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

Mahendrakar, Pranay
SONYTECH
ORCID: 0009-0003-7224-029X
Life of Research
2026-09-05
Pranay Mahendrakar

Abstract

Reinforcement learning with verifiable rewards is the standard route to reasoning-tuned language models, and the field has split over whether its gains leave the training domain. One body of work reports that most models succeeding at mathematics fail to transfer and that single-domain post-training yields no statistically significant out-of-domain improvement; another reports that post-training on constraint-satisfaction puzzles alone raises hard-mathematics accuracy substantially. This paper argues that the two are not in direct contradiction, because the word transfer carries at least four logically independent quantities: whether training on a domain improves others, whether a domain improves when others are trained, whether a domain is preserved rather than degraded, and whether joint training beats separate training. Papers measuring one are routinely cited as evidence about another. The paper then argues that the reconciliation the field has settled on - that transfer follows structural similarity between source and target - is contradicted by the directional findings of the negative result usually cited for it, which reports unstructured domains transferring to structured ones while failing to transfer to each other. Similarity is symmetric; the reported relation is not, and multi-task learning has treated directed, sign-bearing task-affinity matrices as its normal object of study since Taskonomy. Two further problems are set out: the source-side gain may be substantially elicitation of a pretraining-frequent behaviour rather than acquired skill, which would relocate transfer to the base model, and the reported effect sizes sit near a documented seed-to-seed noise floor. What is not known is stated flatly: no published experiment reports a full directed transfer matrix over a fixed domain set, one protocol and more than one model family. Five measurements that would settle the open part are specified.

1. Introduction

Reinforcement learning with verifiable rewards (RLVR) became the default post-training recipe for reasoning models after DeepSeek-R1 reported that reasoning abilities can be incentivised through pure reinforcement learning, without human-labelled reasoning trajectories (DeepSeek-AI et al., 2025). The recipe is narrow by construction: it needs an automatic verifier, and automatic verifiers are easiest to build for mathematics and code. Almost every headline result is therefore obtained by training on mathematics and evaluating on mathematics.

The interesting question is what happens outside that box. If RLVR installs a general reasoning capability, then the mathematics corpus is a convenient scaffold and the gains should appear on tasks that share nothing with it but the demand to reason. If it installs something narrower, then the field has

been buying a domain-specific improvement and describing it as a general one. The question has practical weight, because the entire multi-domain training industry - corpus curation across six domains (Cheng et al., 2025), curriculum design over domain mixtures (Yang et al., 2026), expert merging and dataset mixing (Wang et al., 2026; Wu, S. et al., 2026), and reproducible general-purpose recipes (Pardinas et al., 2026) - presupposes that domains interact in a stable and knowable way.

The published answers conflict, and they conflict at a scale that scale does not explain. On one side, an evaluation of over twenty open-weight reasoning-tuned models across mathematics, scientific question answering, agent planning, coding and instruction following reports that most models that succeed in mathematics fail to transfer their gains to other domains (Huan et al., 2025). A study combining an observational comparison of eighteen open-weight post-trained models against their base models over sixteen benchmarks with an interventional study training three single-domain models from a common base concludes that gains generalise inconsistently and can vanish on domains with different reasoning patterns, and reports that none of its single-domain models achieved a statistically significant improvement on out-of-domain tasks (Hu et al., 2025). Supervised fine-tuning of a 1.7B model on 40,000 plan tuples from ten planning domains yields 82.9 percent valid plans in domain and 0 percent on two unseen domains, and adding a verifier reward reaches saturation in half the supervised training epochs while not improving cross-domain generalisation at all (Belcamino et al., 2026).

On the other side, a 7B model whose entire supervised and reinforcement post-training uses constraint-satisfaction puzzles, with no mathematics problems in the post-training data at all, moves from 16.0 to 36.0 percent pass@32 on OlymMATH-Hard (Maniparambil et al., 2026). Rule-based reinforcement learning on 5,000 synthetic logic problems is reported to produce reasoning skills absent from the logic corpus - reflection, verification, summarisation - and generalisation to the AIME and AMC mathematics benchmarks (Xie et al., 2025), and a recipe that warms up on Knights and Knaves puzzles before applying RLVR is built on the same premise (Shrestha et al., 2025). Prolonged reinforcement learning on a small model across a diverse task set reports gains of 14.7 percent on mathematics, 13.9 percent on coding and 54.8 percent on logic puzzles against strong baselines (Liu, M. et al., 2025). A curated six-domain corpus reports that domains frequently seen during pretraining benefit from cross-domain reinforcement learning, while domains with limited pretraining exposure require in-domain training, and reads this as evidence that RLVR facilitates genuine skill acquisition (Cheng et al., 2025).

Both sides use overlapping model families, overlapping benchmarks and the same optimiser family. Neither side is obviously careless. The usual way to reconcile them, and the way the negative result is most often glossed, is a similarity story: gains move between domains whose reasoning patterns resemble each other, mathematics to code, and stop at the boundary of domains that do not. This paper argues that the similarity story is the wrong reconciliation and that the disagreement is largely an artefact of four things being called transfer.

The argument has four parts. Section 4 separates the four quantities and shows which paper measures which. Section 5 shows that the relation the field is measuring is directed and not symmetric, that the leading negative result reports an asymmetry inside its own findings which a similarity account cannot produce, and that the multi-task learning literature has treated directed transfer matrices as the normal object of study since at least Taskonomy (Zamir et al., 2018). Section 6 argues that the source-side quantity being propagated is partly an elicited pretraining behaviour rather than an acquired skill, which relocates transfer from a property of the training signal to a property of the base model. Section 7 places the reported effect sizes against a documented noise floor.

The strongest objection to this paper appears here rather than at the end. A reader may hold that separating four senses of a word is commentary rather than contribution, and that the field's practitioners already know their measurements differ. Section 11 concedes that the objection has force and states the specific form in which it would be decisive. What the objection does not cover is the

asymmetry argument in Section 5, which is a claim about the content of a published result rather than about vocabulary, and which is checkable against that paper.

2. Scope

This paper is a literature analysis. No experiments were run for it, no model was trained and no benchmark was executed. Every quantitative statement below is attributed to a published source and was verified against that source. Where a number is drawn from a paper's body rather than its abstract, the text says so.

Three boundaries are worth stating before the argument, because each marks a neighbouring question that this paper deliberately does not answer.

