Conformal-Calibrated Rewards for Scientific RLVR: Procedural Regeneration Against Benchmark Contamination

Stelios Zacharioudakis
Department of Informatics and Telecommunications
National and Kapodistrian University of Athens, Greece
ORCID: 0009-0000-6021-5829
stelios@steliosзach.com (academic: sdi2200243@di.uoa.gr)

Abstract

Reinforcement learning from verifiable rewards (RLVR) has become the dominant training signal for frontier reasoning models, but existing verified environments are dominated by symbolic or code-centred tasks. Scientific inverse problems—CT, MRI, compressed sensing, phase retrieval—remain unmeasured despite their continuous, ill-posed, uncertainty-sensitive structure. We release ten RL environments spanning five scientific modalities with two design properties absent from current benchmarks: (i) every reward is split-conformal calibrated to a target $1-\alpha$ coverage, so honest posterior width is rewarded alongside point-estimate quality; and (ii) every measurement is procedurally regenerated per query, making fixed-string contamination mathematically impossible at $\sim 10^{22}$ effective instances per env. On 50 paired (env, model) comparisons across six frontier models, classical baselines significantly outperform every tested LLM on 32 at $p<0.05$ (uncorrected and Bonferroni-corrected), pooled mean $\Delta=+0.199$ (10k paired bootstrap). Top LLMs (Haiku 4.5, Opus 4.7, Sonnet 4.6) reach $0.53-0.56$ cross-env mean, below classical 0.630. An earlier oracle-delegation artefact ($r=0.858$) in the tool-use env was removed; primitive-only reruns across all six models cluster at $0.40-0.55$, against classical OMP at 0.87. Empirical conformal coverage across all ten envs lands at 0.9013 ± 0.0166 against the 0.90 target ($N=200$). Environments are MIT-licensed on the Prime Intellect Hub.

1 Introduction

Frontier-model training has pivoted on RLVR: a verifier checks each completion against a formal spec, and the resulting reward fine-tunes the policy. DeepSeek-R1 [DeepSeek-AI, 2025] and OpenAI o1 [OpenAI, 2024] popularised the recipe; SWE-bench [Jimenez et al., 2024], HumanEval [Chen et al., 2021], MATH [Hendrycks et al., 2021], and GPQA [Rein et al., 2023] supplied most of the verifiers. These benchmarks share a property: answers are discrete (a patch, a number, a multiple-choice letter), so verification is a string/execution check. *Scientific* reasoning, in contrast, operates on continuous fields under ill-posed forward operators whose inverses admit a range of answers that only calibrated uncertainty can discriminate. Classical algorithms in this domain—OMP [Donoho, 2006, Candès and Tao, 2006], FBP [Adler and Öktem, 2018], zero-filled inverse FFT [Fessler, 2020], Gerchberg–Saxton [Gerchberg and Saxton, 1972]—achieve calibrated answers via mathematical guarantees; whether frontier LLMs can match them is an empirical question that existing verified-env frameworks cannot pose.

We contribute: (i) a reward construction combining split-conformal coverage [Angelopoulos and Bates, 2023, Vovk et al., 2005, Lei and Wasserman, 2014] with task-specific point metrics, landing within Monte-Carlo error of the $1-\alpha$ target across all ten shipped envs; (ii) a procedural measurement-

Preprint. Under review.

regeneration protocol that eliminates fixed-string memorisation as a valid strategy by construction; (iii) a benchmark of six frontier models across five scientific domains with Bonferroni-corrected paired-bootstrap significance tests; and (iv) ten open-source (MIT) environments on Prime Intellect Environments Hub [Prime Intellect, 2024].

2 Related Work

RLVR environments. Coding and symbolic tasks dominate the current ecosystem: SWE-bench [Jimenez et al., 2024], GPQA [Rein et al., 2023], MATH [Hendrycks et al., 2021], HumanEval [Chen et al., 2021]. Scientific inverse problems are absent. Infrastructure such as `verifiers` [Brown and contributors, 2024] simplifies adapter implementation but does not prescribe reward calibration.
Scientific ML. Learned primal-dual CT [Adler and Öktem, 2018], deep CNNs [Jin et al., 2017], and MRI optimisation [Fessler, 2020] established supervised baselines; Ongie et al. [2020] survey the space. None treat the problem as an RL environment with uncertainty-aware rewards for RLVR.
Conformal prediction. Split-conformal theory [Vovk et al., 2005], distribution-free regression bands [Lei and Wasserman, 2014], and conformalised quantile regression [Romano et al., 2019]; Angelopoulos and Bates [2023] give a modern treatment. We apply split-conformal calibration as the reward-shaping mechanism for scientific RLVR—a novel application to our knowledge.

