When Does Verifier-Free Reinforcement Learning Improve Math Reasoning? Answering Rate Gates It, and Base Ability Appears to Decide the Rest

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

Reinforcement learning without a ground-truth verifier, using self-consensus signals such as majority vote, improves some base language models on math reasoning and does nothing for others. The common explanation attributes the difference to base math ability. We show instead that the outcome is governed by two *separable* properties of the base model: its parseable-answering rate (how often it emits a template-conformant final answer) and its underlying solution ability. Across three base models on GSM8K, answering rate orders base accuracy, and across these three bases the payoff of reinforcement learning is associated with base ability rather than answering rate. We then isolate the mechanism with two controlled experiments. First, on a base that rarely emits a parseable answer, no reward moves it: a majority-vote reward, a ground-truth verifier, and a purely random reward all leave greedy accuracy unchanged, so the answering-rate gate sits upstream of the reward signal itself. Second, a format prime that raises answering rate lets us intervene on the gate; opening it is not sufficient: reinforcement learning after priming yields a small, single-seed improvement on a capable base, and on a low-ability base it drives answering rate toward one without improving accuracy, reinforcing agreement on answers that are usually wrong. Verifier-free RL therefore appears to require both an open answering-rate gate and sufficient base ability; the gate is necessary but not sufficient, and self-consensus reward optimizes answering, not correctness. All results are from a single seed at 200 evaluation problems; we rest the conclusions on the size of the qualitative effects rather than on replication.

1 Introduction

Reinforcement learning has become a standard tool for improving the mathematical reasoning of large language models [1]. A recent and practically attractive variant removes the ground-truth verifier and rewards the model against a label-free signal derived from its own samples, such as agreement with the majority answer or self-certainty [2, 3, 4]. This verifier-free recipe is appealing because it needs no answer keys, yet its behavior is puzzling: it produces sizable gains on some base models, notably the Qwen family, and fails or collapses on others such as Llama and OLMo [4, 5]. A common reading attributes this to differences in base math ability.

*Corresponding author: deep@lab.cloud

We argue that this reading conflates two properties that can be measured separately, and one of which can be manipulated. The first is the base model's *parseable-answering rate* (for brevity, *answering rate*): how often it produces a template-conformant final answer that a label-free objective can cluster, vote over, or score. The second is its *solution ability*: whether the answers it does produce tend to be correct. In naturally occurring base models these two are correlated, which is why answering rate looks like a proxy for ability. By separating them we obtain a more precise account of when verifier-free RL works.

A label-free objective has nothing well defined to reward if the base rarely produces a clean final answer, so a low answering rate should block the method regardless of the reward's quality. Raising the answering rate should therefore be a prerequisite for the method to have any effect. Whether raising it is *sufficient* is a separate question, and one the literature has posed but not resolved with a controlled intervention that changes answering rate while tracking what else moves [5, 6].

We study three base models (Qwen2.5-3B [7], Llama-3.2-3B [8], and OLMo-2-1124-7B [9], hereafter OLMo-2-7B) on GSM8K [10], measuring greedy pass@1 and parseable-answering rate. Our contributions are:

1. **Answering rate orders base accuracy and pass@8** monotonically across the three models, motivating it as the organizing variable for the reinforcement-learning behavior that follows (§5.1).

2. **A closed answering-rate gate leaves the reward channel without traction.** On the low-answering base, a majority-vote reward, a ground-truth verifier, and a random reward all leave greedy accuracy unchanged (a one-sided 95% bound below about 1.5%). A ground-truth reward does no better than a random one, so the gate sits upstream of the reward (§5.2).

3. **Answering rate is a manipulable variable, but opening the gate is not sufficient, and the payoff is associated with ability.** A format prime raises answering rate; reinforcement learning after priming gives a small directional improvement on the capable base and, on the low-ability base, maximizes answering without improving accuracy. Across the three bases the payoff is ordered by base ability, not by the now-open gate, a correlational pattern over three models (§5.3).

4. **We characterize the self-consensus behavior.** Majority-vote reinforcement learning optimizes agreement, not correctness: it drives answering rate toward one and, when the base's modal answer is wrong, entrenches that answer (§5.3).

2 Related Work

Verifier-free and self-consensus RL. A growing family of methods replaces the verifier with a signal computed from the model's own outputs: majority vote and test-time RL [2, 11], self-certainty [3, 12], semantic-entropy majority [4], and entropy minimization [13]. These works report gains on strong bases and document failures on weaker ones. We provide a controlled account of the boundary between the two regimes.

