Physics-R1: An Audited Olympiad Corpus and Recipe for Visual Physics Reasoning

Shan Yang
Independent Researcher
alexyangshan@gmail.com

Abstract

We audit the multimodal-physics evaluation pipeline end-to-end and document three undetected construction practices that distort how the field measures vision-language reasoning: train–eval contamination, translation drift, and MCQ saturation. (1) Public training pools (UGPhysics-Train, SciInstruct, MMK12) pass single-stage 5-gram-Jaccard audits with zero hits across all six public physics evals; a three-stage audit (Jaccard $\rightarrow$ `mxbai-embed-large` cosine $\rightarrow$ Haiku-4.5 LLM-judge) surfaces $\mathbf{134}$ near-duplicates and $\mathbf{4{,}846}$ paraphrase candidates in SciInstruct alone. (2) A 17-pp Sonnet-4.5 [Anthropic, 2025] delta on 59 paired Estonian-English olympiad problems (30.5% vs. 13.6%; sign test p=0.011, McNemar p=0.021, paired bootstrap 95% CI $[+5.1, +28.9]$ pp). (3) A 46-pp format-and-novelty gradient on identical Sonnet weights between MCQ (79.7% on PhyX) and open-ended olympiad evaluation (33.4% on PHYSOLYM-A). We release four artifacts addressing these gaps: PHYSCORP-A ($6{,}432$-record three-stage-audited multimodal corpus), PHYSR1CORP ($2{,}268$-record closed-form RL pool), PHYSOLYM-A (500-problem, 99.8% novel-source held-out olympiad eval with native difficulty labels and an EN/ET bilingual subset), and Physics-R1, a reference GSPO+DAPO recipe cold-started from Qwen3-VL-8B-Thinking. Across 3 seeds (§5), Physics-R1 lifts the audited corpus over the 8B base by $+18.3$ pp on PHYSOLYM-A liberal ($8.0 \rightarrow \mathbf{26.3} \pm 1.7$; 7.1 pp behind Sonnet 4.5), $+15.7$ pp on PhysReason ($23.9 \rightarrow \mathbf{39.6} \pm 6.4$; ahead of Qwen3-VL-32B and Gemini 2.5 Pro), $+6.9$ pp on OlympiadBench-Physics ($\mathbf{46.2} \pm 1.5$), and $+4.1$ pp on PhyX MCQ ($\mathbf{77.8} \pm 0.3$).

1 Introduction

Multimodal physics reasoning is increasingly tracked via vision-language benchmarks, but how those benchmarks are constructed is rarely audited. Researcher-curated training pools aggregate physics problems from publicly available sources whose paraphrase relationships evade conventional n-gram dedup; multilingual benchmarks distribute English translations of problems first composed in another language; MCQ-format splits saturate against the closed-frontier ceiling. Each represents a methodological gap in how the field constructs benchmarks, and together they distort cross-model comparisons, inflate frontier-model rankings on public leaderboards, and obscure the format-and-novelty axis along which capability actually diverges.

We argue that defensible measurement of multimodal physics reasoning requires an end-to-end audit of the evaluation pipeline. This paper performs that audit, surfaces three measurement findings, and constructs released artifacts directly against the gap each finding identifies. Physics-R1, a reference GSPO+DAPO recipe [Zheng et al., 2025, Yu et al., 2025] cold-started from Qwen3-VL-8B-Thinking [Qwen Team, 2025] and building on MM-Eureka [Meng et al., 2025] and DeepSeek-

Preprint.

R1's binary correctness signal [DeepSeek-AI, 2025, Shao et al., 2024], accompanies the corpus as evidence-of-trainability rather than as the primary contribution: it lifts the audited held-out eval over the 8B base while still trailing the closed frontier (§5.2).

