14.12	<u>68.68</u>
DS-RLVR	Inst	9.54	4.82	10.15	78.92
DS-RLVR	RLVR	7.32	4.77	8.74	<u>81.11</u>
QW-Math	Base	12.85	4.55	12.90	69.14
QW-Math	Math	9.10	4.96	9.56	<u>75.47</u>
QW-Math	RLVR	11.02	4.64	11.27	<u>77.11</u>
QW-Math	Distill	15.93	4.00	15.96	74.07
QW-RLVR	Base	10.35	5.01	10.54	75.93
QW-RLVR	Math	8.80	4.93	9.21	<u>79.74</u>
QW-RLVR	RLVR	4.02	3.39	6.78	<u>87.01</u>
QW-RLVR	Distill	11.94	5.17	12.07	80.87

QW-Math and 87.01% under QW-RLVR. These are differences of 12.43 and 7.27 percentage points in probability concentration. Argmax tokens can express a different mathematical method, while alternative tokens can merely change wording; novelty is assessed at the solution level by the judges.

On Math-generated responses, the concentration gaps are smaller: 72.53% versus 78.38% for DS-Math (5.85 points), and 75.47% versus 77.11% for QW-Math (1.64 points). In both families, the RLVR scorer therefore increases concentration more on its own responses than on Math responses. Each contrast holds the response collection fixed.

At the tested temperature of 1.4, all four generators have lower Cor and Cre than at 0.7 (Appendix B.3). The tested increase in sampling temperature thus reduces both acceptance and accepted-and-novel yield under this protocol.

Connecting the three observations. Higher self-scoring HL and lower OAR accompany the behavioral gap (Tables 2, 3, and 7), but HL has opposite yield associations in the two RLVR generators. Equation 9 and Tables 4 and 6 show why the interpretation requires explicit denominators, thresholds, and scorers: each choice can reverse a diagnostic contrast within the same evaluated response collections. The self-scoring profile alone therefore provides an incomplete account of the yield gap. A useful probability-based account must therefore specify which response-level outcome it explains and which generator–scorer comparison supplies its evidence.

7 Conclusion

On the retained reference-conditioned responses, the tested RLVR models have higher judge correctness acceptance but lower accepted-and-novel yield than their Math counterparts. Higher self-scoring likelihood and lower OAR accompany this gap, yet opposite HL–yield associations in the two RLVR generators limit a uniform explanation based on this profile. The diagnostics identify three choices that shape interpretation: the outcome denominator reverses the QW-RLVR HL–novelty contrast, likelihood thresholds change its conditional direction (Table 4), and scorer choice changes the HESR comparison (Table 6). Evaluations of creative generation in RLVR models should pair behavioral outcomes with probability diagnostics that explicitly state their scorer, likelihood threshold, and outcome denominator.

IMPACT STATEMENT

This study encourages evaluation beyond final-answer accuracy. Its findings concern reference-relative novelty and automated judge acceptance, not general creativity or deployment suitability. These labels and token-probability diagnostics describe model behavior; practical applications require independent validation of mathematical reasoning and originality.

REPRODUCIBILITY STATEMENT

We describe the experimental setup and metrics in Sections 3–5. The appendices provide the prompts, response-class distributions, and temperature and likelihood-threshold comparisons.

AI USE STATEMENT

Generative AI tools assisted with manuscript writing and editing, LaTeX formatting, presentation and interpretation of reported results, and consistency checks of manuscript tables. Experimental language-model roles are described in Section 3. The authors are responsible for the final manuscript.

REFERENCES

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Minhyeok Lee. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*, 11(10):2320, 2023.

Ernesto Quevedo, Jorge Yero Salazar, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach. In *Artificial Intelligence and Applications (CSCE 2024)*, volume 2252 of *Communications in Computer and Information Science*, pp. 154–173. Springer, 2025. doi: 10.1007/978-3-031-86623-4_13.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

Junyi Ye, Jingyi Gu, Xinyun Zhao, Wenpeng Yin, and Guiling Wang. Assessing the creativity of llms in proposing novel solutions to mathematical problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25687–25696, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

A LIMITATIONS

The experiments compare existing checkpoints without training interventions. The checkpoint comparison jointly reflects SFT, RL, pre-training data, prompt compatibility, and other model differences. Only 7B generators from two families are included, DS-Inst is a scorer only, and no Qwen SFT-only generator is evaluated. QW-Distill contributes a scoring comparison on the four generators' responses.

