

The Reassessment That Did Not Travel: What the 2020 Arcade Result About Exploration Bonuses Establishes, the Six Conditions That Made It Informative, and Which of Them the Language-Model Revival Drops

Mahendrakar, Pranay

SONYTECH

ORCID: 0009-0003-7224-029X

Life of Research

2026-09-14

Pranay Mahendrakar

Abstract

Exploration bonuses have returned. Between 2025 and 2026 at least six frameworks added an intrinsic novelty or uncertainty term to reinforcement learning with verifiable rewards for language models, each naming a classical antecedent - prediction error, pseudo-counts, epistemic uncertainty - and each reporting gains. The classical literature those antecedents come from also contains a controlled reassessment. In work published at ICLR 2020, a study held the learning algorithm fixed, tuned every bonus, and compared against plain undirected exploration across the full Atari suite; it reported that bonuses beat the simple scheme on one celebrated game, showed no visible difference from it on the rest of the designated hard-exploration set, and never beat it on games where exploration is not the bottleneck. This paper states what that reassessment establishes and, at comparable length, what it does not; extracts the six design conditions that made it informative; and audits the language-model revival against them. Sixteen papers in the revival were checked mechanically for a citation to it, and none contains one. Four of the six conditions are met by at least one paper in the revival and a fifth in part. The condition the reassessment was built to test - an evaluation arm where exploration is not the bottleneck - is met by none of them in the form it requires, although the pattern that condition exists to detect is already visible in one revival paper's own published table. This paper reports no experiments. It names two failed direct imports that the revival itself reports and does not read as evidence about transfer, states where the analogy breaks on the substrate rather than on the evidence, and specifies the comparison a 2026 survey independently asks for without knowing it has been run once already.

1. Introduction

An intrinsic reward for novelty is one of the oldest ideas in exploration and one of the most frequently re-implemented. In its prediction-error form the agent is paid for reaching states whose consequences its own forward model gets wrong (Pathak et al., 2017). In its pseudo-count form the payment is a decreasing function of a density model's estimate of how often a state has been visited (Bellemare et al., 2016; Ostrovski et al., 2017), or of a hash bucket's count (Tang et al., 2016), or of a supervised estimate derived from coin flips (Lobel et al., 2023). In its random-target form the error is measured against a fixed randomly initialised network, which removes the noisy-transition failure mode by removing the stochasticity from the prediction target (Burda et al., 2018b). The high-water mark of the argument for the idea was a study that removed the extrinsic reward entirely and reported that a purely curiosity-driven agent performs surprisingly well across 54 standard benchmark environments, including the Atari suite (Burda et al., 2018a).

The same literature contains a controlled reassessment of that idea, and the reassessment is not a footnote. Taiga et al. (2021), published as a conference paper at ICLR 2020, fixed the learning algorithm to Rainbow (Hessel et al., 2017), tuned each bonus method's hyperparameters, and ran the resulting agents against Rainbow's own undirected exploration on Montezuma's Revenge, on the set of games designated hard for exploration, and on the whole 60-game Atari suite. Their abstract states the finding: while exploration bonuses lead to higher score on Montezuma's Revenge they do not provide meaningful gains over the simpler epsilon-greedy scheme, and methods that perform best on that game often underperform epsilon-greedy on easy exploration Atari 2600 games. Their conclusion is about the field rather than about any one algorithm: recent gains on Montezuma's Revenge may be better attributed to architecture change rather than better exploration schemes, and the real pace of progress in exploration research for Atari 2600 games may have been obfuscated by good results on a single domain.

Since 2025 the mechanism has been imported into reinforcement learning with verifiable rewards for language models, and into reinforcement learning from human feedback alongside it. A 2026 survey tabulates six of the importing systems and records, for each, the classical signal it adapts: CDE from random network distillation, CD-RLHF from the intrinsic curiosity module, i-MENTOR from random network distillation, MERCI from pseudo-counts, and DACE and INTUITOR from uncertainty (Yang et al., 2026). Its prose refers to five methods where its table lists six; the table is what is counted here. Each of the six reports gains over its baseline. None of the six, and none of the further ten papers checked here, cites Taiga et al.

This paper is an audit, and it has a specific shape. It does not argue that the revival is wrong, because the evidence does not support that and because at least one 2026 result cuts the other way. It argues something narrower and, if correct, more useful: that a negative result is only as informative as the comparison that produced it, that Taiga et al.'s comparison was informative because of six identifiable design choices, and that those choices - not the headline - are the part worth transferring. Three things follow.

First, the reassessment's scope is narrower than its reputation, and stating the scope flatly matters more than repeating the headline. On Montezuma's Revenge itself the bonuses won, clearly and by a large margin. The negative finding is about what happens when the evaluation leaves that one game.

Second, the language-model revival satisfies most of the six conditions and fails the one that carries the result. Sixteen papers were checked. The learning algorithm is routinely held fixed; an undirected baseline is routinely present; bonus hyperparameters are routinely reported. What is routinely absent is an evaluation arm where exploration is not the bottleneck, run under the same protocol as the hard arm - and repeated runs with reported dispersion, which is what makes a small decline on such an arm readable at all.

Third, and this was not anticipated by the question's framing, the pattern the missing arm exists to detect is already legible in the revival's own published numbers, and the revival already contains two reported failures of direct import which it treats as motivation for new methods rather than as evidence about transfer. The most interesting material in this paper is inside the papers it is auditing.

2. Scope, terms, and what this paper is not

Exploration bonus here means an additive term in the reward or advantage that depends on a novelty, count, prediction-error or uncertainty estimate rather than on task success. *Undirected exploration* means a scheme that injects state-independent stochasticity: epsilon-greedy, Boltzmann or softmax sampling, temperature scaling, and entropy regularisation. The grouping is not this paper's invention; Zhang et al. (2025) derive in their own appendix that entropy regularisation yields a Boltzmann policy which

decomposes into a greedy action taken with probability one minus a state-dependent epsilon, and describe this as showing that softmax exploration is a generalized form of epsilon-greedy. That derivation is what licenses reading the two literatures against each other at all, and it comes from inside the revival.

The reassessment means Taiga et al. (2021), the ICLR 2020 paper, whose arXiv record identifies it as the full version of an earlier workshop paper (Taiga et al., 2019). Where the two differ in wording the full version is used and the difference is noted. The classical family as a whole, and its internal divisions by what the intrinsic signal is computed over, are surveyed by Aubret et al. (2019); that survey is the one the 2026 modularised survey cites when it traces the LLM methods back to their antecedents, and no attempt is made to re-run its taxonomy here.