The base model, not the optimizer, decides the outcome. Gandhi et al. [6] show on Countdown that Qwen2.5-3B reaches roughly 60% under RL while Llama-3.2-3B plateaus near 30%, and that priming Llama with reasoning *behavior*, even from incorrect traces, matches Qwen's

trajectory, concluding that cognitive behaviors rather than correctness gate RL. Related pre-RL interventions rescue weak bases through behavior injection [14], structured templates [15], and math-focused mid-training [16]. Our prime is deliberately narrower: it teaches only the answer *format*, not reasoning behavior. This distinction matters for interpreting the contrast. Because a behavior prime transfers reasoning structure and not only a final-answer template, it plausibly supplies solution ability that a bare format prime does not; our narrower intervention isolates the format component, and its failure to rescue a weak base is consistent with ability, which behavior priming supplies there, being the ingredient that format alone does not provide here.

Answering rate as a hidden driver. Liu et al. [17] observe that a Qwen base answers a math question in a usable format almost always, whereas some other bases answer only a small fraction of the time, a pure format artifact independent of skill. Sun et al. [18] name an answering-rate-like quantity as a hidden driver of RL-friendliness, and Zhao et al. [19] show that RL amplifies the format distribution already present in pretraining. Where this line of work identifies answering rate as a correlate of RL-friendliness, we turn it into a variable we intervene on, and we measure both the answering rate and the accuracy the intervention produces rather than assuming ability is unchanged; this lets us separate the gate from ability instead of observing them jointly.

Necessity versus sufficiency, and the verifier boundary. Roy et al. [5] argue that verifier-free RL fails on weak bases because too few correct rollouts exist for a majority to form, suggesting format is necessary but not sufficient. Hu et al. [20] show that with a *ground-truth* reward a base can adopt correct format without priming, which distinguishes the verifier-free setting where no such anchor exists. Our reward-regime sweep and format-prime intervention address both points directly.

Contamination and measurement. Reported verifier-free gains can reflect memorized data rather than reasoning; gains can shrink on clean synthetic sets [21] or under symbolic variation [22], and spurious rewards can appear to help through pretraining priors [23]. Measurement of exploration through pass@k should be treated as a diagnostic rather than an objective [24]. We adopt fixed sampling budgets and report pass@k only as a diagnostic of latent ability.

3 Method

Optimization. We train each base with Group Relative Policy Optimization (GRPO), a critic-free policy-gradient method that normalizes the reward of each sampled completion against the mean reward of its group. We use a group size of 8 and a learning rate of 1×10^{-6}. Prompts ask the model to reason step by step and place the final answer inside a boxed template; the parseable-answering rate is the fraction of completions that emit such a boxed final answer (a `####`-delimited final answer is also accepted).

Reward regimes. To separate the effect of the answering-rate gate from the effect of reward quality, we sweep three rewards over a fixed setup. *Majority vote* (the verifier-free method under study) rewards a completion for agreeing with the modal answer among the group's samples. *Verifier* rewards a completion if its answer matches the ground truth; this is an upper reference that uses label information. *Spurious* assigns a uniformly random binary reward and is a floor that carries

Table 1: Base characterization on GSM8K (greedy, $n = 200$). Answering rate orders both pass@1 and pass@8. Median generation length is in characters.

Base	Answering rate	pass@1	pass@8	Gen. length
Qwen2.5-3B	0.95	0.80	0.96	1057
OLMo-2-7B	0.48	0.43	0.77	1599
Llama-3.2-3B	0.01	0.00	0.14	231

no information. Comparing the three isolates how much of any effect is attributable to the reward being correct rather than merely present.

Format prime. To manipulate answering rate independently of ability, we apply a lightweight supervised fine-tune before RL. The prime pairs training questions with a fixed generic scaffold that carries no problem-specific reasoning and ends with the real final answer of that training question in the boxed template. Because the scaffold contains no derivation, the prime teaches the *habit* of emitting a boxed final answer rather than how to solve a problem. The answers used in the prime come from a pool of training questions disjoint from the evaluation set, so no evaluation answer is leaked. We measure greedy accuracy before and after the prime to observe what the prime changes.

4 Experimental Setup

Models and data. We use three base (non-instruct) models: Qwen2.5-3B, Llama-3.2-3B, and OLMo-2-7B. Training and evaluation use GSM8K; the RL policy is trained on a 512-prompt subsample of the train split, and all accuracy and answering-rate numbers are measured on 200 problems from the test split, which is disjoint from both the RL-train subsample and the 300-example format-prime pool.