The first is the reasoning-boundary question. Whether RLVR expands the set of problems a model can solve within its training domain, or only concentrates probability mass on solutions the base model could already reach, is a separate and well-developed dispute, running from the pass@k inversion result (Yue et al., 2025) through support-constrained analyses (Wu, F. et al., 2025) and overtraining accounts of diversity collapse (Yuan et al., 2026) to rebuttals arguing the boundary does extend once the derivation rather than the answer is scored (Wen et al., 2025). A prior record examined that dispute and argued that its six rebuttals are mutually non-equivalent (Mahendrakar, 2026). The present paper takes no position on it. The two questions are connected - if RLVR only sharpens within domain, cross-domain gain is hard to account for - but they are not the same question, and Section 8 is careful not to import the boundary argument as a premise.

The second is modality. A substantial line studies whether text-only reasoning training transfers to multimodal tasks (Liu, Q. et al., 2025; Zhang, X. et al., 2026) and whether reward design can separate perception from reasoning. Modality transfer raises its own confounds and is out of scope; the domains at issue here are textual reasoning domains.

The third is method advocacy. Several works cited below propose training methods: transfer-aware curricula (Yang et al., 2026), verifier-free extensions to general domains (Yu et al., 2025; Piao and Park, 2026), transfer-based reward shaping (Lee et al., 2026), and fusion of domain experts (Wu, S. et al., 2026; Wang et al., 2026). They are cited for their measurements, not endorsed as methods. This paper makes no claim about which recipe is best.

3. The disagreement as it is usually stated

Stated at the level the field states it, the two positions are these.

The pessimistic position holds that RLVR gains are domain-local. Its strongest evidence is Hu et al. (2025), which runs both an observational and an interventional study and reports that both converge on the same conclusion: that although post-training brings substantial gains on tasks similar to the fine-tuning data, the gains generalise inconsistently and can vanish on domains with different reasoning patterns. Its interventional arm trained three models from DeepSeek-R1-Distill-Qwen-1.5B on three disjoint 40,000-problem sets - mathematics, code, and non-mathematical non-code knowledge-intensive problems - and evaluated all three across all domains. Its supporting evidence is Huan et al. (2025), whose evaluation of over twenty reasoning-tuned models finds that most models succeeding at mathematics fail to transfer, and Belcamino et al. (2026), whose planning model collapses to zero on unseen domains.

The second of those citations is misused whenever it is offered without its other half, and this paper flags it here rather than in a footnote. Huan et al. (2025) follow the observational finding with a controlled experiment on Qwen3-14B using math-only data and different tuning methods, and report

that reinforcement learning-tuned models generalise well across domains while supervised fine-tuning-tuned models often forget general capabilities. Their headline is a negative transfer result; their controlled result locates the failure on the supervised arm. Section 9 returns to what that does to the pooled comparisons.

The optimistic position holds that reasoning training installs something portable. Its strongest evidence is the puzzle-to-mathematics result of Maniparambil et al. (2026), because the post-training data contains no mathematics at all, so the mathematics gain cannot be in-domain improvement under any description. Its supporting evidence is Xie et al. (2025) and Wong et al. (2025), both training on logic puzzles and reporting mathematical gains, and the multi-domain corpus of Cheng et al. (2025), whose cross-domain training benefits Math, Code and Science without in-domain data for the domain being improved.

A third position has appeared more recently and is easy to miss because it does not present itself as a position. Lu et al. (2025), studying RLVR on probabilistic inference over causal graphs, report that RLVR is less effective when tasks are too challenging before fine-tuning and improves generalisation substantially once the model has a basic success rate. That makes transfer conditional on where the model started, which is a variable neither camp reports.

No survey settles the dispute. The nearest is a reward-modelling survey that organises RLVR alongside RLHF, preference optimisation and judge-based rewards under a reasoning-centric taxonomy, and treats distribution shift as one of several core challenges rather than as its subject (Pan et al., 2026). The transfer question is not its object, and nothing located in this search maps the cross-domain results against each other.

Read as competing answers to one question, these are irreconcilable. Read carefully, they are answers to different questions, and the differences are not subtle once named.

4. Four quantities under one word

Let a domain be a set of tasks with a shared verifier and a shared informal label: mathematics, code, logic, tabular reasoning, and so on. The field's papers partition the space differently, which is itself a problem taken up in Section 5.3, but the four quantities below can be stated under any partition.

T1, out-transfer. Training on domain S improves performance on domain T, where T is not in the training data. This is the quantity in the phrase "does reasoning transfer". It is what Maniparambil et al. (2026) measure for the ordered pair (puzzles, mathematics), what Xie et al. (2025) measure for (logic, mathematics), and what the interventional arm of Hu et al. (2025) measures across the nine cells of a three-domain matrix, six of which are off-diagonal.

T2, in-transfer. Domain T improves when a model is trained on domains other than T. This is the quantity in Cheng et al. (2025)'s finding that Math, Code and Science "easily benefit from cross-domain RL training" while Logic, Simulation and Tabular "require in-domain training to achieve meaningful performance gains". Note the grammatical subject. The claim is about which domains receive, not about which domains give.

T3, retention. Domain T does not degrade when a model is trained on S. This is a different quantity from T1 in the way that a null result is different from a negative one, and the distinction is load-bearing in this literature because the alternative to gain is not always neutrality. Hu et al. (2025) report that their mathematics-trained model improves on mathematics and code while performance drops notably on knowledge-intensive tasks. Instruction following degrades as reasoning capacity scales, on a benchmark built for the purpose (Fu et al., 2025). The retention question has its own mature sub-literature comparing reinforcement learning against supervised fine-tuning (Shenfeld et al., 2025; Rojas

Nunez et al., 2026), discussed in Section 9.

T4, joint-training synergy. Training on S and T together outperforms training on each separately. This is the quantity that matters to anyone actually building a general model, and it is what the multi-domain studies measure: the mutual enhancements and conflicts of combined cross-domain training (Li, Y. et al., 2025), the choice between pooling datasets and merging separately trained experts (Wang et al., 2026), and the three-way comparison of merging, mixing and multi-teacher distillation (Wu, S. et al., 2026). T4 can be positive when T1 is zero - two domains can be jointly learnable without either transferring to the other in isolation - and it can be negative when T1 is positive, if the joint objective introduces gradient conflict that the sequential case does not.

These four are logically independent. T1 and T2 are converses over the same relation and differ whenever the relation is not symmetric, which Section 5 argues it is not. T3 is the sign of a quantity T1 leaves unspecified when it reports no significant gain. T4 involves an optimisation interaction that none of the other three contain.