Three things this paper is not. It is not a claim that the pass@k crossover measures a reasoning boundary, or that it does not; that dispute and the five premises it leaves unstated are treated in a prior record (Mahendrakar, 2026a), and the present paper cites Yue et al. (2025) only where a specific number is needed. It is not a claim about which direction of transfer between task domains RLVR supports, treated in a second prior record (Mahendrakar, 2026b). And it is not a benchmark: no experiment was run for it, no model was trained, and every number in it is read off a published table or abstract.

The absence claims in Section 7 rest on a mechanical procedure, stated here so that it can be repeated and, if it is wrong, falsified cheaply. For each named paper the arXiv HTML rendering of the full text was fetched, tags stripped, Unicode decomposed to ASCII so that the diaeresis in the first author's surname cannot hide a match, and the resulting text searched for the strings "taiga", "ali taiga", "2109.11052", "1908.02388", "bonus-based exploration", "bonus based exploration" and "arcade learning". A paper counts as not citing the reassessment when all seven counts are zero, or when the only non-zero count is resolved by reading the passage and found to be something else; one paper of the sixteen falls into the second case and is identified in Section 7 rather than folded silently into the first. The method has one known failure mode, stated here rather than in the limitations: a paper whose HTML rendering omits its bibliography would return seven zeros for the wrong reason. Every text used here was confirmed to contain a bibliography by checking that other reference strings from the same list - "burda", "pathak", "bellemare", "ostrovski" - do occur.

3. What the 2018 result established, and what its own authors said it did not

Burda et al. (2018a) is the paper the revival cites when it wants a licence for the idea, and it is worth reading for what it actually claims. Its three stated contributions are a large-scale study of purely curiosity-driven learning without any extrinsic rewards across 54 standard benchmark environments; an investigation of feature spaces, reporting that random features are sufficient for many popular RL game benchmarks while learned features appear to generalise better; and, third and last, a demonstration of limitations of the prediction-based rewards in stochastic setups.

The third contribution is the one that gets dropped in citation. The celebrated result and its own principal caveat were published in the same abstract. The caveat was then answered by a specific architectural fix - predicting the output of a fixed random network, whose target carries no environment stochasticity at all (Burda et al., 2018b) - and by episodic-memory formulations that measure novelty in steps-to-reach rather than in prediction error, which Savinov et al. (2018) introduce precisely to overcome what they call the couch-potato failure, where an agent instantly gratifies itself by exploiting actions with hardly predictable consequences. The line is still live: Hou et al. (2025) propose rewarding model improvement rather than prediction error or novelty, and report faster intrinsic-reward convergence and more states visited in a maze than baselines in noisy environments.

So the caveat was taken seriously and worked on for seven years. What was not taken seriously to the same degree is the separate question the reassessment asked, which is not whether the bonus is robust to noise but whether the bonus buys anything that a cheaper mechanism does not.

4. What the reassessment measured

Taiga et al. (2021) combined Rainbow with CTS pseudo-counts, PixelCNN pseudo-counts, the intrinsic curiosity module and random network distillation as the bonus arms, and with epsilon-greedy and NoisyNets as the baseline arms, and evaluated on three nested sets: Montezuma's Revenge; the six games Bellemare et al. classified as hard for exploration with sparse rewards, namely Gravitar, Private Eye, Venture, Solaris, Pitfall! and Freeway; and the full 60-game suite. Curves are averaged over five runs with the variance shaded.

On Montezuma's Revenge the bonuses win. The paper states that epsilon-greedy exploration performs poorly here and struggles in such a hard exploration game, that exploration bonuses have a huge impact and all eventually beat the baselines epsilon-greedy and NoisyNets, that only the intrinsic curiosity module is unable to surpass the baselines by a large margin, and that random network distillation, CTS and PixelCNN all average close to 5000 points. It adds, in passing, that a more recent bonus like RND does not lead to higher performance over the older pseudo-count method.

On the remaining hard-exploration games, with hyperparameters tuned on Montezuma, the paper reports that performance of each method on Montezuma's Revenge does not correlate with performance on other hard exploration domains, that there is no visible difference between epsilon-greedy and more sophisticated exploration bonuses, and that epsilon-greedy is on par with every other method and even outperforms them by a significant margin on Gravitar.

On the original Atari training set, four of whose five games are easy-exploration problems, the paper reports that CTS pseudo-counts appear to struggle when the environment is an easy exploration problem and end up performing worse on every game except Seaquest, and that RND and the intrinsic curiosity module are able to consistently match epsilon-greedy exploration but never beat it. It records a result that cuts against the tidy reading: despite its limited performance on Montezuma's Revenge, NoisyNets presented the most consistent improvements.

Two further arms matter. Hyperparameters for CTS and RND were re-tuned on a second training set - Seaquest, Qbert, Pong, Breakout and Asterix - and the paper reports that its conclusions remain valid under that re-tuning. And training was extended from Rainbow's 200 million frames to one billion, with none of the methods surveyed benefiting from the additional samples on the hard-exploration games.

The methodological claim the paper makes about its own field is explicit and is the sentence most worth transferring. It is standard in the exploration community to first tune hyperparameters on Montezuma's Revenge and then evaluate those hyperparameters on the remaining hard exploration games with sparse rewards; this is not in line with the original ALE guidelines, which recommend evaluating on the entire suite (Machado et al., 2017); and the paper supplies empirical evidence that failing to follow the guidelines may interfere with the conclusions of the evaluation.

5. The six conditions that made the reassessment informative

The reassessment is worth more than its headline because of how it was built. Six choices are separable, and each can be checked against any later comparison.

C1. The learning algorithm is held fixed across arms. Only the exploration mechanism varies. Without this, an improvement attributed to a bonus can be an improvement in the optimiser, which is exactly the confound the paper's closing sentence names when it suggests gains may be better attributed

to architecture change.

C2. An undirected baseline is run as its own arm, under the same budget. Rainbow's own schedule, with epsilon decaying from 1 to 0.01 over one million frames, is a full arm rather than an ablation footnote. Note a detail that cuts against the loose reading of this paper and in favour of its conclusion: the bonus arms are tuned and the baselines are not. Each bonus method received a reported hyperparameter sweep; the epsilon-greedy arm ran at Rainbow's standard schedule and the NoisyNets arm, the paper states, was not tuned at all. The comparison is therefore between a tuned bonus and two default simple schemes, and the tuned bonus still did not pay for itself off Montezuma. A reader who wants the opposite conclusion should note that this asymmetry runs the other way for anyone arguing that bonuses help: it makes the reassessment's negative finding harder to dismiss, not easier.