Finding 1: single-stage 5-gram-Jaccard audit reports public physics-VL training pools as clean, but a three-stage audit (Jaccard → mxbai cosine → LLM-judge) surfaces 134 near-duplicates among $4{,}846$ Stage-2 candidates in SciInstruct alone. Across the three published physics-VL training pools we re-audit against six public evals (UGPhysics-Train, SciInstruct's 42K-record en_phy_chem split, MMK12's 15K-record train pool), conventional 5-gram-Jaccard at $J \geq 0.4$ (Stage-1) reports *zero* hits for every pool against all six evals—a single-stage audit calls them all clean. Stage-2 `mxbai-embed-large` cosine at ≥ 0.85 then surfaces $\mathbf{4{,}846}$ paraphrase-class candidate pairs from SciInstruct alone (PhysReason-full $2{,}687$, PhysUniBench-en $1{,}027$ dominant), 9 from UGPhysics-Train, and 66 from MMK12 (Table 2). Stage-3, a Haiku-4.5 LLM-judge, classifies each Stage-2 candidate as a *close duplicate* or a *same-topic neighbor*: of the $4{,}846$ SciInstruct candidates, $\mathbf{134}$ (2.8%) are close duplicates and the duplicate fraction is sharply cosine-driven (100% at $\cos \geq 0.95$, 1.5% at $\cos \in [0.85, 0.87)$). On a $1{,}679$-record researcher-curated sample of PHYSCORP-PRE-AUDIT ($14{,}294$ records) under the field-default within-pool dedup workflow, 345 records ($\mathbf{20.5}\%$) leak at Stage-1 alone against the six public evals (concentrated in PhysUniBench-en, 339, and MMMU-Pro Physics, 20); the joint Stage-1∨Stage-2 sweep on this same sample against an internal analysis eval reaches $\mathbf{8.8}\%$ at the published operating point and 27.1% at $\cos \geq 0.80$ (Table 4).

Finding 2: translation introduces a measurable score delta on identical physics problems. On 59 paired Estonian/English Physics Olympiad problems, Sonnet 4.5 [Anthropic, 2025] attains $\mathbf{30.5}\%$ strict on Estonian originals against only $\mathbf{13.6}\%$ on English translations of the same problems (sign test on 16 discordant pairs p=0.011; McNemar exact p=0.021; bootstrap 95% CI $[+5.1, +28.9]$ pp). Estonian PhO problems were composed in Estonian first; English versions are translations whose physics vocabulary, grammatical case mapping, and subtlety of scope degrade information content. For Sonnet 4.5, whose cross-lingual transfer covers Estonian, published numbers on the English-translation benchmark systematically *underestimate* model ability relative to original-language gold; for models with weaker training in the original language, the relationship is expected to reverse (App. H.4(viii), pre-registered) (§3.2, §5.1).

Finding 3: same-model evaluation across three physics benchmarks reveals a 46-point format-and-novelty gradient. Evaluated in the same week on identical Sonnet 4.5 weights, the score sweeps from $\mathbf{79.7}\%$ on PhyX [Shen et al., 2025] (4-way MCQ) down to $\mathbf{50.4}\%$ liberal on OlympiadBench-Physics [He et al., 2024] and $\mathbf{33.4}\%$ liberal on our held-out audited eval—format-and-novelty alone move the score by 46 points on fixed weights (§3.2; scoring in §5).

Together the three findings imply that defensible physics-VL measurement requires three properties at construction time: a three-stage audit (n-gram Jaccard → embedding cosine → LLM-judge precision filter), original-language gold, and open-ended novel-source evaluation. Four released artifacts instantiate this protocol: (a) PHYSCORP-A, the audited multimodal physics corpus produced by the three-stage pipeline (Algorithm 1), and the closed-form RL training pool PHYSR1CORP on which Physics-R1 is trained (§3); (b) PHYSOLYM-A, the open-ended held-out olympiad benchmark with native difficulty calibration, an EN/ET bilingual subset, and a Sonnet-as-judge protocol whose unjudgeable rate (13.9%) we disclose (§3.2, §5.1); (c) Physics-R1, a reference RL recipe whose audited held-out lift on PHYSOLYM-A validates the corpus as trainable rather than memorized (Table 3); we recommend a binary correctness reward as the default—variance-optimal under GSPO with group-normalized advantages, Goodhart-robust against unit/conservation/format proxies, and harness-portable (§4, properties P1–P4)—and report the dense five-component physics-native reward as a shape ablation; and (d) the audit protocol itself, released as `audit_three_stage.py` with saved best-overlap scores and Stage-3 judge labels (Appendix A). The 3-seed sensitivity sweep (seeds $\{42, 17, 23\}$ on the audited PHYSR1CORP) is reported in Table 3 with $\sigma \leq 3.3$ pp on PUB-OE, OlymBench-Phys, and PHYSOLYM-A, and σ=6.4 pp on PhysReason (seed-42 outlier); the reward-component drop-out ablation (Table 11) is left to follow-up work.