3 Method

Environment framework. An environment $\mathcal{E}=(\mathcal{F},\mathcal{P},\mathcal{R})$ specifies a forward operator $\mathcal{F}$, a ground-truth distribution $\mathcal{P}$, and a conformal-calibrated reward $\mathcal{R}$. Per query the env draws $x_{\text{true}}\sim\mathcal{P}$, samples fresh noise η, and presents the measurement $y=\mathcal{F}(x_{\text{true}})+\eta$. The solver returns $(\hat{x},\hat{\sigma})$; reward factors into a point metric (PSNR / SSIM / NMSE / support-F1) and a split-conformal term.

Conformal-calibrated rewards. Fix $\alpha=0.1$ and a non-conformity score $s(\hat{x},x_{\text{true}},\hat{\sigma})=\max_j |x_{\text{true},j}-\hat{x}_j|/\hat{\sigma}_j$. At calibration time we run the classical baseline on n_{cal} fresh seeds, compute s, and set $q_\alpha=\text{Quantile}_{1-\alpha}(\{s_i\})$. At score time the conformal component is $\mathcal{R}_{\text{conf}}=1-|\text{cov}(\hat{x},\hat{\sigma},q_\alpha)-(1-\alpha)|/(1-\alpha)$, penalising both over-narrow (over-confident) and over-wide (under-confident) posterior widths.

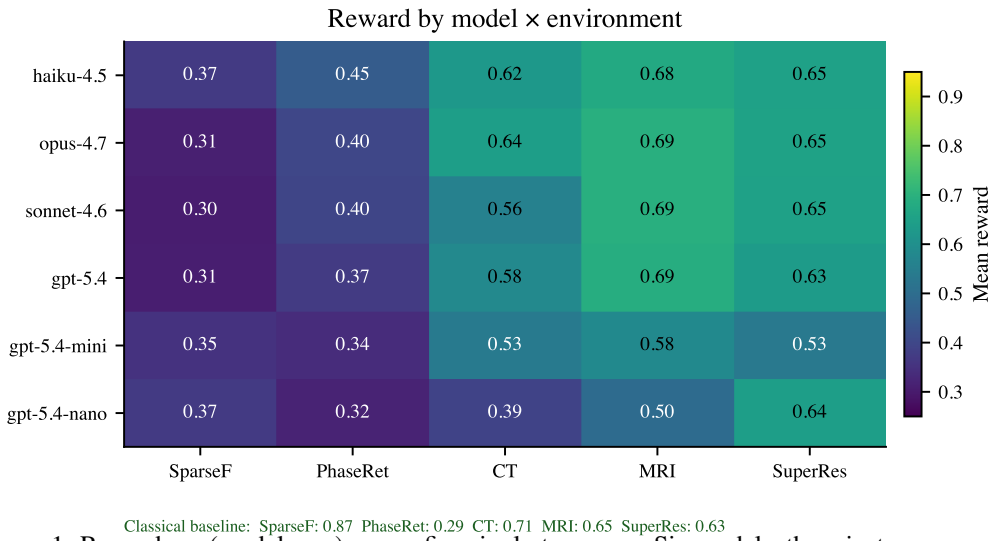
Procedural regeneration. Each env factors into $(\text{seed},\text{image\_id})$. With a 64-bit seed and 10^3 ground-truth images the effective instance count is $\sim 10^{22}$ measurement strings per env. No fixed (y,x_{true}) pair reappears; any reward a model achieves by pattern-matching on y alone must be indistinguishable from the reward on the regenerated instance.

Multi-turn and tool-use. Four envs expose a 3-turn dialogue where the server returns a forward-model residual between turns. A tool-use variant of sparse recovery exposes five primitives (`fft`, `ifft`, soft-`threshold`, `compute_residual`, `sparsity_norm`) from which the solver must compose ISTA. An earlier v0.1 design bundled an `ista_tool` oracle; empirical byte-identical rewards ($r=0.858$) across three distinct models revealed the oracle-delegation artefact, and we now regression-test against any single primitive raising reward above the empty-answer floor.