C3. Bonus hyperparameters are tuned, the tuning set is disclosed, and the tuning is repeated on a second set. The paper reports its sweeps by name and range, states that all bonus-based methods were tuned with respect to final performance on Montezuma's Revenge, and then repeats the tuning for two of them on five other games so that the choice of tuning set cannot be the whole story.

C4. Evaluation includes domains where exploration is not the bottleneck. This is the condition that carries the result. The hard-exploration set was chosen to highlight the benefits of exploration methods and, as the paper says, performance on this set may not be representative of an algorithm's capabilities on the whole suite, because exploration bonuses can negatively impact performance by skewing the reward landscape.

C5. Runs are repeated and dispersion is shown. Five runs per configuration with variance shaded. The reason this is a condition rather than a courtesy is Henderson et al. (2017): non-determinism in standard benchmark environments, combined with variance intrinsic to the methods, can make reported results tough to interpret, and without significance metrics and tighter standardisation it is difficult to determine whether improvements over the prior state of the art are meaningful.

C6. A compute-scaling arm tests whether the advantage is a budget artefact. The 200-million against one-billion frame comparison answers a specific objection - that the bonus needs longer to pay off - and answers it in the negative for the hard-exploration set.

6. What the reassessment does not establish

Being pedantic about the scope of a negative result is not a concession; it is the only way the result stays usable. Five things this one does not license.

It does not license "exploration bonuses do not work." On the single game the field cared about most, they worked, by a margin the paper describes as huge.

It does not license "only undirected exploration is worth its cost." The method the paper reports as presenting the most consistent improvements across the suite is NoisyNets, whose perturbations are learned rather than state-independent. The paper treats NoisyNets as a baseline alongside epsilon-greedy rather than as one of the bonuses under test, and states that it did not tune it, keeping the original hyperparameter. So the most consistent improver in the study was an untuned non-bonus mechanism that lost on Montezuma. The finding is about which mechanism is worth its cost on which distribution of tasks, not about a category.

It does not license conclusions about any bonus not tested. CTS pseudo-counts, PixelCNN pseudo-counts, ICM and RND were tested, against epsilon-greedy and NoisyNets as baselines. Later pseudo-count estimators were not; Lobel et al. (2023) report that their coin-flip estimator is significantly more effective at deducing ground-truth visitation counts than previous work and outperforms existing approaches on most of nine challenging exploration tasks, and that claim is not in the reassessment's

scope. Nor does it license conclusions about non-bonus approaches to the same problem. Ecoffet et al. (2019) had already reported a mean of over 43,000 points on Montezuma's Revenge from an archive-and-return method, against the roughly 5,000 the reassessment measures for the bonuses it tested, and stated in their own abstract that current algorithms perform poorly on hard-exploration domains even those with intrinsic motivation, which was the dominant method there. The two papers are consistent and say different things: one reports that the bonus family was not the frontier, the other that inside the family the differences do not survive leaving one game.

It does not license conclusions outside the Arcade Learning Environment with Rainbow. The paper's final sentence scopes itself to exploration research for Atari 2600 games. A second controlled study, on MiniGrid with four intrinsic rewards at different diversity levels, reaches a differently shaped conclusion: Kayal et al. (2025) report that state counting gives the best exploration performance with low-dimensional observations, that its performance is highly degraded with RGB observations mostly due to representation learning challenges, that maximum entropy is less impacted and therefore more robust despite not always being optimal, and that learning diverse skills with DIAYN does not promote exploration in their environments. Two things about that study matter here, and the second was not what this paper expected to find. It is not a second run of the same comparison: its arms are four intrinsic rewards, chosen to span state, policy and skill-level diversity, rather than a bonus family measured against a simple alternative, and it does not report the easy-domain arm that C4 describes. But one of its four arms, maximum entropy, is an undirected scheme by the grouping adopted in Section 2 - Kayal et al. classify it as an intrinsic reward imposing policy-level diversity, so the reclassification is this paper's and not theirs - and it is the arm they report as the robust one. That is the same shape as NoisyNets in the reassessment: the simplest mechanism in the study is not the peak performer and is the one whose advantage survives a change of setting. Between them the two studies support a claim weaker than either headline and more defensible than both: which exploration mechanism wins has been shown to depend on the observation space and the task distribution, and has not been shown to be a property of the mechanism. It also follows that any description of the Arcade study as the field's only controlled comparison of intrinsic rewards is wrong.

It does not license an inference about how the field responded. The field responded. Dabney et al. (2020) open by stating that recent empirical studies suggest that, when applied to a broader set of domains, some sophisticated exploration methods are outperformed by simpler counterparts such as epsilon-greedy, and propose temporally extended epsilon-greedy as the constructive answer. The reassessment was read where it was published. The claim in this paper is about a different audience.

7. The citation check

Sixteen papers were checked by the procedure in Section 2. Six are the systems the 2026 modularised survey tabulates as intrinsic-motivation methods for RLVR or RLHF: CDE (Dai et al., 2025), CD-RLHF (Sun, H. et al., 2025), i-MENTOR, published under the name IMAGINE (Gao et al., 2025), MERCI (Zhang et al., 2025), DACE (Li, A. et al., 2025) and INTUITOR (Zhao et al., 2025). Four are adjacent exploration work in the same period: COPO (Bai et al., 2025), the entropy-advantage method of Cheng et al. (2025), PREPO (Sun, Y. et al., 2025) and the general exploratory bonus of Li, W. et al. (2025). Three are further exploration papers from the same window: weak-model prefix guidance (Shen et al., 2026), parameter-space exploration (Venkatkrishna et al., 2026), and outcome-based exploration (Song et al., 2025). Three are reviews: the systematic analysis of RLVR exploration mechanisms by Deng et al. (2025), the modularised survey by Yang et al. (2026), and the survey of RL for LLMs under data scarcity by Yu et al. (2026).

In fifteen of the sixteen, every one of the seven probe strings returns a count of zero. The sixteenth is Song et al. (2025), where "bonus-based exploration" occurs once; the passage was read, and it is a

description of the theoretical optimism literature - the sentence says that the majority of those works share the same bonus-based exploration approach as the paper's own historical exploration method - rather than a reference to the reassessment. Its counts for "taiga", for both arXiv identifiers and for "arcade learning" are zero. Every one of the sixteen contains at least one of "burda", "pathak", "bellemare" or "ostrovski", so the bibliographies were present in the fetched text. The classical positive results are imported; the classical reassessment of those results is not.

Two observations qualify this before it is over-read. First, a zero count is evidence about a reference list, not about what an author knows; the reassessment may have been read and judged inapplicable, and Section 12 gives reasons why a careful reader might judge exactly that. Second, the count is a snapshot of September 2026 over a literature that turns over monthly, and a single paper posted last month would change it.