The substitutions in current practice run in every direction. Hu et al. (2025) measure T1 and T3 and are cited for a claim about reasoning generality, which would need T1 over a domain set broad enough to stand for reasoning. Cheng et al. (2025) measure T2 and draw a conclusion about skill acquisition, which is a claim about what the training installed, not about which domains receive. Yang et al. (2026) measure something closer to T4 - the effect of a gradient step on one domain upon the remaining training domains - and report it as transferability, which reads as T1. None of these is a mistake inside its own paper. The mistake is the citation practice that treats them as measurements of one thing.

One consequence deserves emphasis because it dissolves part of the apparent conflict at no cost. Cheng et al. (2025) and Hu et al. (2025) are frequently placed on opposite sides. They are not measuring opposite signs of one quantity. Cheng et al. report that Math has high T2 - it benefits when other domains are trained. Yang et al. (2026) report, in their body text, that math "ranks among the least transferable" domains in their suite and that "math and codegen align weakly, often negatively, with every other domain", which is a statement that Math has low T1. Both can hold simultaneously, and if they do, the field's canonical experiment - train on mathematics, evaluate elsewhere - is being run from the worst available source domain, and its negative results are then generalised to RLVR as such.

4.1 What the substitution costs

The confusion is not merely terminological, because the four quantities support different decisions.

A team choosing what to put in a training mixture needs T4. A team deciding whether to bother collecting data for a low-resource domain needs T2: if the domain receives from others, the collection may be unnecessary. A team deciding which single domain to invest verifier engineering in needs T1, because a high-out-transfer source pays for itself across the suite. A team shipping a model that must retain a capability it is not training on needs T3, and for T3 the relevant comparison is not against the base model's score on the source domain at all.

These are the decisions the multi-domain training literature is actually making. Wang et al. (2026) compare pooling datasets against merging separately trained experts; Wu, S. et al. (2026) compare merging, mixing and multi-teacher on-policy distillation and report that the three paradigms are within 1.4 points of one another on average yet as much as 8.6 points apart on an individual benchmark, and that how they differ from domain to domain follows the relations between domains that can be read off the geometry of their task vectors; Yang et al. (2026) allocate sampling across domains online. Every one of those is a T4 decision. The evidence most often cited to motivate them - that reasoning does or does not transfer - is T1 evidence. A T1 null does not imply a T4 null, and a T4 gain does not license a claim about general reasoning.

The reverse substitution is the one that produces the field's headline. "Does reasoning transfer" is a question about a capability, which would require T1 measured over a domain set broad enough for the answer to generalise beyond the cells tested. The largest available T1 study tests three domains in its interventional arm (Hu et al., 2025). Three domains give six ordered off-diagonal pairs. That is not enough cells to support a claim about reasoning as such, and the paper does not make one; the claim is made on its behalf in citation.

5. The relation is directed, and the similarity reading does not survive

5.1 What the leading negative result actually reports

The similarity reading of Hu et al. (2025) is that gains move between domains with similar reasoning patterns and stop at dissimilar ones. It is the reading the paper's own framing invites. Its abstract says gains "can vanish on domains with different reasoning patterns", and the section reporting the cross-domain analysis is titled "RPT Gains Generalize Across Domains with Similar Reasoning Patterns".

The cells reported under that title do not fit it. The paper divides its domains into two classes. Structured reasoning covers mathematics and code, where in its own words solutions follow deterministic logical steps and require precise syntax and formal semantics. Unstructured reasoning is defined as problem-solving that does not adhere to a fixed sequence of logical operations and lacks a well-defined intermediate representation or symbolic grounding. Its findings are then reported as four separately headed verdicts: structured-to-structured generalisation is effective; structured reasoning patterns that are more foundational tend to exhibit stronger cross-domain transfer; structured-to-unstructured generalisation is limited; and unstructured-to-structured generalisation is promising.

The fourth verdict is the one that matters, and the paper supports it with numbers in both arms. In the observational arm, knowledge-intensive RPT models show substantially higher pass@1 improvements on mathematics, at 21.40 percent, and on code, at 12.16 percent, than on tasks inside the knowledge-intensive domain itself. In the interventional arm, the knowledge-trained model achieves statistically significant gains on mathematics-domain tasks and shows no noticeable degradation on code, while underperforming on tasks within its own domain.

Set the class-level cells side by side. Structured to structured: effective. Structured to unstructured: limited. Unstructured to structured: promising. Unstructured to unstructured: the source's worst target, below both structured targets in the observational arm and an absolute underperformance against the base model in the interventional one. A symmetric similarity account predicts the two within-class cells to beat the two cross-class ones, because similarity at the class level is symmetric and reflexive. Two cells go the other way, and they are the decisive two: the most similar pairing available is the weakest result in the set, and a cross-class pairing is among the strongest.

One qualification belongs on that last cell and it cuts in the argument's favour rather than against it. The unstructured-to-unstructured cell is not a clean off-diagonal entry, because the knowledge-intensive class contains the training domain itself, so part of what is being scored there is in-domain performance. A transfer study would normally exclude the diagonal. Including it here makes the observation stronger, not weaker: the knowledge-trained model is doing worse on the class it was trained on, diagonal included, than on a class it never saw. No account on which transfer decreases with distance can produce that.

Whatever produces that pattern, it is not proximity between source and target. Proximity is symmetric; the reported relation is not. The paper does not in fact rest on proximity when it explains the fourth cell. Its stated interpretation is that unstructured reasoning patterns encompass broader representational complexity and implicitly subsume the essential components of structured reasoning,

functioning as a conceptual superset. That is a containment account, and containment is a directed relation - if U subsumes S then transfer from U to S and transfer from S to U are not the same event, and there is no reason to expect the same sign. So the paper's own mechanism for its hardest cell already requires the relation to be directed. What does not survive is the similarity gloss in the section title, the abstract sentence, and the citations that travel on them.

This is the paper's central factual claim about the literature and it is checkable in one place. The way it could fail should be stated: if the four verdicts are read as one ordered gradient of transfer difficulty rather than as cells of a relation, a reader can hold that the paper is describing a single similarity axis on which unstructured targets simply sit at the hard end. That reading has to explain why the unstructured source does better on distant targets than on its own class, and the containment account is the paper's answer to exactly that, so the reading collapses back into a directed relation. But it is the objection to make.