Metrics and protocol. We report greedy pass@1 and parseable-answering rate, and pass@8 as a diagnostic of latent ability. Generation at evaluation is capped at 640 tokens. With 200 evaluation problems, the binomial standard error on pass@1 is approximately 0.035 near $p = 0.5$, so we treat differences smaller than roughly this size as within noise. All runs use a single seed (1401); we do not report multi-seed error bars, and we mark the small positive effects below as directional accordingly. Reward-regime runs use 300 GRPO steps and format-prime runs use 500 steps.

5 Results

5.1 Answering rate orders accuracy and RL-friendliness

Table 1 reports the three bases. Answering rate spans two orders of magnitude, from 0.95 for Qwen to 0.01 for Llama, and it orders both greedy pass@1 and pass@8 monotonically. Llama's near-zero pass@1 is not primarily a math-ability failure: its pass@8 of 0.14 shows it can occasionally reach a correct answer when it does commit one, and its short median generation length is consistent with a model that trails off without emitting a final answer. OLMo sits between the two on every axis. Answering rate is thus the natural variable along which to organize the RL behavior that follows.

Table 2: Change in greedy pass@1 after 300 GRPO steps ($n = 200$, single seed; the standard error of a before-after change is roughly 0.05, so every cell is within about one standard error of zero and none is individually significant across the nine tests). On the low-answering base every reward leaves accuracy unchanged (point estimate 0.000). No reward-quality separation (verifier vs. spurious) appears on any base.

Base	Majority vote	Verifier (ground truth)	Spurious (random)
Qwen2.5-3B	−0.030	−0.005	−0.010
OLMo-2-7B	+0.040	−0.015	0.000
Llama-3.2-3B	0.000	0.000	0.000

5.2 A closed gate makes the reward channel inert

Table 2 sweeps the three reward regimes over each base at 300 GRPO steps and reports the change in greedy pass@1 relative to the untrained base. Two facts stand out. First, on Llama, the low-answering base, all three rewards leave greedy accuracy unchanged: the point estimate is 0.000 for each, and since the untrained base is already at 0/200, a rule-of-three bound places the true effect below about 1.5% at 95% confidence. A ground-truth verifier does no better than a random reward, and neither does the verifier-free majority vote. When the base almost never emits a parseable answer, the policy receives almost no usable gradient, so the quality of the reward has nothing to act on. This is not simply the optimizer failing to run: the same optimizer applied to this same base after its gate is opened by the format prime (§5.3) moves the answering rate from 0.535 to 0.93, so the closed-gate result reflects an absence of usable signal rather than an inert optimizer. The answering-rate gate therefore sits upstream of the reward signal itself. Second, on the other two bases no reward regime moves pass@1 cleanly at this budget: every value in the table is within about one standard error of zero (the standard error of a before-after change is larger than the single-proportion 0.035, roughly 0.05), and reward quality (verifier versus spurious) does not separate. Even a ground-truth reward does not lift Qwen, which sits near a greedy ceiling. Because the sweep tests nine before-after changes with no multiple-comparison correction, no single cell is individually significant; the load-bearing result here is the qualitative one, the closed-gate row, not any positive cell. These results establish the gate as a necessary condition and show that reward correctness alone does not determine the outcome at this budget.

5.3 Opening the gate is necessary but not sufficient

The format prime lets us raise answering rate by teaching only the answer template, and then ask whether verifier-free RL succeeds; we track accuracy before and after so we can see whatever else the prime moves rather than assume it changes answering rate alone. Table 3 reports the full chain for each base: the untrained base, the primed model, and the model after majority-vote RL on top of the prime.[1]

The prime does raise answering rate on every base, and on Llama, the same low-answering base from §5.2 now viewed with its gate opened, it does so dramatically, from 0.03 to 0.535, lifting greedy pass@1 off the floor from 0.00 to 0.05 with real numeric answers. This confirms that answering rate

[1]The untrained-base columns in Table 3 come from the format-prime runs and differ slightly from Table 1, which reports a separate characterization run with a different decoding setup; both are measured on the same 200 test problems, and the small differences do not affect the within-run comparisons that carry the argument.