What can be said flatly is narrower and still worth saying. The dismissal of the simple alternative in this literature is made on grounds that do not include the one direct empirical test of it. Dai et al. (2025) write that simple heuristics including entropy bonuses and epsilon-greedy policies are either provably suboptimal in theory, citing a guarantee result, or demonstrate debatable effectiveness in complex environments like Deep RL, citing the large-scale study of on-policy design choices by Andrychowicz et al. (2020), and in LLM-based reasoning. Neither citation is the comparison Taiga et al. ran. A theoretical suboptimality result and an empirical finding that a tuned bonus does not beat an untuned default are different kinds of claim, and only the second bears on whether to run the arm.

Zhang et al. (2025) go further in the same direction. Their introduction states that in notoriously difficult exploration tasks like the video game Montezuma's Revenge these shallow methods fail because the chance of discovering long, precise action sequences needed for reward is vanishingly small, and that in contrast deep exploration strategies are both theoretically and empirically superior in such scenarios. The first clause is what the reassessment confirms. The second clause, in the scope "such scenarios," is what the reassessment reports it could not reproduce once "such scenarios" meant more than one game.

8. What the revival actually compares

The audit of the revival against C1 to C6 is done here on the two papers whose full text and configuration tables were read line by line, CDE and MERCI, with two further papers brought in under C5 where the condition is decided by what their experimental sections report. Generalising past those is not warranted and is not attempted.

C1, fixed learning algorithm: met. CDE trains within the verl framework on Qwen3-4B-Base, sharing one GRPO configuration across all GRPO-based arms and one PPO configuration across all PPO-based arms. MERCI trains within verL on Qwen2.5-Math-7B for mathematics and Llama-3.1-8B-Instruct for SQL, on vanilla GRPO and on DAPO (Yu et al., 2025). In both, the bonus is the thing that varies.

C2, undirected baseline as its own arm: met in form, not in the form that tests anything. Both compare against the unaugmented optimiser, which samples at temperature 1.0 during rollout in both cases. By the derivation in MERCI's own appendix that baseline is in the epsilon-greedy family, which makes it the right kind of arm. What is missing is a sweep of its exploration knob. Taiga et al. could leave epsilon at its default because they were arguing that even an untuned default is hard to beat; a paper arguing the opposite direction needs the knob swept, because the claim at stake is that the bonus supplies something temperature and entropy coefficient cannot. MERCI's inclusion of DAPO is the strongest instance of this condition located here: DAPO removes the KL penalty, introduces a clip-higher strategy and dynamic sampling, and is therefore a deliberately exploration-tuned simple baseline rather than a default. MERCI reports gains over it.

C3, tuned bonus with a disclosed tuning set: partly met. Both report bonus hyperparameters - CDE reports its bonus coefficients by name together with a decay schedule, and MERCI reports its coin-flip dimension, its bonus-filtering percentile and its integration weights - and CDE reports a decay-schedule ablation in which all decay schedules outperform the no-decay baseline. Neither text located here states the set on which those values were chosen, which is the part of C3 that does the work, since the reassessment's finding is that the answer changes with the tuning set.

C4, an arm where exploration is not the bottleneck: not met. CDE evaluates on MATH, AMC23, AIME24 and AIME25, and describes all four as challenging. MERCI evaluates on AIME 2024, AIME 2025, Minerva, MATH500, OlympiadBench and College Math, and describes the set as challenging mathematical reasoning benchmarks. The spread in difficulty inside MERCI's set is real and Section 9 uses it. But no paper located here designates an arm as easy-exploration in advance, states what would count as a harmful effect there, and reports it as a first-class result. That is what C4 requires. Describing a benchmark as easier after the fact is a different thing from running the arm.

C5, repeated runs with dispersion: not met in CDE or MERCI. Neither paper's configuration table states a number of random seeds, and the string "seed" does not occur in either fetched text. Both report point estimates. Two papers in the same literature do meet this condition, and the contrast is instructive: Song et al. (2025), whose method is itself a bonus, report random seeds throughout and use error bars, and Venkatkrishna et al. (2026) report mean and standard error for pass@1 across eight rollouts.

C6, a compute-scaling arm: partly met by MERCI. Its appendix scales the vanilla GRPO baseline by continuing training to the 260th step and reports that extended training increases mean@k but substantially degrades pass@k, while its own method does not. That is the right shape of control - it asks whether the baseline simply needed longer - and it is the only instance of C6 located here.

9. The pattern C4 exists to detect is already in the revival's tables

MERCI's main mathematics table reports pass@k and mean@k for the base model, for GRPO and DAPO, and for each with MERCI added. Read along the difficulty axis, the pass@k half contains the shape the reassessment named.

Against vanilla GRPO, adding MERCI moves AIME 2025 from 50.0 to 60.0 and AIME 2024 from 76.7 to 80.0, both at pass@256. On the three less extreme benchmarks it moves Minerva from 64.0 to 63.2, MATH500 from 91.8 to 91.4, and College Math from 49.2 to 48.9, all three downward. OlympiadBench rises from 59.7 to 60.9. The reported average rises from 65.8 to 67.4. Per benchmark, the two AIME suites contribute plus 10.0 and plus 3.3 while the other four contribute minus 0.8, minus 0.4, plus 1.2 and minus 0.3: the gain is concentrated in the two hardest columns and the other four sum to a wash. Against DAPO the same shape is weaker but present: Minerva falls from 66.9 to 66.5 and MATH500 from 92.0 to 91.8, while AIME 2024 rises from 76.7 to 83.3 and College Math rises from 48.3 to 50.2.

Four qualifications, each of which the fair reading requires.

The declines are small - 0.8, 0.4 and 0.3 points - and without seeds or dispersion nothing distinguishes them from noise. That is not a defence of the numbers; it is the same objection in a different place, and it is what C5 exists to prevent.

The AIME suites carry 30 problems each, a figure MERCI states itself when explaining why it evaluates them at k equal to 256. A move from 50.0 to 60.0 on 30 problems at pass@256 is a move of three problems, and pass@256 on a 30-problem set is a coverage measurement over a very small support.

The pattern is metric-dependent, and this is the strongest qualification. MERCI's mean@k half shows gains on every benchmark against both baselines, including Minerva, MATH500 and College Math. So

the decline is confined to the coverage metric and does not appear in average accuracy. Whether that is the signature of a bonus trading coverage for concentration, or is three point estimates moving inside their own noise, is exactly what C5 would decide and cannot be decided from the published table.

MERCI's authors do not hide any of this. Their own summary - that gains are most pronounced on the AIME suites, which are the most challenging, and remain robust on the other datasets - is a fair description of their table. The disagreement here is not about the description. It is that "remain robust" and "declines on three of the four non-AIME pass@k cells against vanilla GRPO" are the same table read against different prior questions, and the reassessment supplies the prior question that makes the second reading worth running as a designed arm.