The outcome is novelty relative to supplied references under two automated judges. It is neither general creativity nor independently verified mathematical originality. The correctness prompt focuses on agreement with the reference result, so invalid intermediate reasoning can survive. The novelty prompt also admits differences in assumptions, generality, or complexity; a changed assumption can alter the original problem without being rejected by the answer-agreement check.

The comparisons concern retained responses under the tested prompts and decoding settings. Regeneration and exclusion can affect the response distribution, so the results do not characterize unfiltered attempts or equal-budget pass@k. Responses sharing a problem are clustered observations. Our response-level rates give more weight to problems with more retained responses; they are not estimates based on independent problem-level trials. We interpret between-model differences as descriptive response-level comparisons. HL describes token likelihood, and HES describes relative entropy-window coverage, so neither directly validates a reasoning mechanism.

B GENERATION AND LABELING DETAILS

B.1 RESPONSE GENERATION

B.1.1 PROMPT

Criteria for evaluating the difference between two mathematical solutions include:
i). If the methods used to arrive at the solutions are fundamentally different, such as algebraic manipulation versus geometric reasoning, they can be considered distinct;
ii). Even if the final results are the same, if the intermediate steps or processes involved in reaching those solutions vary significantly, the solutions can be considered different;
iii). If two solutions rely on different assumptions or conditions, they are likely to be distinct;
iv). A solution might generalize to a broader class of problems, while another solution might be specific to certain conditions. In such cases, they are considered distinct;
v). If one solution is significantly simpler or more complex than the other, they can be regarded as essentially different, even if they lead to the same result.

Given the following mathematical problem:
{problem}

And some typical solutions:
{reference_solutions}

Please output a novel solution distinct from the given ones for this math problem.

Figure 3: Generation prompt with a problem, supplied reference solutions, and criteria for a different solution.

The prompt follows the reference-relative task of Ye et al. (2025). Human-written references define the comparison set, rather than certifying the generated response.

B.1.2 Generation Settings and Retention

We use up to two references, temperature 0.7, top-$p = 1.0$, top-$k = 50$, a 1024-token output cap, and six NVIDIA RTX A5000 GPUs (24 GB each). Multiple responses are sampled for the reference-conditioned prompts. Responses reaching the token cap or containing attempted API calls are regenerated, with at most 20 retries.

The retained response sets contain 1510 DS-Math, 1585 DS-RLVR, 1527 QW-Math, and 1531 QW-RLVR responses. Figure 4 summarizes the regeneration attempts; Table 8 gives the class counts used in the analysis.

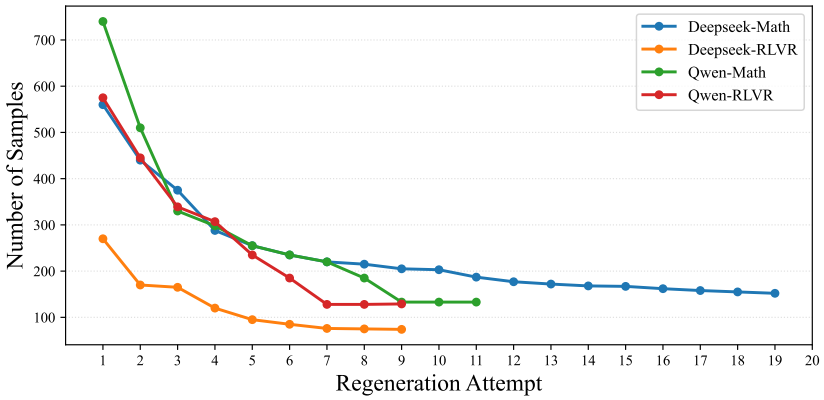


Figure 4: Number of samples requiring regeneration at each attempt for the four generators, with at most 20 retries.

B.2 Labeling of Generated Responses

B.2.1 Judges

The judge versions are Gemini-1.5-Pro (`gemini-1.5-pro-002`) and o4-mini (`o4-mini-2025-04-16`). These are distinct from the checkpoints that score token probabilities.

B.2.2 Correctness and Novelty Rules

Only responses with two positive correctness judgments proceed to novelty assessment. A positive novelty judgment from at least one judge yields Crea; otherwise the accepted response receives Typi. The Fail category includes every response without correctness agreement, including judge disagreement. Crea and Typi summarize the two-judge labeling rules.