One further point should be recorded because it bears on how the result is cited. The paper also states that none of the models post-trained on a single domain exhibits statistically significant improvement on out-of-domain tasks, which is the sentence its negative reputation rests on. That statement and the significant mathematics gain reported for the knowledge-trained model cannot both hold at their face scope, since mathematics is out of domain for that model. The narrower statement is presumably the one intended to be general, but the resolution is not available from the text located here, and it is a reason to read the result tables before citing either sentence alone.

5.2 The multi-task literature had this settled

Directed, asymmetric task relationships are not a discovery. Multitask learning was introduced as a mechanism whereby the inductive bias supplied by related tasks improves generalisation (Caruana, 1997), and the practical study of which tasks help which has produced explicitly directed objects for a decade. Taskonomy computed first-order and higher-order transfer relationships across a bank of visual tasks and organised them into a structure rather than a similarity metric (Zamir et al., 2018). The task-grouping line asks which tasks should be learned together precisely because task objectives compete, and reports that naive joint training often degrades performance (Standley et al., 2020; Fifty et al., 2021) - that is, T4 is routinely negative even among tasks a similarity judgement would group. Negative transfer has its own survey (Zhang, W. et al., 2022).

Three properties of that literature are worth importing wholesale. Transfer is recorded as a matrix indexed by ordered pairs, not as a property of a domain. Sign is reported, not just magnitude, because negative entries are common and informative. And the object is read for its off-diagonal structure: Taskonomy's stated purpose is to identify redundancies so supervision can be reused across related tasks, which presupposes that training directly on a target is not automatically the best use of a fixed supervision budget.

The RLVR transfer literature currently reports row sums, column sums and individual cells from three different matrices, computed on three different domain partitions, and argues about the result. Yang et al. (2026) come closest to the matrix, computing a per-domain transferability score over six domains and observing that the most learnable domain is often not the most transferable. That observation is the correct shape. It is also, as far as this search located, the only place in the RLVR literature where learnability and transferability are separated as distinct axes, and it is reported in service of a curriculum method rather than as a finding about the debate.

5.3 The partition is not fixed, so the cells are not comparable

Three studies at the centre of this dispute partition the space differently. Hu et al. (2025) use mathematics, code and knowledge-intensive reasoning in their interventional arm and a structured/

unstructured cut in their analysis. Cheng et al. (2025) use Math, Code, Science, Logic, Simulation and Tabular. Yang et al. (2026) use math, codegen, logic, simulation, table and stem, which is the same six-way cut. Li, Y. et al. (2025) use mathematics, code generation and logical puzzles.

The cuts are not refinements of one another. "Knowledge-intensive reasoning" in the first is not a cell in the second; "stem" overlaps "Science" but not identically; logic puzzles appear as a training domain in one study and as an evaluation-only probe in another. A transfer claim is a claim about an ordered pair, so a claim stated under one partition does not translate into another without an argument that the cells correspond, and no such argument appears in any source located here. This is a weaker point than Section 5.1 - partitions being incommensurable is a familiar complaint about any comparative literature - but it compounds with the four-quantity problem, because the same paper can be cited for a different quantity under a different partition and the two errors are invisible together.

6. What is being propagated may not be a skill

Sections 4 and 5 argue that the debate's participants are measuring different things. This section argues something stronger and less comfortable: that the source-side quantity may not be the kind of thing that has a transfer profile at all.

The in-domain gains that transfer studies propagate outward are, on the most studied model family, substantially detachable from the correctness of the reward. Qwen2.5-Math-7B gains 21.4 percentage points on MATH-500 when RLVR is run with randomly assigned rewards, against 29.1 points from ground-truth rewards (Shao et al., 2025). The mechanism the authors give is a clipping bias in GRPO that amplifies high-prior behaviours already learned during pretraining, and the case study is code reasoning - reasoning in code without executing it - whose frequency rises from 65 percent to over 90 percent under spurious rewards. Critically for the present argument, the same paper reports that spurious rewards effective for Qwen models often fail to produce gains for other families such as Llama3 or OLMo2. RLVR with a single training example raises MATH500 accuracy for Qwen2.5-Math-1.5B from 36.0 to 73.6 percent, reported by the authors as an 8.6 point improvement beyond format correction, and raises a six-benchmark average from 17.6 to 35.7 percent, reported as a 7.0 point non-format gain (Wang, Y. et al., 2025). That the non-format gain has to be reported separately is itself the point: enough of a headline RLVR number on this family is the model learning to emit an answer the grader can parse that the authors net it out before making their claim. No located transfer study nets it out before propagating the gain outward.

A contamination analysis pushes further, reporting that "pre-training on massive web-scale corpora leaves Qwen2.5 susceptible to data contamination in widely used benchmarks", that the striking results seldom transfer to models like Llama, and that on a leakage-free synthetic arithmetic dataset only accurate reward signals yield steady improvements while random or incorrect rewards do not (Wu, M. et al., 2025). That last clause matters for calibration: the spurious-reward phenomenon is not a claim that reward correctness is irrelevant in general. It is a claim that on contaminated benchmarks and a particular family, a large fraction of the measured gain is elicitation.

If a meaningful share of the source-side gain is the amplification of a behaviour the base model already had at high prior, then the out-of-domain measurement is answering a different question from the one it is asked. It is measuring whether that particular elicited behaviour happens to be useful on the target domain. Under that description, three otherwise puzzling observations line up.

First, family dependence becomes expected rather than anomalous. If transfer is the applicability of an elicited behaviour, and the behaviour is a property of the pretraining distribution, then two families given identical RLVR should show different transfer profiles. That is what is reported (Shao et al., 2025; Wu, M. et al., 2025), and it is a fact that no similarity-between-domains account predicts.

Second, Cheng et al. (2025)'s result reads as the same variable from the target side. Domains frequently seen during pretraining benefit from cross-domain training; domains with limited pretraining exposure require in-domain training. Their reading is that RLVR facilitates genuine skill acquisition, on the ground that low-exposure domains improve when trained directly. The reading is defensible and this paper does not claim it is wrong. But the same numbers support a weaker reading that they do not exclude: cross-domain training is sufficient to surface a capability the pretraining corpus already supports, and insufficient to install one it does not. Under that reading the pretraining distribution is the transfer variable, and the source domain matters mainly by determining which prior behaviour gets amplified.