CDE's table shows the same axis with a different outcome, which is worth recording because it cuts against the tidy version of this paper's thesis. Adding the perplexity bonus to GRPO raises every column, MATH included. The critic-side variant is where the mixed pattern appears: against vanilla PPO's 86.6 on MATH, the two-head variant scores 83.2 and the eight-head variant 85.1, while both improve the AIME columns. On CDE's own evidence the actor-side bonus does not show an easy-set cost and the critic-side one does, in the arm where it is cheapest to see it.

10. Two failed imports, reported inside the revival

The revival contains two reported failures of the direct classical import. Both are published as motivation for a new method. Neither is framed as evidence about whether the classical mechanism transfers, and read together they are the single most informative thing in the recent literature on that question.

Dai et al. (2025) report a preliminary experiment in which they applied count-based exploration to RLVR directly, using SimHash (Tang et al., 2016) to map the embedding of a chain-of-thought response to a discrete hash code and using bucket visitation frequency as a pseudo-count. Their reported result is that this approach proves problematic, that it is difficult to meaningfully characterise a complex reasoning trajectory with a fixed embedding vector, and that in practice most responses collapse into the same or neighbouring hash grids, producing a highly concentrated count distribution that undermines the effectiveness of count-based exploration for RLVR. A classical mechanism was tried in close to its classical form, in one implementation with one choice of embedding and hash, and is reported not to have worked in the new substrate.

Song et al. (2025) report the same kind of negative finding one level up. Their contribution list states that naively adapting UCB does not lead to improved testing performance, which is why they propose refined variants incorporating both positive and negative signals. Their diagnosis of why is a substrate argument: measuring novelty in token space naively results in exponentially many candidates and is intractable, whereas in the verifiable domain the final answer can be used as a proxy, and given a large sample budget the number of distinct answers produced by the base model averages fewer than fifty, a property they name the tractability of the outcome space.

That second paper matters for a further reason. It is the closest thing in the revival to the analysis this paper is conducting, it was written by authors with standing in the classical exploration literature, and it re-derives from first principles the reason a state-novelty bonus does not straightforwardly port. It did not cite the reassessment either. The problem is not that the classical knowledge is unavailable; it is that it is being reconstructed rather than retrieved, one group at a time, with the design lessons left behind.

11. Comparisons that do meet the conditions

Three studies in this substrate do run a comparison of something like the demanding form, and they do

not all point the same way. They are recorded here at the length they deserve, the one that cuts against this paper's thesis first.

Venkatkrishna et al. (2026) compare parameter-space exploration against three action-space exploration baselines: an entropy regularisation term, a method that raises sampling temperature when entropy collapses past a threshold, and a KL penalty on tokens with high covariance between probability and advantage. The effective group size is fixed at 16 rollouts across all methods, results are reported as pass@1 over eight rollouts with standard error, and the comparison is made at near-identical FLOPs. That is C1, C2 in its demanding form, and C5, together. Their reported finding is that all their variants outperform the action-space baselines on average, and that the undirected action-space methods lose to plain GRPO on rates of malformed and incorrect rollouts. Their method is not an exploration bonus, so this is not a test of the bonus family; it is evidence that the tuned-simple-baseline comparison is runnable in this substrate and that it can come out against the simple baseline.

Wang (2026) supplies the strongest form of the reassessment's logic in the language-model substrate, in a study of dense prediction rewards rather than novelty bonuses. Its reported design is 74 preregistered arms varying only the delivery mechanism of one fixed prediction signal under GRPO across ALFWorld, WebShop, a synthetic partially observable environment and three model scales. Two of its reported findings bear directly here: that content-free placebos as a class match or beat the gold signal at both matched seeds in the auxiliary-loss configuration, so that the auxiliary update rather than the signal's content is the regulariser; and that no reward-channel variant measurably beats its matched-normalisation baseline on the two realistic environments. A placebo arm is C2 taken to its conclusion. The caveats are substantial and are stated here rather than in a footnote: this is a single-author preprint, it concerns dense prediction rewards in agentic environments rather than curiosity bonuses in math RLVR, and its central collapse mechanism is specific to standard-deviation normalisation in all-fail groups. It is cited for the design, not as a result about the bonus family.

Schmied et al. (2025) study exploration failures in language-model decision-making and report examining both classic exploration mechanisms such as epsilon-greedy and LLM-specific approaches such as self-correction and self-consistency. Their domains are bandits and Tic-tac-toe rather than RLVR reasoning, so the transfer is partial, but the comparison class is the right one.

12. Where the analogy breaks, and it breaks on the substrate

The case against transferring the reassessment wholesale is real and should be stated by its strongest advocate rather than its critic. Four differences.

The classical setting is a long-horizon Markov decision process over environment states with a fixed transition function. A single RLVR episode is one generation terminated by a verifiable reward, which makes the learning problem closer to a contextual bandit over sequences than to a sequential control problem. Song et al.'s outcome-partitioned bandit formalisation is an explicit statement of this.

The classical premise that makes the Montezuma argument work is that the chance of a long precise action sequence arising under undirected exploration is vanishingly small. That premise assumes a policy close to uniform over actions. An RLVR policy starts from a pretrained model whose prior over sequences is enormously structured, which is why Yue et al. (2025) can report that base models achieve a higher pass@k score than their RLVR-trained descendants at large k. That result suggests the base model may already reach, at a large enough sample budget, much of what a novelty bonus would be paid to reach - though pass@k on a benchmark and coverage of the space a bonus rewards are not the same quantity, and nothing located here measures the second. Whether a bonus therefore re-ranks inside a set the model already covers, or extends it, is the open question, and it is not settled by either literature.

Novelty in the classical setting is defined over a state space that the environment supplies. In RLVR it must be defined over something the designer chooses - token, trajectory, hidden-state hash, or final answer - and the choice is load-bearing. The survey's own table groups the six systems by exactly this axis (Yang et al., 2026), and the two failed imports in Section 10 are both failures of a particular choice of that axis rather than of the bonus principle.

Finally, what counts as easy exploration differs. In Atari it is a property of the game. In RLVR it is a property of the pairing of a base model with a problem set, and it moves as base models improve. An arm that is easy-exploration for Qwen2.5-Math-7B may not be for a weaker model. This makes C4 harder to implement in the new substrate, and it is a real reason the arm is missing rather than an excuse for its absence.

