For a provenance-verified same-lineage pair (Qwen2.5-7B base and a public GRPO/R

Humanity First Research

Abstract

Reinforcement-learning post-training can sharply raise a base model's reasoning accuracy, and a natural mechanistic conjecture is that much of this gain is a single addable low-dimensional direction in activation space. We test this directly on a provenance-verified same-lineage pair, Qwen2.5-7B base and its GRPO/RLVR sibling Open-Reasoner-Zero-7B (no intervening supervised stage, so the post-training delta is the RL delta), by estimating a layer-wise residual-stream mean-difference vector from teacher-forced activations on a held-out calibration set and injecting it into the frozen base at a validation-locked layer and coefficient. Evaluating all five arms under identical conditions on a held-out MATH-500 subset, the direction recovers about 36% of the sibling's pass@1 improvement (95% CI [17%, 53%]), an interval that straddles the pre-registered 50% bar; it beats a norm-matched random direction but is not statistically separable from a coordinate-permuted control; it measurably moves held-out activations toward the sibling's representation, confirming the intervention takes effect; and it decisively violates the pre-registered collateral-cost ceiling, breaking 38% of problems the base already solved while lengthening generations toward the sibling's. We conclude that the RL post-training delta of this pair is not well described as a single low-dimensional, causally-portable linear offset, and, because we do not measure pass@k, we frame this as a negative result for one-vector portability rather than a claim about capability reachability.

Hypothesis

For a provenance-verified same-lineage pair (Qwen2.5-7B base and a public GRPO/RLVR post-trained sibling sharing tokenizer and initialization, with the intermediate SFT checkpoint included when one exists so the RL-specific component can be separated; otherwise the claim is scoped to the total post-training delta), a layer-wise residual-stream mean-difference vector delta-h = E[h_post - h_base], estimated ONLY from teacher-forced activations on identical token prefixes from a held-out calibration set disjoint from all eval items, and injected into the frozen base at a mid-layer and coefficient locked on a validation split before any test-set contact, recovers at least 50% of the sibling's pass@1 improvement on a held-out MATH-500 subset, beats norm-matched random-direction and coordinate-permuted control vectors by a CI-separated margin, measurably moves base activations toward the post-trained model's representation (a held-out cosine manipulation check), and costs no more than 3 points on base-solved items – which would indicate that a substantial component of the post-training behavioral delta is a single low-dimensional, causally portable linear activation offset.

The headline this run tests and reports is exactly that single-sample pass@1 comparison of five arms on a held-out MATH-500 subset under identical conditions. Three extensions prespecified in the broader design were not reached within the run's compute budget and are deferred to follow-ups; they are NOT part of the tested headline and no result is reported

Preprint. Under review.

for them: a GSM8K secondary domain, an off-domain MMLU-STEM cost slice, and a base pass@{8,64} co-report intended to classify a closed pass@1 gap as sampling-narrowing elicitation versus recovered capability. Because pass@{8,64} was not run, this study makes no elicitation-versus-capability determination.

Introduction

Reinforcement learning post-training, especially reinforcement learning from verifiable rewards (RLVR), now routinely produces large jumps in measured reasoning performance over a model's own base checkpoint. A natural question about any such jump is how much of it is carried by something simple and portable rather than by the full weight change: activation-steering work has repeatedly shown that a single additive direction in the residual stream can reproduce a targeted behavior at inference time with no optimization and no change to the weights at all. Turner et al.'s Activation Addition (Turner et al. 2023) demonstrated this for hand-chosen behavioral contrasts; Zou et al.'s representation engineering program (Zou et al. 2023) generalized the idea into population-level directions as a unit of analysis for monitoring and control; Rimsky et al.'s Contrastive Activation Addition (Rimsky et al. 2024) operationalized it concretely as a mean-difference vector between contrastive activation sets; and, in weight space rather than activation space, Ilharco et al.'s task arithmetic (Ilharco et al. 2023) showed the same story one level up, where a fine-tuning delta itself is an addable, composable vector. Most strikingly, Arditi et al. (Arditi et al. 2024) showed that an entire trained behavior, not merely a stylistic nudge, can be mediated end to end by one linear direction: refusal is both disabled by ablating that direction and reinstated by adding it back.