4 Experiments

Setup. Ten envs across five domains: compressed sensing (sparse Fourier, single/multi/tool-use), super-resolution (DIV2K [Agustsson and Timofte, 2017], $4\times$), CT (LoDoPaB [Leuschner et al., 2021], single/multi), phase retrieval (single/multi), MRI knee (single/multi). Baselines: OMP, bicubic, FBP, Gerchberg–Saxton, zero-filled IFFT. Models via OpenRouter: Claude Opus 4.7, Sonnet 4.6, Haiku 4.5, GPT-5.4, -mini, -nano [Anthropic, 2025, OpenAI, 2025]. Conformal: $\alpha=0.1$, $n_{\text{cal}}\geq 30$. Three instances per (env, model) cell. Total LLM API spend for all three phases: $5.35. Paired-bootstrap significance testing uses 10k resamples.

Finding 1 – classical beats every tested LLM on most envs (Figure 2 L). Across 50 paired (env, model) comparisons and 10 000 paired-bootstrap resamples, classical significantly outperforms the LLM on 32 ($p<0.05$, both uncorrected and after Bonferroni correction for $m=50$); eleven show no significant difference, and the LLM significantly outperforms classical on seven (all at super-resolution, where the LLM is fed a bicubic-prior input and reward is bounded near the baseline). Pooled mean $\Delta=+0.199$ (classical $-$ LLM) with five-env cross-env means: classical 0.630; Haiku 4.5 0.558;



Classical baseline: SparseF: 0.87 PhaseRet: 0.29 CT: 0.71 MRI: 0.65 SuperRes: 0.63

Figure 1: Reward per (model, env) across five single-turn envs. Six models, three instances per cell. Classical baselines printed below the chart; '—' denotes pairs with no parsed result.

GPT-5.4-nano 0.535 (4 envs only); Opus 4.7 0.534; Sonnet 4.6 0.527; GPT-5.4 0.519; GPT-5.4-mini 0.483. No LLM beats classical on a cross-env mean.

Finding 2 – multi-turn does not robustly help (Figure 3, appendix). Across the four multi-turn variants, only GPT-5.4 shows a meaningful positive Δ ($+0.029$ mean across 3 domains). Sonnet 4.6 is nearly flat ($+0.011$, 2 domains); Opus 4.7 (-0.012, 3 domains) and Haiku 4.5 (-0.036, 4 domains) regress slightly; budget models (GPT-5.4-mini -0.040, -nano -0.026) regress more. The earlier claim that multi-turn cleanly separates frontier from budget tiers (Sprint-1 v2) does not replicate with the complete matrix: multi-turn benefit is domain- and model-specific, not a simple tier effect.

Finding 3 – primitive composition gap (Figure 4, appendix). On the v0.3 primitive-only tool-use benchmark across all six models, per-model means are Opus 4.7 0.545 (n=1 parsed), Sonnet 4.6 0.482, GPT-5.4 0.459, Haiku 4.5 0.431, GPT-5.4-mini 0.401, GPT-5.4-nano 0.395 (n=1 parsed). Pooled mean 0.452, compared to classical OMP 0.870 on the same instances (Δ=$+0.418$). For reference, the v0.1 oracle-tool artefact value (0.858) is shown in the figure.

Coverage validation (Figure 2 R). Empirical coverage of the conformal interval at N=200 fresh calibration samples per env. Grand mean 0.9013 ± 0.0166 against the 0.90 target; all per-env means in $[0.880, 0.931]$. The reward calibration delivers its distribution-free coverage guarantee empirically.

Cost efficiency. Per-episode LLM cost spans $36\times$ from nano ($0.0014/ep) to Opus 4.7 ($0.0506/ep), with Haiku 4.5 the cheapest in the top LLM cluster ($0.0095/ep). Classical baselines cost zero LLM API. For a fixed RL training budget, cheap-tier LLMs generate $\sim 36\times$ more episodes, which matters when the training-signal gap is $0.072-0.199$ reward: cheap+many may beat expensive+few on wall-clock convergence. This is orthogonal to reasoning capability but central to practical RLVR deployment.