13. The survey asks for an experiment that has been run once already

Yang et al. (2026) tabulate the six intrinsic-motivation methods with their classical counterparts and then write the recommendation this paper exists to answer: that the methods in their table each adapt a different classical signal - prediction error, pseudo-counts, uncertainty - but that no existing work evaluates them on a shared benchmark, and that a systematic comparison would clarify which signals are complementary and which are redundant. That sentence says "five methods" of a table with six rows, which is worth noting only because it is the sentence this paper is answering. A web search conducted for this paper located no such shared-benchmark comparison as of September 2026; that is weak evidence of absence and is offered as such.

The survey is correct that the comparison is missing and correct that it is worth running. What it does not say, because it does not cite the reassessment, is that this comparison was run once before in the previous substrate, that its answer was negative in a specific and scoped way, and that the design which made that answer informative is available off the shelf. A shared-benchmark comparison of six bonuses against each other, without C2 and without C4, would produce a ranking within the family and would not answer whether the family is worth its cost. That is the outcome the reassessment was built to prevent, and it is the outcome the survey's recommendation, taken literally, would produce.

14. What would settle the open part

Five measurements, in decreasing order of how much they would change.

An easy-exploration arm, designated in advance. Partition a problem set by base-model pass@1 measured before training, declare the high-pass@1 partition the arm where exploration is not the bottleneck, and report bonus against no-bonus on that partition as a first-class result with the same k, the same decoding parameters and the same number of seeds as the hard partition. The reassessment's finding is that this arm is where the cost of a bonus becomes visible. Nothing located here reports it.

A swept undirected baseline. Sweep rollout temperature and entropy coefficient over a grid of comparable size to the bonus-scale sweep, and report the best simple configuration rather than the default. The claim under test is that the bonus supplies something undirected stochasticity cannot; that claim is only tested against a tuned undirected arm. MERCI's DAPO comparison is the closest existing instance and covers only clip-higher and dynamic sampling. The components for a stronger arm exist: Cui et al. (2025) report an empirical relation between policy entropy and downstream performance across their runs, and propose two targeted interventions, Clip-Cov and KL-Cov, that restrict updates to high-covariance tokens. One of those is already used as a baseline by Venkatkrishna et al. (2026). What has not been located is a bonus paper that uses it as one.

Seeds and dispersion. Three or more seeds per configuration with the spread reported. Without this, none of the small easy-set declines in Section 9 is interpretable in either direction, and the argument of

this paper is as blocked by their absence as the papers' own claims are.

A compute-matched control. Bonus arms cost extra forward passes, extra parameters, or extra rollouts. Report the baseline given the same budget spent on more rollouts or longer training. MERCI's extended-training appendix is the existing template.

The signal-content control. Replace the bonus with a content-free signal of the same magnitude and variance schedule and report whether the gain survives. The placebo arms in Wang (2026) are the design; nothing in the curiosity-bonus literature located here has run it. Of the five, this is the cheapest and, if the placebo matched the bonus, the most decisive.

15. What is not known

It is not known whether exploration bonuses improve RLVR beyond what a tuned undirected scheme achieves, because the comparison has not been run in a form that would answer it.

It is not known whether the easy-set declines in Section 9 are real. The papers report point estimates without dispersion, and this paper cannot supply the missing runs.

It is not known whether a bonus in this substrate discovers reasoning the base model would not otherwise produce, or re-ranks within its existing support. Yue et al. (2025) and Wen et al. (2025) disagree about what the pass@k evidence shows on that question, and the disagreement is analysed elsewhere (Mahendrakar, 2026a); no bonus paper located here reports an experiment designed to separate the two for its own bonus.

It is not known whether the sixteen papers' authors are unaware of the reassessment or have judged it inapplicable. A reference list does not record reading, and Section 12 gives grounds on which a careful reader could reasonably have set it aside.

It is not known which definition of novelty - token, trajectory, hidden-state hash, or outcome - is the right one for this substrate. The two failed imports in Section 10 rule out two specific choices in two specific implementations and rule out nothing in general.

16. Threats to this paper's contribution

The most serious is that the audit's central claim is a claim about absence, established from reference lists over a literature that moves monthly. Sixteen papers are a sample, not a census, and the strings searched are seven. A paper posted between the search and the reading of this one could make Section 7 false. The method is stated in Section 2 precisely so that it costs a reader minutes rather than hours to check.

The second is that the paper leans hard on two tables, one of which it reads against its authors' own summary. If MERCI's three easy-set pass@k declines are noise - which is entirely possible, and which the paper says twice - then Section 9 loses its empirical leg and the argument reverts to a methodological one about what should be measured. The methodological argument survives, but it is weaker without the illustration, and it would then rest on Sections 7, 10 and 13.

The third is that Section 13's claim of novelty is bounded by a survey published three months ago. Yang et al. (2026) already identify the missing comparison and already map each revival method to its classical counterpart. What this paper adds to that is the design conditions and the scoped statement of what the earlier answer was. A reviewer could fairly say that the survey did the mapping and this paper supplied a bibliography entry. The reply is that the bibliography entry changes what the recommended experiment should look like, and Section 13 states how; whether that is enough is a judgement this paper cannot make for its reader.

The fourth is asymmetry of scrutiny. The reassessment is read here with care and its limits are stated at length, but it is one paper, five seeds, one learning algorithm and one benchmark suite, published in 2020 on hardware and at scales that no longer resemble current practice. Its own field has moved: Lobel et al. (2023) report a pseudo-count estimator that postdates it and outperforms the methods it tested. A version of this paper written by a defender of the revival would put that sentence in its abstract.

The fifth is that Section 11's strongest design, the placebo arm, comes from a single-author preprint whose subject is a different reward family. If Wang (2026) does not survive scrutiny, the fifth measurement in Section 14 loses its worked example, though not its rationale.

17. Conclusion

A negative result is a statement about a comparison. The 2020 reassessment of exploration bonuses is usually remembered, when it is remembered, as a verdict on bonuses; it is better read as a demonstration that the verdict depends on six choices about how the comparison is built - one of which the bonus papers examined here reliably make, three of which they make in part, and two of which they do not make at all. The one that matters most is the arm where exploration is not the bottleneck, because that is where a mechanism that skews the reward landscape has to pay for itself, and it is the arm nobody in the sixteen papers checked here runs by design.

The revival is not thereby refuted. On its own evidence the actor-side perplexity bonus improves every benchmark it is tested on, a trajectory-level count bonus improves mean accuracy everywhere and coverage on the hardest suites, and one 2026 comparison run against three undirected baselines at matched group size and matched cost comes out against the simple schemes. What can be said is that the field is answering a question it has not yet posed in the form that would settle it, and that it is reconstructing, one group at a time, conclusions that are already in the literature it cites for its premises but not for its cautions. Two papers in the revival report that the direct classical import fails, and both report it as a reason to build something new rather than as evidence about transfer. A survey asks for the shared comparison without knowing the shape of the answer the last one gave.