B.2.3 Reference Solutions and Response Labels

Human-written solutions provide the reference set. The two LLM judges assign labels to model-generated responses using the correctness and novelty prompts above.

Given the following mathematical problem:
{problem}

Reference solutions:
{reference_solutions}

New solution:
{new_solution}

Please output "YES" if the new solution leads to the same result as the reference solutions; otherwise, output "NO".
YES or NO?

Figure 5: Correctness prompt. It asks whether the new solution reaches the same result as the references; it does not explicitly require a step-by-step proof audit.

Criteria for evaluating the novelty of a new mathematical solution include:
1. If the new solution used to arrive at the solutions is fundamentally different from reference solutions, such as algebraic manipulation versus geometric reasoning, it can be considered novel;
2. Even if the final results are the same, if the intermediate steps or processes involved in reaching those solutions vary significantly, the new solution can be considered novel;
3. If the new solution relies on different assumptions or conditions, it should be considered novel;
4. A solution might generalize to a broader class of problems, while another solution might be specific to certain conditions. In such cases, they are considered distinct;
5. If the new solution is significantly simpler or more complex than the others, it can be regarded as essentially novel, even if they lead to the same result.

Given the following mathematical problem:
{problem}

Reference solutions:
{reference_solutions}

New solution:
{new_solution}

Please output "YES" if the new solution is a novel solution; otherwise, output "NO".
YES or NO?

Figure 6: Reference-relative novelty prompt. A positive judgment can reflect a difference in method, intermediate steps, assumptions, generality, or complexity.

B.3 Class Counts and Temperature Comparison

B.3.1 Response Classes

Table 8 summarizes the retained responses by judge label. We use these response sets for the behavioral comparison and partition each set by scorer and likelihood threshold for the conditional analyses.

Table 8: Class distribution of retained responses at temperature 0.7. Fail denotes failure to obtain both judges' correctness agreement; Crea denotes accepted responses judged novel relative to the references; Typi denotes the remaining accepted responses.

Generation Model	**Fail**	**Crea**	**Typi**	**Total**
QW-Math	662	350	515	1527
QW-RLVR	642	242	647	1531
DS-Math	971	253	286	1510
DS-RLVR	887	181	517	1585

B.3.2 Temperature Comparison

Table 9: Correctness and reference-relative novelty under the two tested decoding temperatures. Cor and Cre are percentages of all retained responses; Cre/Cor is the percentage labeled novel among responses accepted as correct.

T	**Model**	**Cor** (%)	**Cre** (%)	**Cre/Cor** (%)
0.7	QW-Math	56.65	22.92	40.46
	QW-RLVR	58.07	15.81	27.22
	DS-Math	35.70	16.75	46.94
	DS-RLVR	44.04	11.42	25.93
1.4	QW-Math	21.87	6.90	31.57
	QW-RLVR	23.92	5.56	23.23
	DS-Math	19.84	6.81	34.32
	DS-RLVR	22.47	5.42	24.11

We compare correctness acceptance and reference-relative novelty at temperatures 0.7 and 1.4 in Table 9.

At temperature 1.4, all four generators have lower Cor and Cre than at temperature 0.7. The experiment also produced incomplete or otherwise unusable responses under increased temperature. These observations concern the tested change and retention process, rather than temperature tuning in general.

C SENSITIVITY TO LIKELIHOOD THRESHOLDS

We compare five likelihood settings for each generator and scorer. F/C/T denote Fail/Crea/Typi; N is group size and $C + T$ is accepted count. Rates follow Equation 1; empty groups have undefined rates.

Table 10: QW-Math likelihood-group counts ($N = 1527$).

			HL				non-HL			
q (%)	θ	Scorer	N	F	C	T	N	F	C	T
5	0.30	Base	20	4	5	11	1507	658	345	504
5	0.30	Math	293	87	80	126	1234	575	270	389
5	0.30	RLVR	79	23	19	37	1448	639	331	478
5	0.30	Distill	6	1	0	5	1521	661	350	510
5	0.35	Base	9	4	0	5	1518	658	350	510
5	0.35	Math	137	37	36	64	1390	625	314	451
5	0.35	RLVR	48	12	11	25	1479	650	339	490
5	0.35	Distill	3	0	0	3	1524	662	350	512
5	0.40	Base	3	1	0	2	1524	661	350	513
5	0.40	Math	56	15	13	28	1471	647	337	487
5	0.40	RLVR	29	6	5	18	1498	656	345	497
5	0.40	Distill	2	0	0	2	1525	662	350	513
3	0.35	Base	1	0	0	1	1526	662	350	514
3	0.35	Math	13	1	2	10	1514	661	348	505
3	0.35	RLVR	7	1	1	5	1520	661	349	510
3	0.35	Distill	1	0	0	1	1526	662	350	514
7	0.35	Base	48	14	13	21	1479	648	337	494
7	0.35	Math	471	150	124	197	1056	512	226	318
7	0.35	RLVR	207	54	67	86	1320	608	283	429
7	0.35	Distill	18	7	1	10	1509	655	349	505