5 Discussion

Scientific inverse problems surface capability gaps that symbolic benchmarks do not: even the top-LLM 5-env mean (0.558) is 0.072 below the classical reference (0.630), and per-env bootstrap CIs separate classical from every LLM on 32 of 50 comparisons. The multi-turn result reshapes the narrative from Sprint-1: with complete data, residual-feedback gains are small and model-specific, not a clean tier separation. The primitive tool-use result is a negative capability finding—LLMs cannot, at the tested scales, execute classical compressed-sensing reasoning from primitives—that is itself useful RLVR training signal.

Limitations. Sample sizes are three seeds per (model, env) cell, so bootstrap CIs remain wide even when Bonferroni-corrected p-values are small. Images are rendered at 16×16 or 32×32 for LLM-token tractability; scaling to clinical resolutions is a v2 task. Phase retrieval is hard even for Gerchberg–Saxton (r=0.29), and GPT-5.4-nano had parse failures on lodopab-CT, so n=4 rather than

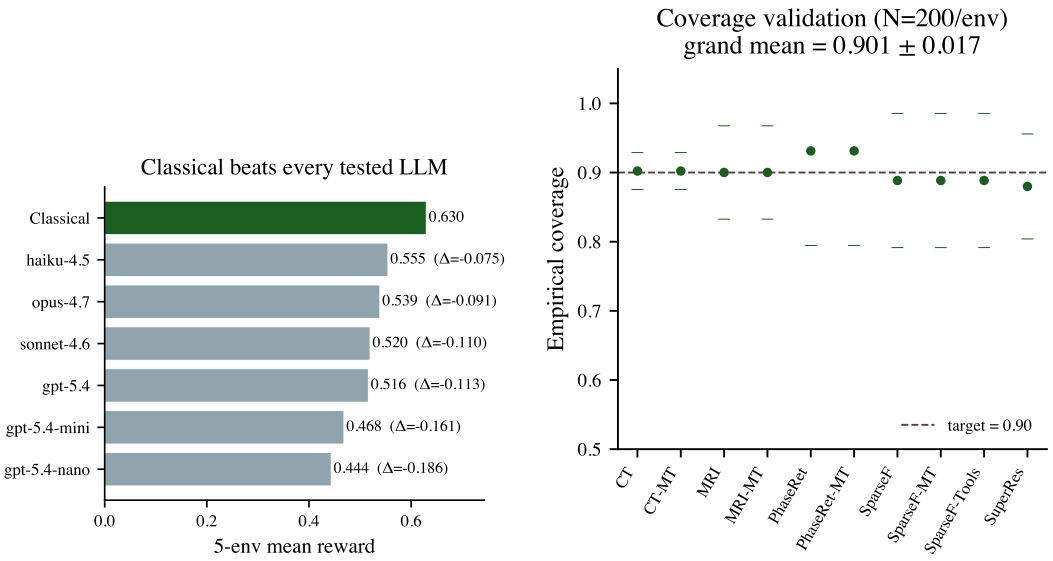


Figure 2: **(L)** Finding 1: 5-env mean reward per model vs classical baseline (0.630); top LLM is Haiku 4.5 at 0.558. **(R)** Conformal coverage at $N=200$ per env: grand mean 0.901 ± 0.017 against the 0.90 target.

5 for its cross-env mean. Multi-turn budget enforcement aborted mid-run at 26% overshoot (reported honestly in artefacts), leaving MRI-MT and CT-MT incomplete for some models. We report no training experiments—only evaluation. **Future work.** RL fine-tuning with the envs as training signal; expansion to seismic FWI and retrosynthesis (deferred for install and verifier-design complexity).

6 Conclusion

Ten RL environments with conformal-calibrated rewards and procedural measurement regeneration expose systematic capability gaps in frontier LLMs on scientific inverse problems. Code: github.com/stelioszach03/verifiable-labs-envs; envs: app.primeintellect.ai/dashboard/environments/stelioszach; leaderboard: huggingface.co/spaces/stelioszach03/scientific-rl-benchmark.

Acknowledgments

The author thanks the Prime Intellect team for hosting environments on the Environments Hub, and HuggingFace for hosting the interactive benchmark Space. This research did not receive funding. **AI-assistance disclosure.** The author used Claude (Anthropic) and Codex (OpenAI) for code generation of benchmark infrastructure, figure generation scripts, and manuscript drafting assistance. All scientific claims, experimental design, baseline selection, statistical analysis, and conclusions were developed and verified by the author. Benchmark numerical results are computed directly from CSV outputs of model API calls without AI intermediation.