The cheapest correction is not a new algorithm. It is an easy-exploration arm, three seeds, and a swept temperature.

References

- Andrychowicz, M., Raichuk, A., Stanczyk, P., Orsini, M., Girgin, S., Marinier, R., Hussenot, L., Geist, M., Pietquin, O., Michalski, M., Gelly, S., & Bachem, O. (2020). What Matters In On-Policy Reinforcement Learning? A Large-Scale Empirical Study. arXiv:2006.05990. doi:10.48550/arXiv.2006.05990. [PREPRINT]
- Aubret, A., Matignon, L., & Hassas, S. (2019). A survey on intrinsic motivation in reinforcement learning. arXiv:1908.06976. doi:10.48550/arXiv.1908.06976. [PREPRINT]
- Bai, C., Zhang, Y., Qiu, S., Zhang, Q., Xu, K., & Li, X. (2025). Online Preference Alignment for Language Models via Count-based Exploration. International Conference on Learning Representations (ICLR) 2025. arXiv:2501.12735. doi:10.48550/arXiv.2501.12735.
- Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., & Munos, R. (2016). Unifying Count-Based Exploration and Intrinsic Motivation. arXiv:1606.01868. doi:10.48550/arXiv.1606.01868. [PREPRINT]
- Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., & Efros, A. A. (2018a). Large-Scale Study of Curiosity-Driven Learning. arXiv:1808.04355. doi:10.48550/arXiv.1808.04355. [PREPRINT]
- Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2018b). Exploration by Random Network Distillation. arXiv:1810.12894. doi:10.48550/arXiv.1810.12894. [PREPRINT]
- Cheng, D., Huang, S., Zhu, X., Dai, B., Zhao, W. X., Zhang, Z., & Wei, F. (2025). Reasoning with Exploration: An Entropy Perspective. AAAI Conference on Artificial Intelligence (AAAI) 2026. arXiv:2506.14758. doi:10.48550/

arXiv.2506.14758.

Cui, G., Zhang, Y., Chen, J., Yuan, L., Wang, Z., Zuo, Y., Li, H., Fan, Y., Chen, H., Chen, W., Liu, Z., Peng, H., Bai, L., Ouyang, W., Cheng, Y., Zhou, B., & Ding, N. (2025). The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models. arXiv:2505.22617. doi:10.48550/arXiv.2505.22617. [PREPRINT]

Dabney, W., Ostrovski, G., & Barreto, A. (2020). Temporally-Extended epsilon-Greedy Exploration. arXiv:2006.01782. doi:10.48550/arXiv.2006.01782. [PREPRINT]

Dai, R., Song, L., Liu, H., Liang, Z., Yu, D., Mi, H., Tu, Z., Liu, R., Zheng, T., Zhu, H., & Yu, D. (2025). CDE: Curiosity-Driven Exploration for Efficient Reinforcement Learning in Large Language Models. arXiv:2509.09675. doi:10.48550/arXiv.2509.09675. [PREPRINT]

Deng, J., Chen, J., Chen, Z., Cheng, D., Bai, F., Zhang, B., Min, Y., Gao, Y., Zhao, W. X., & Wen, J. R. (2025). From Trial-and-Error to Improvement: A Systematic Analysis of LLM Exploration Mechanisms in RLVR. arXiv:2508.07534. doi:10.48550/arXiv.2508.07534. [PREPRINT]

Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O., & Clune, J. (2019). Go-Explore: a New Approach for Hard-Exploration Problems. arXiv:1901.10995. doi:10.48550/arXiv.1901.10995. [PREPRINT]

Gao, J., Pan, L., Wang, Y., Zhong, R., Lu, C., Wang, M., Cai, Q., Jiang, P., & Zhao, X. (2025). Navigate the Unknown: Enhancing LLM Reasoning with Intrinsic Motivation Guided Exploration. arXiv:2505.17621. doi:10.48550/arXiv.2505.17621. [PREPRINT]

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2017). Deep Reinforcement Learning that Matters. Thirty-Second AAAI Conference on Artificial Intelligence (AAAI) 2018. arXiv:1709.06560. doi:10.48550/arXiv.1709.06560.

Hessel, M., Modayil, J., van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., & Silver, D. (2017). Rainbow: Combining Improvements in Deep Reinforcement Learning. Thirty-Second AAAI Conference on Artificial Intelligence (AAAI) 2018. arXiv:1710.02298. doi:10.48550/arXiv.1710.02298.

Hou, Z., An, Z., & Du, W. (2025). Beyond Noisy-TVs: Noise-Robust Exploration Via Learning Progress Monitoring. International Conference on Learning Representations (ICLR) 2026. arXiv:2509.25438. doi:10.48550/arXiv.2509.25438.

Kayal, A., Pignatelli, E., & Toni, L. (2025). The impact of intrinsic rewards on exploration in Reinforcement Learning. arXiv:2501.11533. doi:10.48550/arXiv.2501.11533. [PREPRINT]

Li, A., Yuan, Z., Zhang, Y., Liu, S., & Wang, Y. (2025). Know When to Explore: Difficulty-Aware Certainty as a Guide for LLM Reinforcement Learning. arXiv:2509.00125. doi:10.48550/arXiv.2509.00125. [PREPRINT]

Li, W., Oh, C., & Li, S. (2025). General Exploratory Bonus for Optimistic Exploration in RLHF. International Conference on Learning Representations (ICLR) 2026. arXiv:2510.03269. doi:10.48550/arXiv.2510.03269.

Lobel, S., Bagaria, A., & Konidaris, G. (2023). Flipping Coins to Estimate Pseudocounts for Exploration in Reinforcement Learning. International Conference on Machine Learning (ICML) 2023. arXiv:2306.03186. doi:10.48550/arXiv.2306.03186.

Machado, M. C., Bellemare, M. G., Talvitie, E., Veness, J., Hausknecht, M., & Bowling, M. (2017). Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents. arXiv:1709.06009. doi:10.48550/arXiv.1709.06009. [PREPRINT]

Mahendrakar, P. (2026a). The Crossover Does Not Carry the Claim: Pass@k Inversion, the Reasoning Boundary, and Five Premises the RLVR Debate Leaves Unstated. Zenodo. doi:10.5281/zenodo.22063128

Mahendrakar, P. (2026b). Transfer Is a Directed Relation: Four Quantities Under One Word in the Cross-Domain RLVR Debate, and Why the Structural-Similarity Reading Does Not Survive Its Own Evidence. Zenodo. doi:10.5281/zenodo.22313641

Ostrovski, G., Bellemare, M. G., Oord, A., & Munos, R. (2017). Count-Based Exploration with Neural Density Models. arXiv:1703.01310. doi:10.48550/arXiv.1703.01310. [PREPRINT]

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven Exploration by Self-supervised Prediction. International Conference on Machine Learning (ICML) 2017. arXiv:1705.05363. doi:10.48550/arXiv.1705.05363.