Third, the asymmetry in Section 5.1 acquires a second candidate mechanism, which competes with the containment account Hu et al. (2025) give. On the containment reading, unstructured reasoning subsumes structured reasoning, so training on the superset supplies the subset. On an elicitation reading, structured post-training amplifies behaviours - explicit multi-step derivation, verification against a checkable target - that a knowledge-intensive task does not reward and may actively penalise, while unstructured post-training amplifies behaviours a structured task can still use. The two make different predictions and are separable: containment predicts the asymmetry survives a change of model family, elicitation predicts it does not. Neither has been tested against the other, and this paragraph is a conjecture rather than a finding.

The honest limit of this section is that the elicitation results are concentrated on one family and mostly on mathematics benchmarks with documented contamination. They do not establish that RLVR gains are generally elicitation. They establish that on the family most transfer studies build on, the source-side quantity is not cleanly a learned skill, and that no located transfer study controls for this.

7. The measurement floor

The effect sizes in this debate are small, and the benchmarks are noisy in a way that has been quantified.

A standardised re-evaluation of reasoning methods reports that current mathematical reasoning benchmarks are highly sensitive to decoding parameters, random seeds, prompt formatting and even hardware and software configuration, and that most reinforcement learning approaches yield only modest improvements far below prior claims, with a marked tendency to overfit on small benchmarks such as AIME'24 (Hochlehnert et al., 2025). In the paper's body, pass@1 standard deviation across seeds is reported to range from 5 to 15 percentage points; a change in a single question shifts pass@1 by 2.5 to 3.3 points; and performance on AIME'24 varied by up to 8 percent for one model and 6 percent for another across compute clusters.

Transfer deltas are typically smaller than that. This does not invalidate the transfer results, and it would be a scope error to say that it does. Hu et al. (2025) applied Cochran-Mantel-Haenszel testing at $p < 0.05$, which is exactly the right response to a noisy instrument, and their central negative claim is correspondingly modest: none of their single-domain models exhibits a statistically significant improvement on out-of-domain tasks. That is a statement about the absence of a detectable effect, not a demonstration that the effect is zero. Their sixteen-benchmark suite is larger than the small benchmarks Hochlehnert et al. single out, which helps. The general point stands regardless: any design aiming to detect a two-point transfer effect on a benchmark whose seed-to-seed standard deviation is five points or more is underpowered by construction, and an underpowered null is weak evidence for a negative conclusion. No located transfer study reports a power analysis or a minimum detectable effect.

The asymmetry in how this cuts is worth stating. Noise makes the negative results weaker than they read, because failure to detect is what an underpowered design produces by default. It leaves the one

large positive transfer result - 16.0 to 36.0 percent pass@32 on OlymMATH-Hard (Maniparambil et al., 2026) - relatively untouched, because a noise floor of a few points does not manufacture a twenty-point gap. What the noise floor most damages is the middle of the distribution, which is where most of the argued-over cells sit.

8. What the positive results establish

The positive-transfer results are real and this paper does not discount them. It is worth being precise about what each licenses, because they are usually cited in a bundle.

Maniparambil et al. (2026) is the cleanest case in the located literature, because the post-training data contains no mathematics. A 7B model post-trained only on constraint-satisfaction puzzles moves from 16.0 to 36.0 percent pass@32 on OlymMATH-Hard, decomposed by the authors into roughly 7 points from puzzle supervised fine-tuning, a further 6 from vanilla GSPO reinforcement learning, and a further 7 from a diversity bonus. Two features of this deserve more attention than they get. The reinforcement learning stage is reported to suppress exploratory primitives such as hypothesising and backtracking, and the third component of the gain comes from a novelty bonus introduced to counteract that suppression - so even in the field's best transfer demonstration, the RL step is doing something that had to be corrected before the transfer was fully realised. And the direction is puzzles to mathematics: from a domain Cheng et al. (2025) classify as low-pretraining-exposure to one they classify as high. That is precisely the direction their account predicts should work.

Xie et al. (2025) and Wong et al. (2025) run the same direction, logic puzzles to mathematics. Shrestha et al. (2025) use it as a deliberate warm-up stage. The consistency is itself informative: the field's positive transfer demonstrations cluster in one direction of one ordered pair type, and that clustering is not evidence of general transferability. It is evidence about one cell.

Liu, M. et al. (2025) report large multi-domain gains from prolonged training, but the training set is itself diverse, so the result speaks to T4 rather than T1. Absolute Zero (Zhao et al., 2025) and the verifier-free extensions (Yu et al., 2025; Piao and Park, 2026) widen the set of domains RLVR can be applied to, which changes the supply of source domains without settling how they interact. Chu et al. (2025) report that reinforcement learning generalises where supervised fine-tuning memorises, on rule and visual variants in two controlled environments; this is a claim about within-task variant generalisation and is not the same claim as cross-domain transfer, though it is often cited as though it were.

Finally, the conditional from Lu et al. (2025) noted in Section 3 bites hardest here. If RLVR is ineffective on tasks the model cannot already partly do, and improves generalisation substantially once a basic success rate exists, then transfer measured from a source the model was near-zero on and transfer measured from a source it was moderately good at are not the same measurement. No located study reports the base model's success rate on the source domain as a controlled variable, which means the positive and negative results may differ in where they started rather than in what they found.

9. The comparison class problem

A transfer number is a difference, and the field is not consistent about what it is a difference from.

Two positions are in print on whether reinforcement learning or supervised fine-tuning generalises better, and they are not straightforwardly reconcilable. On one side, a cluster of results converging from behaviour, mechanism and circuits. Chu et al. (2025) report that RL, particularly with outcome-based rewards, generalises across rule-based textual and visual variants where SFT memorises. Huan et al. (2025), in the controlled half of their study on Qwen3-14B with math-only data, report that RL-tuned models generalise well across domains while SFT-tuned models often forget general capabilities, with

latent-space and token-distribution analyses attributing the difference to representation drift under SFT. Shenfeld et al. (2025) give a mechanism for the same asymmetry: the degree of forgetting tracks the KL divergence between the fine-tuned and base policy on the new task, and on-policy RL is implicitly biased toward KL-minimal solutions. A circuit-level follow-up reports the behavioural account is mirrored mechanistically (Rojas Nunez et al., 2026).

Against that, Hochlehnert et al. (2025) report that under their standardised protocol supervised fine-tuning methods show consistently stronger generalisation than RL methods in the settings they study. The apparent contradiction is partially resolved by scope: their comparison is about generalisation across mathematics evaluations under controlled decoding, whereas the other three are about retention of non-target capabilities. Those are T1-within-domain and T3 respectively, which Section 4 separates. The resolution is only partial, because both are routinely invoked in the same argument about whether RL generalises.