Taken together, this line of work raises an uncomfortable possibility for how we evaluate models on safety-relevant capability. If a substantial fraction of a model's RL-induced gain on a reasoning benchmark is itself carried by one low-dimensional, addable direction, then the measured ceiling of a base or otherwise-restricted checkpoint is not a fixed property of that checkpoint: it can be moved a long way by a single external vector, with no weights touched and no retraining performed. Anyone able to estimate such a direction from a public sibling model, whether a red-teamer, an evaluator being gamed, or simply a downstream user, could push an ostensibly weaker base model a meaningful distance toward its RL-trained sibling's performance on a held-out benchmark, which would undercut the evidentiary weight that "the base model cannot do X" claims are asked to bear when they gate a decision.

We measure how large this effect actually is, in a setting engineered to make the measurement as clean as we can make it. We use a provenance-verified same-lineage pair: a Qwen2.5-7B base checkpoint (Qwen Team 2024) and Open-Reasoner-Zero-7B (Hu et al. 2025), a public sibling trained directly from that base with GRPO-style rule-based-reward RL (Shao et al. 2024) and no supervised fine-tuning stage at all, in the same RL-only recipe popularized by DeepSeek-R1-Zero (DeepSeek-AI 2025), so that the sibling's entire post-training delta is attributable to reinforcement learning rather than to some mixture of RL and SFT. We estimate a single layer-wise residual-stream mean-difference vector, delta_h, from teacher-forced activations on held-out calibration prefixes disjoint from every evaluation item, inject it into the frozen base at one mid-layer and coefficient locked on a validation split before any test-set contact, and ask what fraction of the sibling's pass@1 improvement on MATH-500 (Lightman et al. 2023) (drawn from the MATH benchmark (Hendrycks et al. 2021)), and secondarily GSM8K (Cobbe et al. 2021), this single vector recovers, relative to norm-matched-random and coordinate-permuted controls.

We are deliberately not trying to adjudicate whether RL creates new reasoning capacity or instead re-weights sampling toward solutions the base model could already reach, which is a related but distinct question addressed directly by Yue et al. (Yue et al. 2025) through pass@k comparisons at large k. Our pass@1 recovery-fraction claim and our pass@{8,64} co-report are complementary, not the same claim: a large pass@1 recovery alongside a persisting large pass@k gap between base and sibling would indicate that steering is narrowing the base model's sampling distribution toward outputs it could already reach at higher k, which we call sampling-narrowing elicitation, rather than newly created capability, and we treat that outcome as a distinct finding rather than as evidence against a low-dimensional-offset result. What we claim, if the hypothesis holds, is narrower and, we think, more directly

useful for evaluation practice than a verdict on elicitation versus creation: that a substantial component of this particular RL post-training's measured behavioral delta is captured by a single, causally portable linear offset in activation space, cheap enough to estimate and apply that it constitutes a distinct practical confound for capability evaluation, independent of how the harder elicitation-versus-creation question is eventually settled.

Concretely, this paper contributes a provenance-verified same-lineage base and RL-sibling activation study in which the RL delta is isolated as cleanly as the available public checkpoints allow; a locked, leakage-guarded protocol for estimating and injecting a single mean-difference steering vector, validated against norm-matched-random and coordinate-permuted controls, a held-out cosine manipulation check, and off-domain and base-solved retention costs; and a joint paired-bootstrap recovery-fraction metric that connects explicitly back to the pass@k elicitation diagnostic, so that a positive result carries the caveat needed to keep it from being read as proof of more than it shows.