Savinov, N., Raichuk, A., Marinier, R., Vincent, D., Pollefeys, M., Lillicrap, T., & Gelly, S. (2018). Episodic Curiosity through Reachability. International Conference on Learning Representations (ICLR) 2019. arXiv:1810.02274. doi:10.48550/arXiv.1810.02274.

- Schmied, T., Bornschein, J., Grau-Moya, J., Wulfmeier, M., & Pascanu, R. (2025). LLMs are Greedy Agents: Effects of RL Fine-tuning on Decision-Making Abilities. arXiv:2504.16078. doi:10.48550/arXiv.2504.16078. [PREPRINT]
- Shen, X., Zhang, H., Li, P., Wang, Y., & Zhao, D. (2026). Boosting LLM Exploration via Weak-Model Guidance in RLVR. arXiv:2608.27420. doi:10.48550/arXiv.2608.27420. [PREPRINT]
- Song, Y., Kempe, J., & Munos, R. (2025). Outcome-based Exploration for LLM Reasoning. arXiv:2509.06941. doi:10.48550/arXiv.2509.06941. [PREPRINT]
- Sun, H., Chai, Y., Wang, S., Sun, Y., Wu, H., & Wang, H. (2025). Curiosity-Driven Reinforcement Learning from Human Feedback. Annual Meeting of the Association for Computational Linguistics (ACL) 2025. arXiv:2501.11463. doi:10.48550/arXiv.2501.11463.
- Sun, Y., Guo, J., Kok, S., Wang, Z., Wen, Z., & Zhang, Z. (2025). Efficient Reinforcement Learning for Large Language Models with Intrinsic Exploration. arXiv:2511.00794. doi:10.48550/arXiv.2511.00794. [PREPRINT]
- Taiga, A. A., Fedus, W., Machado, M. C., Courville, A., & Bellemare, M. G. (2019). Benchmarking Bonus-Based Exploration Methods on the Arcade Learning Environment. Second Exploration in Reinforcement Learning Workshop, International Conference on Machine Learning (ICML) 2019. arXiv:1908.02388. doi:10.48550/arXiv.1908.02388.
- Taiga, A. A., Fedus, W., Machado, M. C., Courville, A., & Bellemare, M. G. (2021). On Bonus-Based Exploration Methods in the Arcade Learning Environment. International Conference on Learning Representations (ICLR) 2020. arXiv:2109.11052. doi:10.48550/arXiv.2109.11052.
- Tang, H., Houthoofd, R., Foote, D., Stooke, A., Chen, X., Duan, Y., Schulman, J., De Turck, F., & Abbeel, P. (2016). #Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning. Advances in Neural Information Processing Systems (NIPS) 2017. arXiv:1611.04717. doi:10.48550/arXiv.1611.04717.
- Venkatkrishna, V., Daheim, N., & Gurevych, I. (2026). Parameter Exploration for RLVR via Variational Learning. arXiv:2608.09805. doi:10.48550/arXiv.2608.09805. [PREPRINT]
- Wang, Y. (2026). The Dark Room in the Reward Channel: Dense Prediction Rewards Collapse GRPO-Trained LLM Agents -- and The Channel, Not the Content, Decides What Works. arXiv:2607.21273. doi:10.48550/arXiv.2607.21273. [PREPRINT]
- Wen, X., Liu, Z., Zheng, S., Ye, S., Wu, Z., Wang, Y., Xu, Z., Liang, X., Li, J., Miao, Z., Bian, J., & Yang, M. (2025). Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs. arXiv:2506.14245. doi:10.48550/arXiv.2506.14245. [PREPRINT]
- Yang, Z., Jiang, Y., Chen, T. C., Yang, L., Wong, A., Gao, C., Kooi, J. E., Li, Z., Shi, J., Qiu, K., Huang, Q., Zu, X., Yang, S., Zhang, H., Wong, N., Ilievski, F., Yu, S., Plaat, A., Ren, Z., Hoogendoorn, M., & Francois-Lavet, V. (2026). Modularized Reinforcement Learning on LLMs: From MDP Creation to Exploration and Learning. arXiv:2606.21943. doi:10.48550/arXiv.2606.21943. [PREPRINT]
- Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, W., Zhu, H., Zhu, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W. Y., Zhang, Y. Q., Yan, L., Qiao, M., Wu, Y., & Wang, M. (2025). DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv:2503.14476. doi:10.48550/arXiv.2503.14476. [PREPRINT]
- Yu, Z., Mou, Y., Yan, J., Luo, J., Chen, C., Wei, X., Liu, Y., Sun, H., Zhang, Y., Xu, J., Bian, Y., Zhang, M., Ye, W., He, T., Yang, J., Zheng, G., Wu, Z., Zhang, B., Bai, L., & Luo, X. (2026). A Survey of Reinforcement Learning for Large Language Models under Data Scarcity: Challenges and Solutions. Annual Meeting of the Association for Computational Linguistics (ACL) 2026. arXiv:2604.17312. doi:10.48550/arXiv.2604.17312.
- Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Yue, Y., Song, S., & Huang, G. (2025). Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? Advances in Neural Information Processing Systems (NeurIPS) 2025 (Oral). arXiv:2504.13837. doi:10.48550/arXiv.2504.13837.
- Zhang, X., Li, R., Zhou, Z., Li, L., Qin, Y., Li, K., Sun, X., Tan, X., Qu, C., & Qi, Y. (2025). Count Counts: Motivating Exploration in LLM Reasoning with Count-based Intrinsic Rewards. arXiv:2510.16614. doi:10.48550/arXiv.2510.16614. [PREPRINT]
- Zhao, X., Kang, Z., Feng, A., Levine, S., & Song, D. (2025). Learning to Reason without External Rewards. International Conference on Learning Representations (ICLR) 2026. arXiv:2505.19590. doi:10.48550/arXiv.2505.19590.

The literature search, drafting and citation verification for this paper were carried out with AI assistance under the author's direction. Every citation was machine-verified against the arXiv record before inclusion, and every quantitative claim was read back against the cited source's own abstract or, where a claim is drawn from a paper's body or a published table, against the located passage. The absence claims in Section 7 were produced by fetching each named paper's full text and searching it mechanically; the procedure is stated in Section 2 so that it can be repeated. The author is responsible for the final text and for all claims made in it.