Table 1: **Released artifacts vs related benchmarks across eight axes.** *Audit:* 2-stage (n-gram+embedding) / 1-stage / orig. (constructed-novel) / none. *T/T leak:* train→test joint-stage ($J \geq 0.4 \vee \cos \geq 0.85$) audit against six public physics evals; ✓ all 6 = clean. *Diff:* organizer difficulty. *X-L:* paired cross-lingual. *Use:* E/T = eval/train. *RL-ready:* closed-form gold + audit-clean + RL recipe. "·" = eval-only; "n/r" = train pool, no cross-corpus audit. Only this work reports train/test contamination: after re-audit cleanup, PHYSCORP-A (6,432) and PHYSR1CORP (2,268) are clean against **all six** evals (Table 2).

Benchmark	Size	Format	MM	Audit	T/T leak	Diff	X-L	Use	RL-ready
Physics-domain benchmarks									
PHYBench [Qiu et al., 2025]	500	open+EED	–	orig.	·	–	–	E	–
PhysUniBench [Wang et al., 2025b]	3,304	open MM	✓	1-stage	·	✓	–	E	–
UGPhysics [Xu et al., 2025]	5,520	open text	–	1-stage	n/r	–	EN/ZH	T	–
PhysReason [Zhang et al., 2025]	1,200	step open MM	✓	none	·	–	–	E	–
OlympiadBench [He et al., 2024]	8,952	open MM	✓	none	·	–	EN/ZH	E	–
Olympiad / formal / contamination-by-design									
PutnamBench [Tsoukalas et al., 2024]	1,692	Lean/Isab.	–	orig.	·	✓	–	E	–
OIBench [Zhu et al., 2025]	250	open code	–	2-stage	·	✓	EN/ZH	E	–
FrontierMath [Glazer et al., 2024]	290	open math	–	orig.	·	✓	–	E	–
HLE [Phan et al., 2025]	2,500	expert exam	✓	orig.	·	–	–	E	–
Multimodal / multi-domain									
MMLU-Pro [Wang et al., 2024]	12,032	10-MCQ	–	none	·	–	–	E	–
MMMU-Pro Phys [Yue et al., 2024b]	60	10-MCQ MM	✓	none	·	–	–	E	–
SciInstruct [Zhang et al., 2024]	254,K	SFT instr.	–	1-stage	n/r	–	–	T	–
This work (train→test cross-corpus audit reported; Table 2)									
PHYSCORP-A (ours)	6,432	open+MCQ MM	✓	2-stage	✓ **all 6**	✓	✓	T	✓
PHYSR1CORP (ours)	2,268	MCQ + num MM	✓	2-stage	✓ **all 6**	✓	✓	T	✓
PHYSOLYM-A (ours)	500	open MM novel	✓	2-stage	· (eval; clean)	✓	EN/ET	E	–

✓= present; – = absent or not reported. "2-stage" audit = pairwise 5-gram-Jaccard *and* embedding-cosine against external corpora and held-out splits.