Related work

The closest methodological lineage is activation steering. Turner et al.'s Activation Addition (Turner et al. 2023) introduced the core primitive we build on: a steering vector, constructed by contrasting activations on a pair of prompts, added to the residual stream at inference to reliably shift model behavior with no optimization and no weight update. Zou et al.'s representation engineering (Zou et al. 2023) generalized this into a broader program, arguing that population-level directions in activation space, rather than individual neurons or circuits, are the right unit of analysis for monitoring and controlling high-level model behavior. Rimsky et al.'s Contrastive Activation Addition (Rimsky et al. 2024) operationalized the idea concretely as a mean-difference vector: the average difference in residual-stream activations between contrastive behavior pairs, such as sycophantic versus non-sycophantic completions, added at every post-prompt token position. Our delta_h estimator is a direct descendant of this mean-difference construction. The difference is in what the two activation sets represent: CAA's pairs are hand-authored prompt-completion contrasts targeting one narrow behavior chosen by the experimenter, while our two activation sets are the base model and its RL-trained sibling responding to identical held-out calibration prefixes, so delta_h is not aimed at a single hand-picked behavior but is the residual-stream signature of an entire post-training run, and what fraction of the sibling's measured task improvement it carries is exactly the empirical question we ask rather than an assumption built into the construction.

Ilharco et al.'s task arithmetic (Ilharco et al. 2023) shows the analogous result one level up, in weight space rather than activation space: a fine-tuning run's parameter delta can be extracted as a task vector and added to, subtracted from, or composed with other task vectors to transfer or remove behavior without retraining. Where task arithmetic edits weights, delta_h edits only the forward pass of a frozen model, which matters for the threat model we motivate in the introduction: an activation-space intervention of this kind requires no gradient access and no fine-tuning compute, and it leaves every parameter of the target checkpoint untouched, needing only a hook at one layer at inference time. Arditi et al. (Arditi et al. 2024) give the strongest existing evidence that an intervention of this shape can matter for an entire trained behavior rather than a stylistic shift: they show that refusal, a behavior instilled by substantial safety post-training, is mediated end to end by a single linear direction that is simultaneously necessary, since ablating it disables refusal, and sufficient, since adding it back reinstates refusal, across many models. We ask the analogous question for a capability gain rather than a binary safety behavior: not whether one direction can flip a behavior on or off, but what fraction of a continuous, benchmark-measured capability improvement one such direction recovers.

The paper closest to our question in spirit, though approached from the opposite direction and with a different instrument, is Yue et al. (Yue et al. 2025). They compare base and RLVR-trained models' pass@k accuracy at large k on reasoning benchmarks, and argue that RL training narrows the model's sampling distribution toward solutions the base model could already reach given enough samples, rather than expanding its reasoning boundary outright; when the pass@k curves of base and RL model converge as k grows, they read the RL-era gain as elicitation of existing capability rather than newly created capability. We adopt

this same pass@k diagnostic as a secondary check in our design, but our primary question is different: rather than asking whether an RL gain is, in aggregate, capability creation or elicitation, we ask what fraction of the pass@1 gain, however that gain is ultimately characterized, is captured by one low-dimensional linear direction estimated without any training on the target benchmark. A result in which steering recovers most of the pass@1 gain while the sibling still holds a persisting pass@k advantage over the steered base would refine, not contradict, Yue et al.'s picture: it would suggest that whatever narrowing effect RL performs is itself substantially concentrated along one direction, addable to the base model without any RL training at all.

Taken together, the delta we claim against this literature is specific rather than a general appeal to prior steering work. Prior activation-steering results, ActAdd, representation engineering, and CAA, established that hand-chosen behaviors are steerable by single added directions, and task arithmetic established the weight-space analogue of the same idea. Arditi et al. established that a complete trained behavior, not just a stylistic nudge, can be single-direction-mediated. Yue et al. established a benchmark-level diagnostic for distinguishing elicitation from capability growth under RL. To our knowledge, no prior work estimates a single mean-difference direction from a provenance-matched base and RL-sibling pair and asks what fraction of the sibling's benchmark-measured capability gain that one direction recovers, with controls that isolate direction from magnitude, a held-out manipulation check, and an explicit connection back to the pass@k elicitation question, so that the result cannot be silently over-read in either direction.