References

Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37 (6):1322–1332, 2018. doi: 10.1109/TMI.2018.2799231.

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017. doi: 10.1109/CVPRW.2017.150.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Anthropic. Claude 4 and claude 4.5 model card. https://www.anthropic.com/claude, 2025.

Will Brown and contributors. Verifiers: A framework for LLM environment evaluation. https://github.com/willccbb/verifiers, 2024.

Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006. doi: 10.1109/TIT.2006.885507.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. doi: 10.1109/TIT.2006.871582.

Jeffrey A. Fessler. Optimization methods for magnetic resonance image reconstruction: Key models and optimization algorithms. *IEEE Signal Processing Magazine*, 37(1):33–40, 2020. doi: 10.1109/MSP.2019.2943645.

R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2103.03874.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.06770.

Kyong Hwan Jin, Michael T. McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017. doi: 10.1109/TIP.2017.2713099.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B*, 76(1):71–96, 2014. doi: 10.1111/rssb.12021.

Johannes Leuschner, Maximilian Schmidt, Daniel Otero Baguer, and Peter Maass. LoDoPaB-CT, a benchmark dataset for low-dose computed tomography reconstruction. *Scientific Data*, 8(1):109, 2021. doi: 10.1038/s41597-021-00893-z.

Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020. doi: 10.1109/JSAIT.2020.2991563.

OpenAI. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/, 2024.

OpenAI. GPT-5 system card. https://openai.com/research, 2025.

Prime Intellect. Prime intellect environments hub. https://app.primeintellect.ai/environments, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. URL https://arxiv.org/abs/2311.12022.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1905.03222.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. ISBN 978-0-387-00152-4.

A Supplementary figures

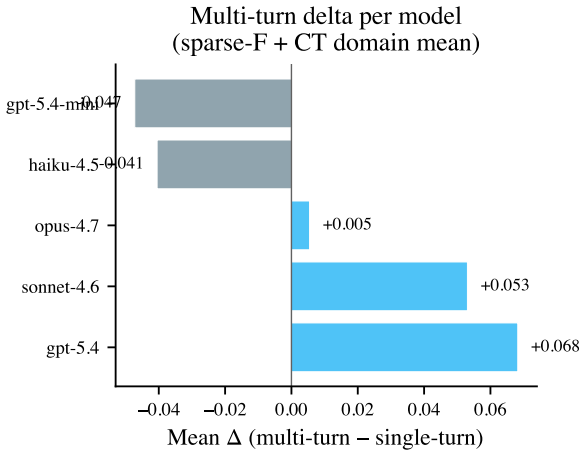


Figure 3: Mean reward delta (multi-turn minus single-turn) per model averaged over the multi-turn envs with paired data. GPT-5.4 is the only model with a meaningfully positive delta; all others are near-zero or negative. The clean frontier-vs-budget tier separation observed in Sprint-1 does not replicate on the complete matrix.

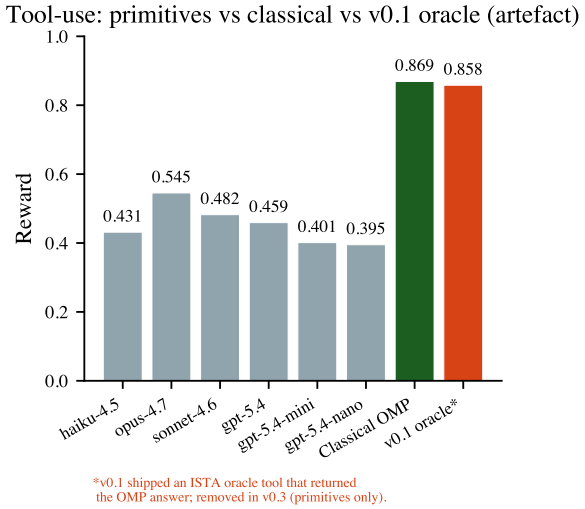


Figure 4: Tool-use rewards on sparse Fourier recovery (v0.3 primitive tools only). Six LLMs cluster in $[0.40, 0.55]$; classical OMP on the same instances reaches 0.87. The v0.1 oracle-artefact value (0.858) is shown for contrast: it matched classical precisely because the tool itself returned the classical answer.