The comparison-class problem has a second face. Huan et al. (2025)'s observational finding - most models that succeed at mathematics fail to transfer - is drawn from open-weight models whose post-training recipes differ, and their own controlled experiment locates the transfer failure on the SFT arm. Hu et al. (2025)'s observational arm compares eighteen post-trained models against their base models, and its interventional arm builds on DeepSeek-R1-Distill-Qwen-1.5B, itself a distillation product. A pooled comparison across recipes measures the average effect of whatever those recipes did, and if the SFT component is where transfer is lost, a pooled negative result about "reinforcement post training" is attributing an effect to the wrong stage. Hu et al. do not claim otherwise; their observational arm is explicitly observational, and their interventional arm is what carries their causal claim. But the observational finding is what gets cited.

10. Five measurements that would settle the open part

None of what follows requires a new method. All five are measurement work on existing artifacts, and the infrastructure exists: Reasoning Gym supplies over 100 procedural data generators and verifiers across algebra, arithmetic, computation, cognition, geometry, graph theory, logic and games, with adjustable complexity (Stojanovski et al., 2025), and Guru supplies a 92,000-example six-domain verifiable corpus (Cheng et al., 2025).

M1. A directed transfer matrix under one protocol. For a fixed domain set of n domains, train n single-domain models from one base under one recipe and one compute budget, and evaluate all n models on all n domains. Report the full n-by-n matrix with signs and confidence intervals, not row or column summaries. This is the object Taskonomy built for visual tasks (Zamir et al., 2018) and it does not exist for RLVR domains. Hu et al. (2025) come closest at $n = 3$. What it would settle: whether the relation is asymmetric, which is currently inferred from two sentences in one paper's body.

M2. The same matrix on at least three model families. Qwen, Llama and OLMo at minimum, given that the spurious-reward and contamination results are family-split along exactly that line (Shao et al., 2025; Wu, M. et al., 2025). What it would settle: whether the transfer matrix is a property of the training signal or of the pretraining distribution. If the matrices differ substantially in structure and not merely in scale, transfer is not a property of RLVR and the field's central question is mis-specified. It would also separate the two accounts of the structured/unstructured asymmetry set out in Section 6, since a containment relation between reasoning types should hold across families and an elicited pretraining behaviour should not.

M3. Source-side decomposition before propagation. For each source domain, report the in-domain gain decomposed into the format-compliance share and the remainder, following the decomposition Wang, Y. et al. (2025) report for one-shot RLVR, and report the spurious-reward control - the gain

obtained with randomised rewards on the same source. Then measure out-transfer separately for the two components. What it would settle: whether elicited behaviour and acquired capability have different transfer profiles, which Section 6 conjectures and nothing in the located literature tests.

M4. Power analysis before the null. For every reported absence of transfer, state the minimum detectable effect given the benchmark size, seed variance and number of seeds, using the variance figures now available (Hochlehnert et al., 2025). What it would settle: which of the field's negative results are evidence of absence and which are absence of evidence. This is the cheapest of the five and would change how several existing results should be read.

M5. Base-rate-controlled sources. Following the conditional reported by Lu et al. (2025), stratify source domains by the base model's pre-training success rate on them and report transfer within strata. What it would settle: whether the apparent conflict between the puzzle-to-mathematics successes and the mathematics-to-elsewhere failures is a direction effect or a base-rate effect. The puzzle results start from a low-competence source and target a high-competence domain; the mathematics results do the reverse. No located study varies one while holding the other.

A sixth item is worth naming although it is not a measurement: the field needs an agreed domain partition, or explicit crosswalks between the partitions in use. Absent that, M1 through M5 will produce five more incomparable matrices.

11. Threats to this paper's contribution

The most serious objection is that the four-quantity separation in Section 4 is a relabelling of distinctions competent researchers already draw internally, and that naming them adds vocabulary rather than knowledge. The objection has real force. Every paper cited here is internally consistent about what it measured; the confusion documented in Section 4 lives in the citation practice between papers, not inside any of them, and a reader who regards inter-paper citation slippage as a normal cost of a fast-moving field will not find the section worth its length. The reply is that Section 4 is load-bearing for Section 5 rather than freestanding - the T1/T2 distinction is what makes the Cheng et al. and Yang et al. results compatible rather than contradictory, and that compatibility is not obvious without it - but a reviewer who accepts that and still calls the section commentary is not being unreasonable.

The second objection concerns Section 5.1, which is the paper's most specific factual claim and its most exposed. The asymmetry argument rests on the four headed verdicts in Section 4.2 of Hu et al. (2025) and on the pass@1 figures quoted with them, read as cells of a relation rather than as points on one gradient. Section 5.1 states the objection to that reading and why it does not hold up, but the present analysis worked from the paper's text and its stated numbers, not from a reproduction of its result tables or its figures. The unresolved conflict recorded at the end of Section 5.1 - a statement that no single-domain model improved significantly out of domain alongside a statement that the knowledge-trained model gained significantly on mathematics - is itself a reason to read the tables before relying on any single sentence from that paper, including the ones this argument uses. If the asymmetry does not hold, Section 5.2's import of the task-affinity framing loses its motivating case; Sections 4 and 6 survive, but the paper's sharpest point does not. This is stated here rather than buried because it is the first thing a hostile reviewer should check.

A related concession belongs with it. Because Hu et al. (2025) advance a containment account of their own for the cell in question, the claim in Section 5.1 is narrower than the paper's title suggests on a fast reading. What is contradicted is the similarity gloss that the result's abstract sentence, its section title and its downstream citations carry, not the analysis inside the paper, which already treats the relation as directed. A reviewer who reads the title as accusing that paper of an error would be reading it

wrongly, and the title would be partly to blame.

The third is source quality. Of the 2026 preprints carrying weight in this paper, Yang et al. (2026) supplies the learnability-versus-transferability separation and the claim that mathematics ranks among the least transferable domains, and Maniparambil et al. (2026) supplies the best positive transfer demonstration. Neither is refereed as of this writing. Section 4's argument does not depend on either; Section 5.2's closing observation and Section 8's opening case do. Eight sources here are refereed at recognised venues - Nature (DeepSeek-AI et al., 2025), ICLR 2026 (Hu et al., 2025), NeurIPS 2025 (Yue et al., 2025; Stojanovski et al., 2025), COLM 2025 (Hochlehnert et al., 2025), AAAI 2026 (Wu, M. et al., 2025), EMNLP 2025 (Shrestha et al., 2025) and TMLR 2026 (Pan et al., 2026) - and five more are the classical transfer anchors, published in Machine Learning, CVPR, ICML, NeurIPS and IEEE/CAA Journal of Automatica Sinica. Of the remaining twenty-eight entries, twenty-seven are unreplicated preprints and one is a prior record of the author's own.