Table 3: Format-prime causal chain ($n = 200$, single seed). Each base is shown as untrained base, after the format prime, and after majority-vote RL on the primed model (500 steps). The prime raises answering rate on every base and reduces accuracy on the two capable bases; after the prime, RL gives a small directional gain on the able bases and, on the low-ability base, drives answering rate toward one without improving accuracy. Deltas of ± 0.02 to ± 0.05 are within the roughly 0.05 standard error of a before-after change.

	pass@1			Answering rate		
Base	Base	Primed	Post-RL	Base	Primed	Post-RL
Qwen2.5-3B	0.805	0.265	0.305	0.965	0.985	0.985
OLMo-2-7B	0.470	0.115	0.135	0.520	0.860	1.000
Llama-3.2-3B	0.000	0.050	0.010	0.030	0.535	0.930

is a manipulable variable rather than a fixed property of the base. The prime is also double edged, and it does not leave ability untouched: on the two capable bases the generic scaffold overwrites their working chains of thought and degrades accuracy sharply, from 0.805 to 0.265 for Qwen and from 0.47 to 0.115 for OLMo, while raising their answering rate. Answering habit and reasoning ability are therefore separable, moving by different, base-dependent amounts under the same intervention, and cheaply teaching format to a model that already reasons can damage the reasoning.

The central comparison is the RL step after the gate is open. On Llama, majority-vote RL drives answering rate from 0.535 to 0.93, yet pass@1 does not improve (it moves from 0.05 to 0.01, a change within noise). With the gate open but ability low, the modal answer the model agrees with is usually wrong, so optimizing agreement entrenches that answer: answering rate rises toward one while accuracy does not improve. On the two bases with more ability, RL after the prime instead nudges pass@1 up, by 0.02 on OLMo and 0.04 on Qwen, both directional at this single seed and within the same noise band. Ordered by base ability (pass@8), the RL-after-prime effect runs from the low-ability base (no gain) to the more able ones (small directional gains), and not by the now-open gate, which exceeds 0.86 after RL on all three; with three bases this ordering is correlational rather than a manipulation of ability. Opening the gate is necessary for RL to have traction but does not by itself make it help.

6 Discussion

The experiments give a compact account of verifier-free RL. A label-free objective needs parseable answers to operate on, so a low answering rate leaves the channel without traction: on the low-answering base a ground-truth verifier moves accuracy no more than a random reward, with no detectable change from either. This is the sharpest form of the gate we observe, because it holds even when the reward is a perfect ground-truth verifier, and it is not an artifact of an idle optimizer, since the same optimizer moves this base substantially once the gate is opened. Answering rate is therefore a necessary condition, and it is upstream of the reward.

It is not, however, sufficient. When we open the gate on a low-ability base with the format prime, majority-vote RL does not improve accuracy; it maximizes the quantity it actually optimizes, namely agreement with the model's own modal answer. On a base whose modal answer is usually wrong, this raises answering rate toward one without improving accuracy: the training signal improves

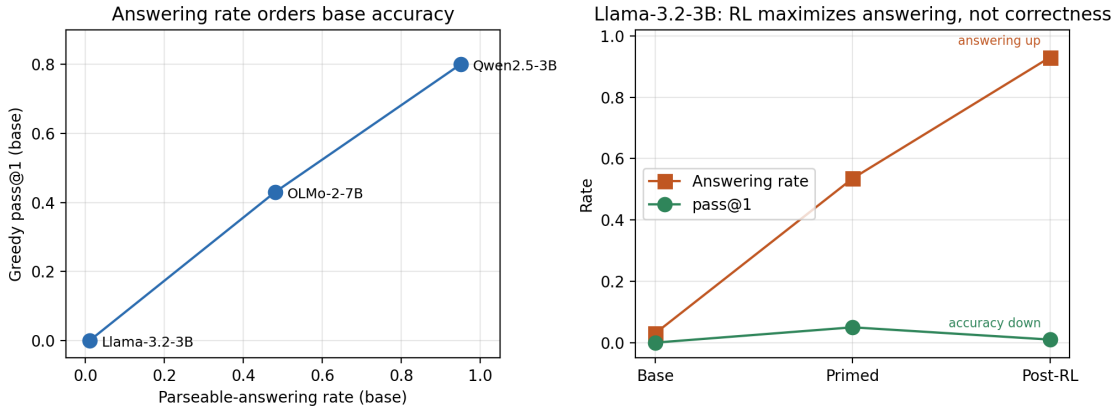

Figure 1: Left: the three measured bases, ordered by answering rate against greedy pass@1 (axes are rates in $[0, 1]$); answering rate and accuracy are correlated across these bases but are separable, as the prime shows. Right: on the low-ability base, majority-vote RL after priming drives answering rate toward one while pass@1 stays near zero, the self-consensus behavior. Values are from Tables 1 and 3.