2 Related Work

Rule-based RL for reasoning. DeepSeek-R1 [DeepSeek-AI, 2025] established that simple rule-based rewards (binary correctness + format) suffice to train competitive math reasoners directly from a base model without SFT, using GRPO [Shao et al., 2024]. MM-Eureka [Meng et al., 2025] extended the recipe to VLMs with a difficulty curriculum; DAPO [Yu et al., 2025] added decoupled clipping and dynamic sampling; GSPO [Zheng et al., 2025] replaced token-level with sequence-level importance weighting. Physics-R1 inherits MM-Eureka's structural choices and the binary correctness reward unchanged: although physics intermediate steps carry units, conservation laws, and symbolic equations that *a priori* admit per-step verification, we find that under GSPO with group-normalized advantages a binary reward is variance-optimal and robust to the within-wrong-group Goodhart channel that physics-native shaping opens (§4); the dense physics-native reward is reported as an ablation.

Physics QA benchmarks. PhyX [Shen et al., 2025], OlympiadBench-Physics [He et al., 2024], UGPhysics [Xu et al., 2025], PhysReason [Zhang et al., 2025], MMMU/MMMU-Pro [Yue et al., 2024a,b], MMK12 [Meng et al., 2025], PHYBench [Qiu et al., 2025], and PhysUniBench [Wang et al., 2025b] are the canonical references. Top entries cluster within ten points of the closed-frontier ceiling on MCQ formats; only PHYBench, OIBench, and PutnamBench publish a contamination protocol, and none publish the three-stage (n-gram, embedding, LLM-judge) pairwise audit we introduce in §3.3. Table 1 maps our released audited corpus and PHYSOLYM-A against related benchmarks on seven axes.

Contamination audits and other prior work. PutnamBench [Tsoukalas et al., 2024], FrontierMath [Glazer et al., 2024], HLE [Phan et al., 2025], and EnigmaEval [Wang et al., 2025a] provide release-policy templates and dismissal grounds; methodological work spans n-gram audits [Sainz et al., 2023], the rephrased-samples failure mode [Yang et al., 2023] (which our Stage 2 catches), embedding-based detection [Singh et al., 2024], and performance-based detection [Dekoninck et al., 2024]; the survey of Ravaut et al. [2024] consolidates these. We import the math template, adding the embedding-cosine pass because physics statements (units, vectors, figure refer-

ences) are more paraphrase-sensitive than typical math problems—a sensitivity Table 4 quantifies. PhysBench [Chow et al., 2025] evaluates intuitive-physics dynamics from video, orthogonal scope. Multilingual benchmarks have proliferated [Xuan et al., 2025, Ahuja et al., 2024, Wu et al., 2025]; our cross-lingual finding (§5.1) differs methodologically by evaluating identical 59 problems in original Estonian and English translation on the same closed model with paired tests, isolating a within-problem effect aggregate benchmarks cannot.

3 Data: The Audited Corpus and Held-Out Olympiad Eval

Released artifacts: PHYSCORP-A (6,432-record audited corpus, including 1,609 first-ML-format olympiad problems—Estonian PhO with native 1–10 difficulty + 201 EN/ET bilingual, Kevin Zhou's handouts, 7 international olympiads); PHYSR1CORP (2,268-record closed-form RL pool, MCQ and numerical only); the held-out PHYSOLYM-A eval (§3.2); the Physics-R1 recipe (Algorithms 2, 3); and the audit pipeline (Algorithm 1, Table 4). All ship under per-source licenses (Table 5) on HuggingFace+GitHub+Zenodo with Croissant 1.0 metadata.