The fourth is that Section 6's elicitation argument may prove too much. If the source-side gain on Qwen2.5-Math is substantially elicitation, the same doubt applies to the puzzle-to-mathematics transfer results, which are also obtained on 7B models of comparable provenance. The paper does not claim those results are artefacts, and Wu, M. et al. (2025)'s leakage-free control is direct evidence that reward correctness does matter once contamination is removed. But an argument that undermines the source-side measurement undermines it symmetrically, and Section 8's credence in the positive results is therefore not fully consistent with Section 6's scepticism about the negative ones. Naming that tension is better than smoothing it.

Finally, this paper reports no new measurement. Its claims are about what published measurements mean and what they do not cover. A reader who wants the transfer question answered rather than reframed will not find an answer here, because on the evidence located there is not one to give.

12. Conclusion

The cross-domain RLVR literature is not as contradictory as it looks, and it is less informative than it looks. Four quantities travel under the word transfer - whether a domain gives, whether it receives, whether it survives, and whether joint training pays - and the debate's two sides are substantially measuring different ones. The reconciliation the field has reached for, structural similarity between source and target, is in tension with the directional findings of the leading negative result that is usually cited for it: as reported, the most similar available pairing is the one that fails and a cross-class pairing is one that succeeds, and a symmetric relation cannot produce that. Directed, asymmetric, sign-bearing transfer matrices are the ordinary object of study in multi-task learning and have been since Taskonomy; the RLVR literature has been arguing about scattered cells of a matrix nobody has computed.

Two things are not known and should be stated without hedging. Nobody has reported a full directed transfer matrix over a fixed domain set, a fixed protocol and more than one model family, so it is not currently possible to say whether the reported disagreements are about models or about instruments. And nobody has separated the share of a source-domain gain that is elicited pretraining behaviour from the share that is acquired capability, and then measured whether the two propagate differently - which means the field does not know whether it is studying a property of RLVR or a property of the corpora its base models were pretrained on.

Revision note

Version 1.1 (29 September 2026) replaces version 1.0. Passages that followed the wording of cited abstracts too closely are now either quoted, with their citation, or rewritten in new words, and a few

qualifiers the sources state (such as "may" and "often") were restored in the process. The argument, the references and the reported values are unchanged.

References

Belcamino, V., Attolino, N., Capitanelli, A., & Mastrogiovanni, F. (2026). On the Generalization Gap in LLM Planning: Tests and Verifier-Reward RL. arXiv:2601.14456. [PREPRINT]

Caruana, R. (1997). Multitask Learning. Machine Learning, volume 28, pp. 41-75. doi:10.1023/A:1007379606734

Cheng, Z., Hao, S., Liu, T., Zhou, F., Xie, Y., Yao, F., Bian, Y., Zhuang, Y., Dey, N., Zha, Y., Gu, Y., & Zhou, K. (2025). Revisiting Reinforcement Learning for LLM Reasoning from A Cross-Domain Perspective. arXiv:2506.14965. [PREPRINT]

Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q. V., Levine, S., & Ma, Y. (2025). SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. arXiv:2501.17161. [PREPRINT]

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., & Zhang, X. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Nature 645, 633-638. arXiv:2501.12948.

Fifty, C., Amid, E., Zhao, Z., Yu, T., Anil, R., & Finn, C. (2021). Efficiently Identifying Task Groupings for Multi-Task Learning. Advances in Neural Information Processing Systems (NeurIPS) 2021. arXiv:2109.04617.

Fu, T., Gu, J., Li, Y., Qu, X., & Cheng, Y. (2025). Scaling Reasoning, Losing Control: Evaluating Instruction Following in Large Reasoning Models. arXiv:2505.14810. [PREPRINT]

Hochlehnert, A., Bhatnagar, H., Udandarao, V., Albanie, S., Prabhu, A., & Bethge, M. (2025). A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Reproducibility. Conference on Language Modeling (COLM) 2025. arXiv:2504.07086.

Hu, C., Zhu, Y., Kellermann, A., Biddulph, C., Waiwitlikhit, S., Benn, J., & Kang, D. (2025). Breaking Barriers: Do Reinforcement Post Training Gains Transfer To Unseen Domains? International Conference on Learning Representations (ICLR) 2026. arXiv:2506.19733.

Huan, M., Li, Y., Zheng, T., Xu, X., Kim, S., Du, M., Poovendran, R., Neubig, G., & Yue, X. (2025). Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning. arXiv:2507.00432. [PREPRINT]

Lee, H., Abbasloo, S., Tack, J., & Shin, J. (2026). Beyond Correctness: Learning Robust Reasoning via Transfer. arXiv:2602.08489. [PREPRINT]

Li, Y., Pan, Z., Lin, H., Sun, M., He, C., & Wu, L. (2025). Can One Domain Help Others? A Data-Centric Study on Multi-Domain Reasoning via Reinforcement Learning. arXiv:2507.17512. [PREPRINT]

Liu, M., Diao, S., Hu, J., Lu, X., Dong, X., Zhang, H., Bukharin, A., Zhang, S., Zeng, J., Sreedhar, M. N., Shen, G., & Mosallanezhad, D. (2025). Scaling Up RL: Unlocking Diverse Reasoning in LLMs via Prolonged Training. arXiv:2507.12507. [PREPRINT]

Liu, Q., Zhang, S., Qin, G., Ossowski, T., Gu, Y., Jin, Y., Kiblawi, S., Preston, S., Wei, M., Vozila, P., Naumann, T., & Poon, H. (2025). X-Reasoner: Towards Generalizable Reasoning Across Modalities and Domains. arXiv:2505.03981. [PREPRINT]

Lu, B., Zhao, H., Sun, S., Peng, H., Ding, R., & Mei, H. (2025). Generalization of RLVR Using Causal Reasoning as a Testbed. arXiv:2512.20760. [PREPRINT]

Mahendrakar, P. (2026). 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

Maniparambil, M., Karuvally, A., Sejnowski, T., & Reid, F. (2026). When RL Suppresses Its Own Vocabulary: Recovering Reasoning Diversity in Puzzle-to-Math Transfer. arXiv:2605.29190. [PREPRINT]

Pan, P., Liang, Y., & Lin, S. (2026). Reward Modeling for Reinforcement Learning-Based LLM Reasoning: Design, Challenges, and Evaluation. Transactions on Machine Learning Research (TMLR) 2026. arXiv:2602.09305.