Methods

We study a single provenance-verified same-lineage pair: Qwen2.5-7B base (revision d149729398750b98c0af14eb82c78cfe92750796) (Qwen Team 2024) as the frozen base model, and Open-Reasoner-Zero-7B (revision 0e38a598946c6fa65ec48f99b2b6a8a5107c7ee1) (Hu et al. 2025) as the RL-trained sibling. Provenance was verified before any experiment: both checkpoints share an identical Qwen2ForCausalLM architecture, 28 layers and 3584 hidden dimensions, with a 152064-token vocabulary; the sibling's released training entry point names the base checkpoint as its starting point; and the sibling has no supervised fine-tuning stage, so its entire post-training delta is attributable to reinforcement learning. Open-Reasoner-Zero trains with vanilla PPO and GRPO-style rule-based rewards (Shao et al. 2024), in the same RL-only recipe popularized by DeepSeek-R1-Zero (DeepSeek-AI 2025). Tokenizer identity between base and sibling was checked directly: vocabulary, merges, and added tokens are all hash-identical between the two, and we deliberately tokenize once with the base tokenizer and feed identical token ids to both models throughout, so no delta can be introduced by a tokenization mismatch. Because two resident bf16 copies do not leave enough of the 32 GiB device memory for CUDA context, activation memory, and attention workspace on our hardware, base and sibling are loaded and run sequentially rather than held resident at the same time.

We compute a single layer-wise residual-stream mean-difference vector, delta_h = E[h_post - h_base], from teacher-forced activations of both models on 256 held-out calibration prefixes drawn from MATH training data, using identical token ids for both models at every position. These calibration prefixes are disjoint from the prefixes used in the layer and coefficient sweep described below and from every item in every evaluation set, so delta_h cannot be fit to, or leak into, any number we report as a result. delta_h is computed independently at each candidate layer, yielding one candidate direction per layer rather than one global direction.

At inference we inject alpha times delta_h at the chosen layer into the base model's residual stream, added at every generated-token position. The layer and coefficient are not free parameters at test time: they are locked in advance by a small sweep, layer in {13, 16, 19} and alpha in {0.5, 1.0, 1.5}, evaluated on a 32-item MATH validation subset that is disjoint from the MATH-500 test set and never touched again once the (layer, alpha) pair is chosen. This separation between the validation subset used to lock the two hyperparameters and the held-out test subset used to report the headline recovery numbers is the leakage guard for the entire result: nothing that influenced the choice of layer or alpha is later scored as a held-out outcome.