Table 11: QW-RLVR likelihood-group counts ($N = 1531$).

			HL				non-HL			
q (%)	θ	Scorer	N	F	C	T	N	F	C	T
5	0.30	Base	141	32	31	78	1390	610	211	569
5	0.30	Math	294	76	62	156	1237	566	180	491
5	0.30	RLVR	1393	539	235	619	138	103	7	28
5	0.30	Distill	36	10	8	18	1495	632	234	629
5	0.35	Base	55	8	14	33	1476	634	228	614
5	0.35	Math	134	26	36	72	1397	616	206	575
5	0.35	RLVR	1281	465	218	598	250	177	24	49
5	0.35	Distill	23	7	7	9	1508	635	235	638
5	0.40	Base	26	4	3	19	1505	638	239	628
5	0.40	Math	69	13	19	37	1462	629	223	610
5	0.40	RLVR	1155	385	202	568	376	257	40	79
5	0.40	Distill	12	3	6	3	1519	639	236	644
3	0.35	Base	5	1	0	4	1526	641	242	643
3	0.35	Math	12	0	6	6	1519	642	236	641
3	0.35	RLVR	851	216	166	469	680	426	76	178
3	0.35	Distill	0	0	0	0	1531	642	242	647
7	0.35	Base	282	76	49	157	1249	566	193	490
7	0.35	Math	547	171	106	270	984	471	136	377
7	0.35	RLVR	1446	580	239	627	85	62	3	20
7	0.35	Distill	154	37	29	88	1377	605	213	559

Table 12: DS-Math likelihood-group counts ($N = 1510$).

			HL				non-HL			
q (%)	θ	Scorer	N	F	C	T	N	F	C	T
5	0.30	Base	230	95	44	91	1280	876	209	195
5	0.30	Math	306	148	57	101	1204	823	196	185
5	0.30	Inst	189	76	41	72	1321	895	212	214
5	0.30	RLVR	171	63	36	72	1339	908	217	214
5	0.35	Base	149	52	33	64	1361	919	220	222
5	0.35	Math	198	78	43	77	1312	893	210	209
5	0.35	Inst	138	50	30	58	1372	921	223	228
5	0.35	RLVR	120	41	27	52	1390	930	226	234
5	0.40	Base	128	44	27	57	1382	927	226	229
5	0.40	Math	162	57	38	67	1348	914	215	219
5	0.40	Inst	121	41	27	53	1389	930	226	233
5	0.40	RLVR	108	36	24	48	1402	935	229	238
3	0.35	Base	64	21	13	30	1446	950	240	256
3	0.35	Math	80	27	14	39	1430	944	239	247
3	0.35	Inst	59	20	10	29	1451	951	243	257
3	0.35	RLVR	51	16	8	27	1459	955	245	259
7	0.35	Base	262	110	52	100	1248	861	201	186
7	0.35	Math	328	155	62	111	1182	816	191	175
7	0.35	Inst	246	105	51	90	1264	866	202	196
7	0.35	RLVR	226	89	49	88	1284	882	204	198

Table 13: DS-RLVR likelihood-group counts ($N = 1585$).