Pardinas, R., Kamalloo, E., Vazquez, D., & Drouin, A. (2026). Apriel-1.5-OpenReasoner: RL Post-Training for General-Purpose and Efficient Reasoning. arXiv:2604.02007. [PREPRINT]

Piao, S., & Park, S. (2026). GeneralThinker: Domain-General Reasoning through Likelihood-Guided Answer-

Conditioned Optimization. arXiv:2605.27934. [PREPRINT]

Rojas Nunez, J., Sawant, V., Allen, N., Amgalanbaatar, N., Zongo, Y., Sharma, V., & Chaudhary, M. (2026). Mechanistic origins of catastrophic forgetting: why RL preserves circuits better than SFT? arXiv:2605.28860. [PREPRINT]

Shao, R., Li, S. S., Xin, R., Geng, S., Wang, Y., Oh, S., Du, S. S., Lambert, N., Min, S., Krishna, R., Tsvetkov, Y., & Hajishirzi, H. (2025). Spurious Rewards: Rethinking Training Signals in RLVR. arXiv:2506.10947. [PREPRINT]

Shenfeld, I., Pari, J., & Agrawal, P. (2025). RL's Razor: Why Online Reinforcement Learning Forgets Less. arXiv:2509.04259. [PREPRINT]

Shrestha, S., Kim, M., Nepal, A., Shrestha, A., & Ross, K. (2025). Warm Up Before You Train: Unlocking General Reasoning in Resource-Constrained Settings. Empirical Methods in Natural Language Processing (EMNLP) 2025. arXiv:2505.13718.

Standley, T., Zamir, A. R., Chen, D., Guibas, L., Malik, J., & Savarese, S. (2020). Which Tasks Should Be Learned Together in Multi-task Learning? International Conference on Machine Learning (ICML) 2020. arXiv:1905.07553.

Stojanovski, Z., Stanley, O., Sharratt, J., Jones, R., Adefioye, A., Kaddour, J., & Kopf, A. (2025). Reasoning Gym: Reasoning Environments for Reinforcement Learning with Verifiable Rewards. Advances in Neural Information Processing Systems (NeurIPS) 2025. arXiv:2505.24760.

Wang, H., Long, X., Li, Z., Xu, Y., Li, T., & Tang, Y. (2026). To Mix or To Merge: Toward Multi-Domain Reinforcement Learning for Large Language Models. arXiv:2602.12566. [PREPRINT]

Wang, Y., Yang, Q., Zeng, Z., Ren, L., Liu, L., Peng, B., Cheng, H., He, X., Wang, K., Gao, J., Chen, W., & Wang, S. (2025). Reinforcement Learning for Reasoning in Large Language Models with One Training Example. arXiv:2504.20571. [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. [PREPRINT]

Wong, Z. H., Deng, J., He, R., Chen, Z., You, Q., Dong, H., Liang, H., Shen, C., Cui, B., & Zhang, W. (2025). LogicPuzzleRL: Cultivating Robust Mathematical Reasoning in LLMs via Reinforcement Learning. arXiv:2506.04821. [PREPRINT]

Wu, F., Xuan, W., Lu, X., Liu, M., Dong, Y., Harchaoui, Z., & Choi, Y. (2025). The Invisible Leash: Why RLVR May or May Not Escape Its Origin. arXiv:2507.14843. [PREPRINT]

Wu, M., Zhang, Z., Dong, Q., Xi, Z., Zhao, J., Jin, S., Fan, X., Zhou, Y., Lv, H., Zhang, M., Fu, Y., & Liu, Q. (2025). Reasoning or Memorization? Unreliable Results of Reinforcement Learning Due to Data Contamination. Association for the Advancement of Artificial Intelligence (AAAI) 2026. arXiv:2507.10532.

Wu, S., Yang, K., Cai, Y., Xu, X., Wang, P., Wang, J., Liu, J., Lyu, J., Chen, Y., Yang, S., & Xiao, Y. (2026). Consolidating RLVR Capabilities Across Domains: A Deep Dive into Fusion Paradigms. arXiv:2608.27409. [PREPRINT]

Xie, T., Gao, Z., Ren, Q., Luo, H., Hong, Y., Dai, B., Zhou, J., Qiu, K., Wu, Z., & Luo, C. (2025). Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning. arXiv:2502.14768. [PREPRINT]

Yang, Y., Liu, J., He, Y., Zhang, L., Scholkopf, B., & Jin, Z. (2026). Transferability for General Reasoning: An Automated Curriculum for Multi-Domain RLVR. arXiv:2606.25178. [PREPRINT]

Yu, T., Ji, B., Wang, S., Yao, S., Wang, Z., Cui, G., Yuan, L., Ding, N., Yao, Y., Liu, Z., Sun, M., & Chua, T. (2025). RLPR: Extrapolating RLVR to General Domains without Verifiers. arXiv:2506.18254. [PREPRINT]

Yuan, S., Chen, J., Zheng, J., Li, M., Feng, L., Wang, D., Xiang, T., Liu, T., & An, B. (2026). Understanding Diversity Collapse in RLVR via the Lens of Overtraining. arXiv:2606.15455. [PREPRINT]

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. arXiv:2504.13837.

Zamir, A., Sax, A., Shen, W., Guibas, L., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling Task Transfer Learning. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018. arXiv:1804.08328.

Zhang, W., Deng, L., Zhang, L., & Wu, D. (2022). A Survey on Negative Transfer. IEEE/CAA Journal of Automatica Sinica 2022. arXiv:2009.00909.

Zhang, X., Li, X., Wang, R., Miao, R., Wang, Z., Roth, D., & Li, C. (2026). Chart-RL: Generalized Chart Comprehension via Reinforcement Learning with Verifiable Rewards. arXiv:2603.06958. [PREPRINT]

Zhao, A., Wu, Y., Yue, Y., Wu, T., Xu, Q., Yue, Y., Lin, M., Wang, S., Wu, Q., Zheng, Z., & Huang, G. (2025). Absolute Zero: Reinforced Self-play Reasoning with Zero Data. arXiv:2505.03335. [PREPRINT]

(c) 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

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 API and Crossref 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, against the located passage. The author is responsible for the final text and for all claims made in it.

DOI: 10.5281/zenodo.22313640