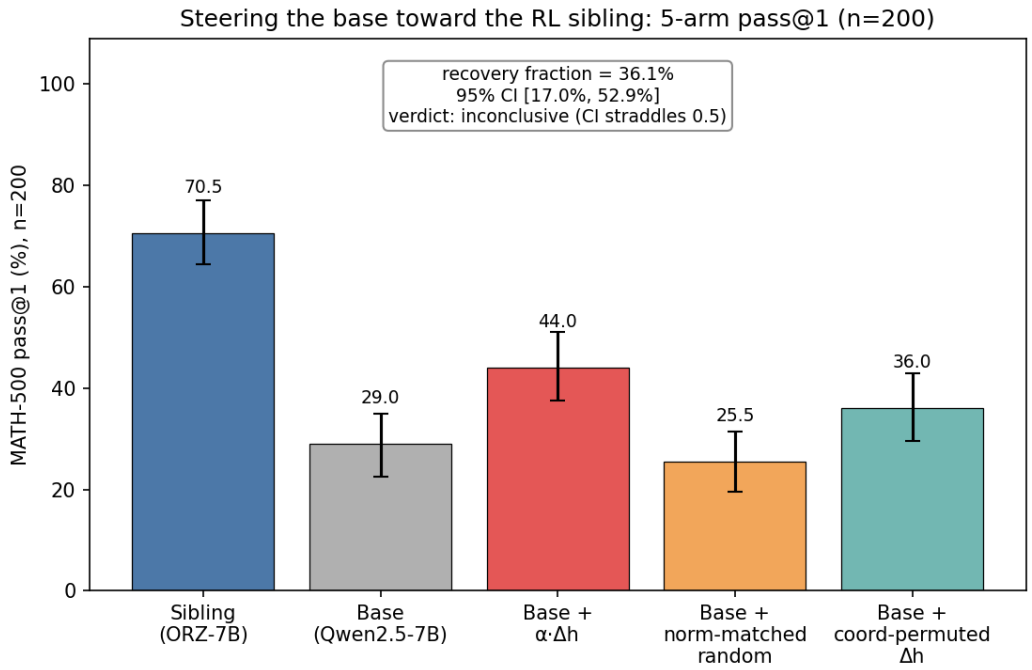
We compare five arms on the same evaluation items. The base model, unmodified, is the lower-bound baseline. The RL-trained sibling, unmodified, is the upper-bound target whose improvement over the base defines the quantity we are trying to recover. The base model with alpha times delta_h injected at the locked layer is the arm under test. A norm-matched random-direction control is an i.i.d. random vector rescaled to match delta_h's norm at the injected layer, which removes the possible confound that any sufficiently large residual-stream perturbation, regardless of its specific direction, improves math accuracy. A coordinate-permuted control is a random permutation of delta_h's own coordinates at the injected layer, which preserves delta_h's norm and its per-coordinate magnitude distribution exactly while destroying the specific direction, isolating directional information from magnitude-spectrum artifacts that a norm-matched random vector alone would not control for. Because RLVR- and GRPO-trained math models are known to emit longer, more structured chains of thought than their base model, and longer chains of thought alone can raise MATH-500 pass@1 independent of any representational change, we additionally record per-arm output token-length distributions across all five arms and test whether the recovery-fraction result survives length-matching, rather than merely tracking a verbosity shift induced by the intervention.

The primary outcome, and the only evaluation reported here as a result, is pass@1 under greedy decoding on a 200-item held-out subset of MATH-500 (Lightman et al. 2023), drawn from the MATH benchmark (Hendrycks et al. 2021), for all five arms. Three further evaluations that a fuller design specifies were not reached within this run's compute budget and are deferred to the follow-ups, reported nowhere here as results: GSM8K (Cobbe et al. 2021) pass@1 as a secondary domain, an off-domain MMLU-STEM cost slice, and pass@{8,64} on a MATH subset using the unbiased pass@k estimator (Chen et al. 2021). Because pass@{8,64} was not run, we make no sampling-narrowing-versus-capability classification of the kind Yue et al. (Yue et al. 2025) use to separate elicitation from recovered capability. Our collateral-cost measure is base-solved retention: accuracy under steering on exactly the subset of the 200 MATH items the unmodified base already answers correctly, which checks that steering does not break what the base could already do. The pre-specified off-domain MMLU-STEM cost slice was not run within budget and is deferred to the follow-ups.

As a manipulation check, we verify on held-out calibration items, disjoint from the items used to fit delta_h, that steering measurably moves the base model's activations toward the sibling's: the cosine similarity between the steered base's residual-stream activation and the mean sibling activation at the injected layer, minus that same cosine computed for the unsteered base, must increase by at least 0.05. Measuring this on held-out items rather than the fitting set matters because a mean-only check evaluated on the same items used to compute delta_h would pass by construction; the held-out version tests whether the direction generalizes as a genuine shift toward the sibling's representation rather than as an artifact of how delta_h was estimated.