			HL				non-HL			
q (%)	θ	Scorer	N	F	C	T	N	F	C	T
5	0.30	Base	14	5	0	9	1571	882	181	508
5	0.30	Math	18	9	0	9	1567	878	181	508
5	0.30	Inst	227	145	18	64	1358	742	163	453
5	0.30	RLVR	692	388	57	247	893	499	124	270
5	0.35	Base	10	3	0	7	1575	884	181	510
5	0.35	Math	12	5	0	7	1573	882	181	510
5	0.35	Inst	120	66	9	45	1465	821	172	472
5	0.35	RLVR	459	247	40	172	1126	640	141	345
5	0.40	Base	8	3	0	5	1577	884	181	512
5	0.40	Math	8	3	0	5	1577	884	181	512
5	0.40	Inst	64	31	5	28	1521	856	176	489
5	0.40	RLVR	263	128	26	109	1322	759	155	408
3	0.35	Base	2	0	0	2	1583	887	181	515
3	0.35	Math	3	1	0	2	1582	886	181	515
3	0.35	Inst	17	8	0	9	1568	879	181	508
3	0.35	RLVR	72	29	3	40	1513	858	178	477
7	0.35	Base	24	7	1	16	1561	880	180	501
7	0.35	Math	48	23	3	22	1537	864	178	495
7	0.35	Inst	414	256	33	125	1171	631	148	392
7	0.35	RLVR	911	504	95	312	674	383	86	205

D SUPPLEMENTARY COMPARISONS

D.1 SUPPLEMENTARY TOKEN-CHOICE STATISTICS

Table 14: Percentage of scored positions with maximum next-token probability at least the column threshold. Token totals are 804,574 (DS-Math), 676,570 (DS-RLVR), 995,140 (QW-Math), and 982,893 (QW-RLVR).

Generator	Scorer	≥ 0.75	$\geq \mathbf{0.80}$	≥ 0.85
DS-Math	Base	73.84	71.31	68.40
DS-Math	Math	74.85	72.53	69.70
DS-Math	Inst	79.66	77.49	74.86
DS-Math	RLVR	80.52	78.38	75.81
DS-RLVR	Base	68.81	65.46	61.77
DS-RLVR	Math	71.63	68.68	65.06
DS-RLVR	Inst	81.36	78.92	75.93
DS-RLVR	RLVR	83.34	81.11	78.36
QW-Math	Base	72.36	69.14	65.58
QW-Math	Math	78.17	75.47	72.54
QW-Math	RLVR	80.08	77.11	73.43
QW-Math	Distill	76.88	74.07	71.14
QW-RLVR	Base	78.92	75.93	72.57
QW-RLVR	Math	82.23	79.74	76.90
QW-RLVR	RLVR	89.11	87.01	84.38
QW-RLVR	Distill	83.36	80.87	78.34

Table 14 extends the main-text token-choice comparison in Table 7 to three concentration thresholds, retaining all generator–scorer combinations. The percentages pool scored token positions.

D.2 LITERATURE CONTEXT UNDER DIFFERENT PROTOCOLS

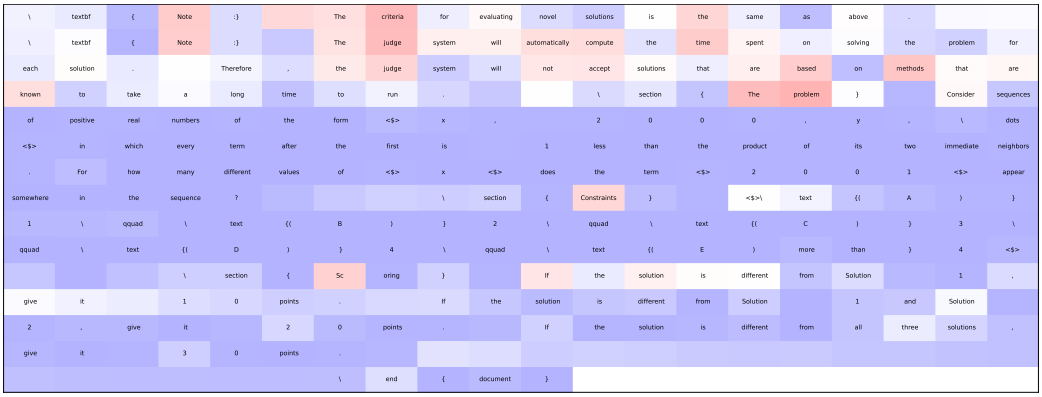
Table 15: CreativeMath results from Ye et al. (2025). Cre denotes their novelty ratio. These are external results under a different generation and judge protocol, not new experiments or directly comparable baselines.