and the surface behavior looks more decisive, yet the model is no more correct. The same RL step nudges a capable base upward instead (directionally, by 0.02 and 0.04 at a single seed), whose modal answer is more often correct. The payoff of verifier-free RL thus appears to depend jointly on an open answering-rate gate and sufficient base ability. We manipulate only the gate, through the prime; ability is not manipulated but varied by the choice of base, so the gate half of this statement is causal while the ability half is a correlational pattern across three models.

This reconciles several observations in the literature. Answering rate looks like a strong predictor of RL-friendliness because, in naturally occurring bases, answering habit and ability are correlated; our prime separates them, and across the three bases it is ability, not the now-open gate, that is associated with whether reinforcement learning pays off. It is also consistent with reports that a ground-truth reward can induce correct formatting during reinforcement learning [20], a correctness signal the verifier-free setting lacks, though even that would help only once the base answers often enough to receive any gradient at all.

7 Limitations

This study was scoped around a pre-specified quantitative test: that the format prime would recover at least half of the gap between the strong base and a weak-format base while leaving base accuracy nearly unchanged. That criterion was not met, and the pre-specified alternative, that raising answering rate is necessary but not sufficient, is the outcome we report; we present it as the finding, not as a shortfall. Our study uses a single evaluation seed (1401), so the small positive post-prime RL effects on OLMo and Qwen (+0.02 and +0.04) are directional and not significance tested; the results we emphasize are the qualitative ones (the closed-gate row, where no reward produces a detectable change, and the large answering-rate rise against a flat, near-zero accuracy on the primed low-ability base), which are not marginal. All reported numbers, including the headline rows, come from this single seed; we rest the qualitative conclusions on the size of the effects, a point-estimate zero across all three reward types and an answering-rate rise above 0.3 against a flat, near-zero

accuracy, rather than on replication. The positive post-prime effects are also measured relative to the primed model rather than the untrained base: because the prime itself degrades the capable bases, the net change from the untrained base is a large decrease, for example Qwen from 0.805 to 0.305, so the reinforcement-learning step recovers only part of the accuracy the prime removes. We evaluate on GSM8K only; difficulty strata such as MATH levels and harder competition benchmarks are not covered, and whether the gate and ability decomposition holds across difficulty is untested here. We do not run a post-cutoff clean synthetic set, so we do not separately exclude memorization as a contributor to absolute accuracy; this matters most for the ability ordering that anchors the correlational half of our account, since a contaminated benchmark could inflate a base's apparent ability, whereas the closed-gate and answering-rate results do not depend on absolute accuracy being uncontaminated. The format prime is a single realization of a broad design space; a prime that raises answering rate without damaging the reasoning of a capable base would sharpen the ability-scaling comparison. Finally, we study one optimizer (GRPO) and one verifier-free reward family (majority vote, with verifier and random controls); other label-free objectives may interact with the gate differently.

8 Availability

All artifacts, including the training and evaluation code, configurations, and the exact evaluation recipe, are available from the authors on request.

9 Conclusion and Future Work

On GSM8K, under GRPO with a majority-vote reward, the success of verifier-free RL is predicted by two separable properties of the base model: its parseable-answering rate and its solution ability. A low answering rate is a hard gate, where every reward we tried, including a ground-truth verifier, leaves the model unchanged, and raising the answering rate alone does not unlock gains. Whether these two properties generalize beyond this setting is the main open question, and the experiments that would settle it are concrete: an instruction-tuned reference that has a high answering rate with ability intact; a milder prime that raises answering rate without harming reasoning, to sharpen the ability comparison; a ground-truth-reward arm after priming, to separate the self-consensus failure from a general one; difficulty-stratified and contamination-clean evaluations; and additional seeds and label-free reward families against the same gate-and-ability decomposition.

References

[1] Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, Fangfu Liu, Xiang Xu, Jiaze Ma, Xuekai Zhu, Ermo Hua, Yihao Liu, Zonglin Li, Huayu Chen, Xiaoye Qu, Yafu Li, Weize Chen, Zhenzhao Yuan, Junqi Gao, Dong Li, Zhiyuan Ma, Ganqu Cui, Zhiyuan Liu, Biqing Qi, Ning Ding, and Bowen Zhou. A Survey of Reinforcement Learning for Large Reasoning Models. 2025. URL https://arxiv.org/abs/2509.08827.