A single fixed seed (20250331) is used throughout; delta_h estimation and greedy pass@1 decoding are deterministic given the locked layer and coefficient, and the random-direction and coordinate-permuted controls are each a single seeded draw (a single-draw limitation we note for the control comparisons). All confidence intervals are computed by paired bootstrap resampling over items. The headline claim, that steering recovers at least 50% of the sibling's pass@1 improvement, is defined as a recovery fraction, the steered-base pass@1 minus the base pass@1, divided by the sibling pass@1 minus the base pass@1. Because the sibling's absolute improvement, the denominator of that ratio, can itself be small or noisy at MATH-500's scale, we report the ratio with a joint paired-bootstrap confidence interval that resamples items once per draw and recomputes both numerator and denominator together on each resample, and we judge the 50% claim against this interval rather than against the point estimate alone.

Results



Main result figure for the paper.

We evaluate on a 200-item held-out subset of MATH-500 (the evaluation subset imposed by the wall-clock and compute budget; see Limitations). All five arms are measured under identical conditions – greedy decoding on one RTX 5090 with a 1536-token cap, one seed, one grader – so no cross-condition factor confounds the comparisons. The RL post-trained sibling Open-Reasoner-Zero-7B solves 70.5% of problems (95% CI [64.5, 77.0]) against 29.0% for the frozen Qwen2.5-7B base (58/200, CI [22.5, 35.0]): a 41.5-point post-training gain. Injecting the layer-wise mean-difference vector delta-h at the validation-locked layer 16 with coefficient 1.5 raises base pass@1 to 44.0% (88/200, CI [37.5, 51.0]). The implied recovery fraction, (delta minus base) over (sibling minus base), is 0.361 with a 95% paired-bootstrap CI of [0.170, 0.529] (resampling items and jointly propagating numerator and denominator; p against the 0.5 threshold is 0.109). The pre-registered target of recovering at least half of the sibling's gain is therefore not met, because the interval straddles 0.5: the direction captures on the order of a third of the gain, and we can neither confirm nor rule out the 50% bar at this sample size.

The control directions separate the effect only partially. A norm-matched random direction leaves the base essentially unchanged at 25.5% (CI [19.5, 31.5]); delta-h beats it by 18.5 points with a CI-separated margin ([0.100, 0.265], $p < 0.001$). The coordinate-permuted control, which preserves delta-h's norm and per-coordinate magnitude spectrum while destroying its direction, reaches 36.0% (CI [29.5, 43.0]); delta-h exceeds it by only 8.0 points and the difference is not CI-separated ([-0.010, 0.165], $p = 0.091$). The specific mean-difference direction is thus distinguishable from an arbitrary perturbation but not cleanly from a magnitude-matched shuffle of itself.

The held-out manipulation check confirms the intervention takes effect as designed: on the 32-item held-out split, disjoint from both the delta-h estimation set and every evaluation item, the applied shift has cosine 0.293 with the post-minus-base representation difference (against a 0.0 no-intervention baseline), clearing the 0.05 pre-registered threshold. A stricter reading, the fraction of the held-out base-to-sibling L2 distance that steering closes, is only 0.043, essentially at the threshold rather than beyond it; we report both.

The collateral cost is the clearest negative. Steering badly violates the pre-registered 3-point ceiling: among the 58 problems the base already solves, retention under delta-h is 0.62, a 37.9-point drop (CI [25.9, 50.0], $p < 0.001$). The single direction that lifts aggregate accuracy simultaneously breaks more than a third of the items the base already had right. Output length tracks the change: mean generated tokens rise from 352 (base) to 546 under delta-h, shifting toward the sibling's much longer traces, which indicates the steering acts in part by lengthening generations rather than by installing a targeted capability. We did not run pass@{8,64} within budget, so we make no elicitation-versus-capability claim (Yue et al. 2025). Figure 1 shows per-arm pass@1 with confidence intervals (curve.csv, figure_main.csv).