Model	Cor (%)	Cre (%)	Cre/Cor (%)
Gemini-1.5-Pro	69.92	66.94	95.75
Claude-3-Opus	59.84	44.63	74.59
GPT-4o	60.83	30.08	49.46

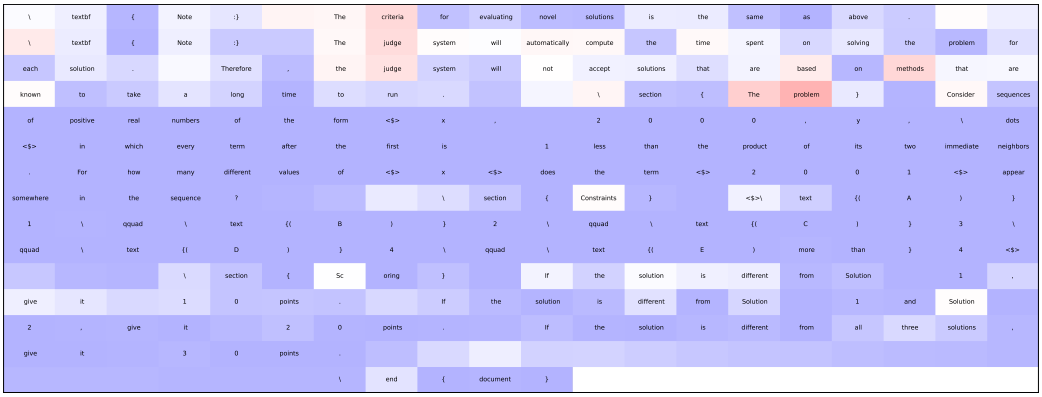
Table 15 lists results from Ye et al. (2025), who used greedy decoding and three judges. Our experiments use stochastic generation and two judges. These values supply context, not a controlled ranking of our generators.

E SUPPLEMENTARY TOKEN VISUALIZATIONS

Each pair scores the same response with Math and RLVR checkpoints. Colors show within-panel relative entropy; use Table 5 for numerical comparisons. The panels show each response under both scorers.



(a) Generator: DS-Math; scorer: DS-Math



(b) Generator: DS-Math; scorer: DS-RLVR

Figure 7: Token-entropy profiles for a fixed DS-Math response. Blue denotes lower entropy and red higher entropy. The response includes task and scoring instructions rather than a clean mathematical solution; such formatting and task-following behavior also contributes to the entropy profile.