[2] Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning

Ding, and Bowen Zhou. TTRL: Test-Time Reinforcement Learning. 2025. URL https://arxiv.org/abs/2504.16084.

[3] Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to Reason without External Rewards. 2025. URL https://arxiv.org/abs/2505.19590.

[4] Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right Question is Already Half the Answer: Fully Unsupervised LLM Reasoning Incentivization. 2025. URL https://arxiv.org/abs/2504.05812.

[5] Shuvendu Roy, Hossein Hajimirsadeghi, Mengyao Zhai, and Golnoosh Samei. You Need Reasoning to Learn Reasoning: The Limitations of Label-Free RL in Weak Base Models. 2025. URL https://arxiv.org/abs/2511.04902.

[6] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive Behaviors that Enable Self-Improving Reasoners, or, Four Habits of Highly Effective STaRs. 2025. URL https://arxiv.org/abs/2503.01307.

[7] Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 Technical Report. 2024. URL https://arxiv.org/abs/2412.15115.

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Angela Fan, et al. The Llama 3 Herd of Models. 2024. URL https://arxiv.org/abs/2407.21783.

[9] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Nathan Lambert, et al. 2 OLMo 2 Furious. 2024. URL https://arxiv.org/abs/2501.00656.

[10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. 2021. URL https://arxiv.org/abs/2110.14168.

[11] Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, and ShaoGuo Liu. ETTRL: Balancing Exploration and Exploitation in LLM Test-Time Reinforcement Learning Via Entropy Mechanism. 2025. URL https://arxiv.org/abs/2508.11356.

[12] Pengyi Li, Matvey Skripkin, Alexander Zubrey, Andrey Kuznetsov, and Ivan Oseledets. Confidence Is All You Need: Few-Shot RL Fine-Tuning of Language Models. 2025. URL https://arxiv.org/abs/2506.06395.

[13] Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing Confidence Alone Improves Reasoning. 2025. URL https://arxiv.org/abs/2505.22660.

[14] Zhepeng Cen, Yihang Yao, William Han, Zuxin Liu, and Ding Zhao. Behavior Injection: Preparing Language Models for Reinforcement Learning. 2025. URL https://arxiv.org/abs/2505.18917.

[15] Jinyang Wu, Chonghua Liao, Mingkuan Feng, Shuai Zhang, Zhengqi Wen, Haoran Luo, Ling Yang, Huazhe Xu, and Jianhua Tao. TemplateRL: Structured Template-Guided Reinforcement Learning for LLM Reasoning. 2025. URL https://arxiv.org/abs/2505.15692.

[16] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling. 2025. URL https://arxiv.org/abs/2506.20512.

[17] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like Training: A Critical Perspective. 2025. URL https://arxiv.org/abs/2503.20783.

[18] Shaoning Sun, Mingzhu Cai, Huang He, Bingjin Chen, Siqi Bao, Yujiu Yang, Hua Wu, and Haifeng Wang. Distributional Clarity: The Hidden Driver of RL-Friendliness in Large Language Models. 2026. URL https://arxiv.org/abs/2601.06911.

[19] Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. Echo Chamber: RL Post-training Amplifies Behaviors Learned in Pretraining. 2025. URL https://arxiv.org/abs/2504.07912.

[20] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-Reasoner-Zero: An Open Source Approach to Scaling Up Reinforcement Learning on the Base Model. 2025. URL https://arxiv.org/abs/2503.24290.

[21] Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, Yanwei Fu, Qin Liu, Songyang Zhang, and Qi Zhang. Reasoning or Memorization? Unreliable Results of Reinforcement Learning Due to Data Contamination. 2025. URL https://arxiv.org/abs/2507.10532.

[22] Jian Yao, Ran Cheng, and Kay Chen Tan. VAR-MATH: Probing True Mathematical Reasoning in LLMS via Symbolic Multi-Instance Benchmarks. 2025. URL https://arxiv.org/abs/2507.12885.

[23] Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious Rewards: Rethinking Training Signals in RLVR. 2025. URL https://arxiv.org/abs/2506.10947.

[24] Yang Yu. Pass@k Metric for RLVR: A Diagnostic Tool of Exploration, But Not an Objective. 2025. URL https://arxiv.org/abs/2511.16231.