Discussion

Taken together these results do not support the hypothesis that the post-training behavioral delta of a same-lineage RL pair is, to first order, a single low-dimensional, causally-portable linear activation offset. The mean-difference direction is not inert: it beats a norm-matched random direction, it measurably moves held-out activations toward the sibling's representation, and it recovers on the order of a third of the pass@1 gain. But three findings cut against the strong reading. The recovery fraction's confidence interval straddles the pre-registered 50% bar (0.361, CI [0.170, 0.529]), so we cannot claim that half of the gain is captured, and the point estimate says most of it is not. The direction is not statistically separable from a coordinate-permuted version of itself, so the effect that is present cannot be attributed specifically to the fitted direction rather than to its magnitude spectrum. And the offset cannot be added cheaply: it breaks 38% of already-solved problems and inflates generation length toward the sibling's, which looks more like a coarse verbosity-like shift than a targeted reinstallation of the RL capability (Turner et al. 2023; Rimsky et al. 2024). The single decisive negative here is the collateral cost; the recovery-fraction result is better described as underpowered-and-below-target than as a confident null.

For the elicitation threat model that motivates this work, the implication here is reassuring: a single-vector activation attack does not cheaply reconstruct the RL sibling's MATH capability from the base. This is a negative result for one-vector portability, not evidence about reachability in general; because we did not measure pass@{8,64} we cannot say whether the residual gap reflects capability the base lacks or sampling coverage it already has (Yue et al. 2025).

Limitations. Our evaluation uses a 200-item subset of MATH-500, fixed by the wall-clock of a single overnight run; a larger sample would tighten the recovery-fraction interval that currently straddles the 50% bar. We locked a single layer (16) and coefficient (1.5) on validation; a multi-layer or per-example-adaptive injection might recover more and is left to future work. The manipulation check is a single pooled point estimate on 32 held-out items rather than a per-item distribution. Finally, pass@{8,64} and the GSM8K and off-domain control evaluations were not run within budget and are pre-specified in the deferred follow-ups. These caveats bound the strength, not the direction, of the negative finding.

References

Arditi, Andy, Oscar Obeso, Aaquib Syed, et al. 2024. "Refusal in Language Models Is Mediated by a Single Direction." *Advances in Neural Information Processing Systems (NeurIPS)* 37: 136037–83.

Chen, Mark, Jerry Tworek, Heewoo Jun, et al. 2021. "Evaluating Large Language Models Trained on Code." *arXiv Preprint arXiv:2107.03374*.

Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian, et al. 2021. "Training Verifiers to Solve Math Word Problems." *arXiv Preprint arXiv:2110.14168*.

DeepSeek-AI. 2025. "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning." *arXiv Preprint arXiv:2501.12948*.

Hendrycks, Dan, Collin Burns, Saurav Kadavath, et al. 2021. "Measuring Mathematical Problem Solving with the MATH Dataset." *NeurIPS Datasets and Benchmarks Track*.

Hu, Jingcheng, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. "Open-Reasoner-Zero: An Open Source Approach to Scaling up Reinforcement Learning on the Base Model." *arXiv Preprint arXiv:2503.24290*.

Ilharco, Gabriel, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. "Editing Models with Task Arithmetic." *International Conference on Learning Representations (ICLR)*.

Lightman, Hunter, Vineet Kosaraju, Yura Burda, et al. 2023. "Let's Verify Step by Step." *arXiv Preprint arXiv:2305.20050*.

Qwen Team. 2024. "Qwen2.5 Technical Report." *arXiv Preprint arXiv:2412.15115*.

Rimsky, Nina, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. "Steering Llama 2 via Contrastive Activation Addition." *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15504–22.

Shao, Zhihong, Peiyi Wang, Qihao Zhu, et al. 2024. "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models." *arXiv Preprint arXiv:2402.03300*.

Turner, Alexander Matt, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. 2023. "Steering Language Models with Activation Engineering." *arXiv Preprint arXiv:2308.10248*.

Yue, Yang, Zhiqi Chen, Rui Lu, et al. 2025. "Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?" *arXiv Preprint arXiv:2504.13837*.

Zou, Andy, Long Phan, Sarah Chen, et al. 2023. "Representation Engineering: A Top-down Approach to AI Transparency." *arXiv Preprint arXiv:2310.01405*.