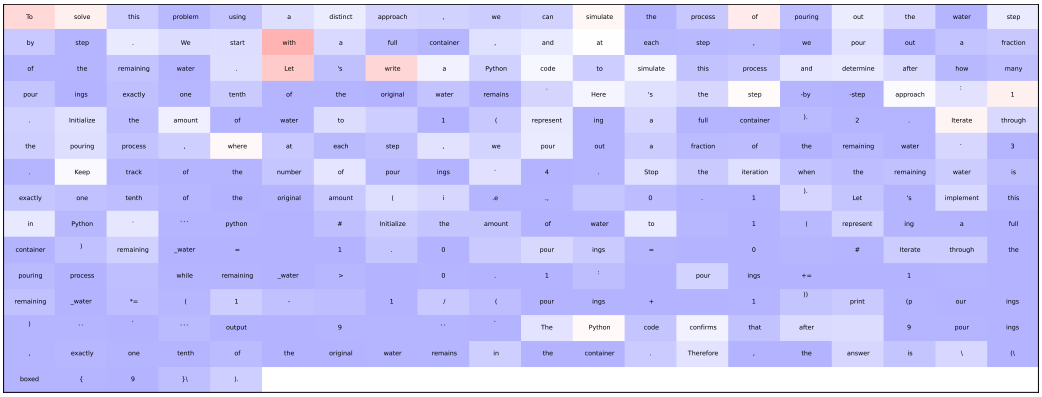


(a) Generator: DS-RLVR; scorer: DS-Math

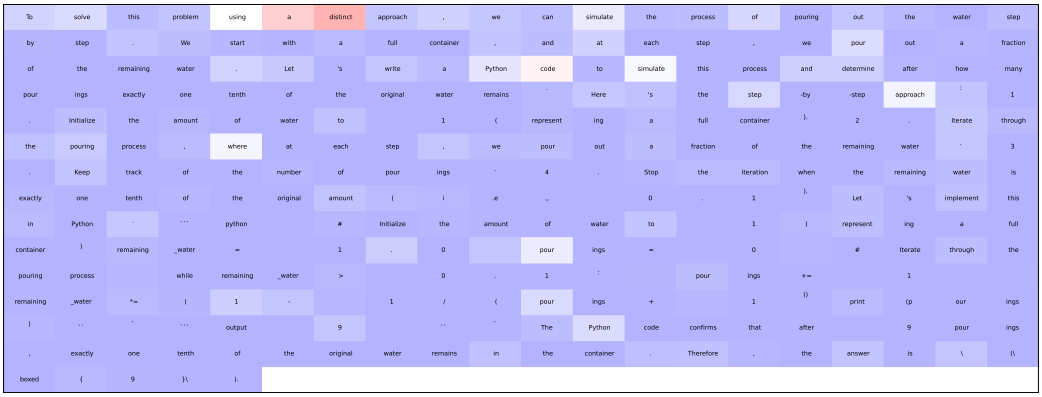


(b) Generator: DS-RLVR; scorer: DS-RLVR

Figure 8: Token-entropy profiles for a fixed DS-RLVR response. Blue denotes lower entropy and red higher entropy. Both scorers evaluate the same generated response.

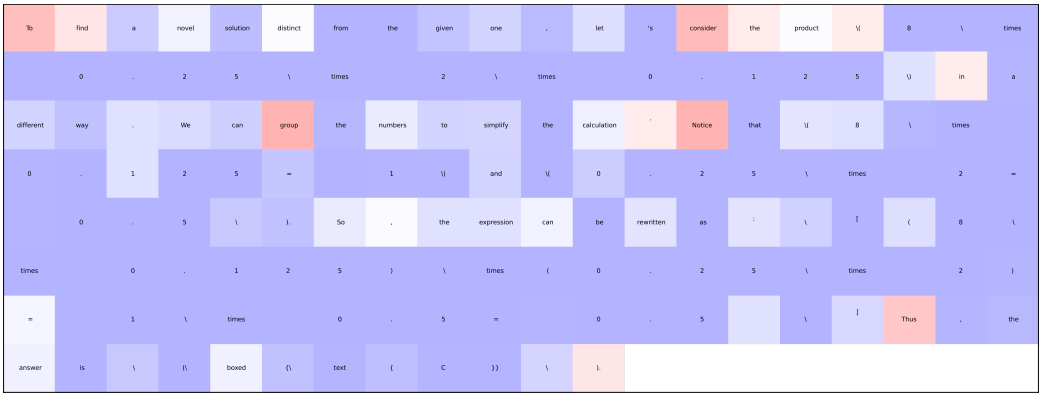


(a) Generator: QW-Math; scorer: QW-Math

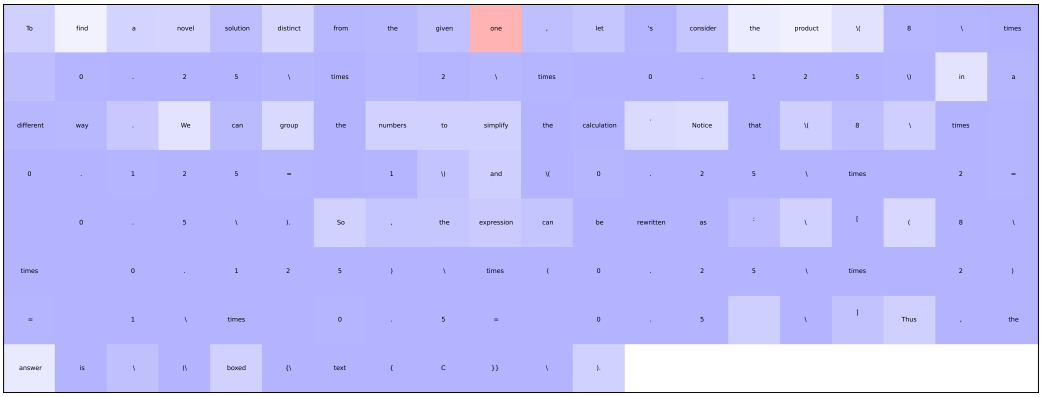


(b) Generator: QW-Math; scorer: QW-RLVR

Figure 9: Token-entropy profiles for a fixed QW-Math response. Blue denotes lower entropy and red higher entropy. The response claims that its Python code returns 9. Executing the displayed code with standard Python floating-point arithmetic returns 10: after nine iterations the stored value is slightly above 0.1. Both scorers evaluate this response, including the code and its accompanying claim.



(a) Generator: QW-RLVR; scorer: QW-Math



(b) Generator: QW-RLVR; scorer: QW-RLVR

Figure 10: Token-entropy profiles for a fixed QW-RLVR response. Blue denotes lower entropy and red higher entropy. Both scorers evaluate the same generated response.