3.1 Training Corpus Composition

The corpus is drawn from nine source families (Table 9). Five are repackaged from existing benchmarks under documented licenses (UGPhysics [Xu et al., 2025], OpenStax College and University Physics [OpenStax, 2024], Physics Stack Exchange [Stack Exchange Inc., 2024], an MMMU+o1-CoT seed [Yue et al., 2024a], PhysReason [Zhang et al., 2025]); four contribute first-ML-format material: the Estonian Physics Olympiad collection [Estonian Physics Olympiad, 2018] (418 problems, 2004–2018, with organizer-issued 1–10 difficulty labels and a 201-problem bilingual EN+ET subset), Kevin Zhou's olympiad handouts [Zhou, 2018] (692 problems, with native point values 1–5 and a 3.2% advanced flag; some problems are drawn from books or other olympiad archives with inline attribution preserved per record, see Appendix E), and refreshed scrapes of seven international olympiads (IPhO [International Physics Olympiad, 2025], NBPhO [NBPhO Committee, 2025], EuPhO [EuPhO Committee, 2025], APhO [Asian Physics Olympiad Committee, 2025], USAPhO [American Association of Physics Teachers, 2025], INPhO [Homi Bhabha Centre for Science Education, 2025], IYPT). Source families ship under a mix of CC BY 4.0, CC BY-SA 4.0, public-domain by competition policy (Estonian PhO, IPhO, NBPhO, EuPhO, APhO, USAPhO, INPhO), CC BY-NC 4.0 (Kevin Zhou's handouts; written grant 2026-05-03), and CC BY-NC-SA 4.0 (UGPhysics); per-source licenses are listed in Appendix E (Table 5) and carried through to each released record. The full 14,294-record pre-audit pool is released as PHYSCORP-PRE-AUDIT so that downstream users can reproduce the audit; PHYSCORP-A is the 6,432-record subset that survives all three stages plus a re-audit against PhysReason-full and PhysUniBench-en (804 records dropped, dominated by PhysReason-full 540 and PhysUniBench-en 186). The released pool is disjoint from PhyX, MMMU-Pro Physics, OlympiadBench-Physics, UGPhysics-Train, PhysReason-full, PhysUniBench-en, and PHYSOLYM-A at the joint operating thresholds. The candidate-to-release cleanup for PHYSR1CORP is detailed in §3.3.

LLM-touched-statement subset disclosure. Of the 2,268 records in PHYSR1CORP, approximately 73 (3.2%) have LLM-touched problem statements: $\sim$ 11 are derived from a 85-record Claude-generated synthetic-MCQ augmentation pool (3 verbatim, 8 numeric paraphrases), and $\sim$ 62 are numeric-variation paraphrases of real PHYSCORP-A records (e.g., variant problem constants). The remaining $\sim$ 2,195 records have unmodified problem statements from the nine source families. LLM augmentation is documented per-distribution in the Croissant metadata's `syntheticDataDescription` field; the held-out PHYSOLYM-A eval contains no synthetic problem content.

3.2 PHYSOLYM-A: Held-Out Olympiad Eval

Standard physics-VL benchmarks no longer resolve frontier-class differences: PhyX clusters top entries within ten points of the 80% ceiling; OlympiadBench-Physics predates the contamination-audit discipline; UGPhysics is itself a candidate for audited training data, not held-out evaluation. Physics-R1's stopping rule and reward-component ablation depend on a held-out signal that is non-saturating and contamination-clean against the training pool.

Table 2: **Train/test contamination across released physics-VL training pools, three-stage audit.** Rows: public physics eval splits; columns: training pools (three competitor, two cleaned ours). Cells: **Stage-1 / Stage-2 raw / Stage-3 near-dup** pair counts (Algorithm 1). Stage-1 = 5-gram Jaccard ≥ 0.4, Stage-2 = `mxbai-embed-large-v1` cosine ≥ 0.85 (high recall over close-content pairs), Stage-3 = Haiku-4.5 LLM judge separating each Stage-2 candidate into *close duplicate* (paraphrase / numeric variation of the same problem) vs. *same-topic neighbor* (related physics, distinct setup). Competitor pools (UGPhysics-Train: 200-record annotated subset; SciInstruct: en_phy_chem 42,352-record subset of 254 K; MMK12: 15,608-record MM-Eureka train pool) report $\mathbf{0}/\mathbf{6}$ Stage-1 hits; Stage-2 surfaces $\mathbf{4{,}846}$ close-content pairs in SciInstruct, 9 in UGPhysics-Train, and 66 in MMK12. **Stage-3 LLM-judge separates close duplicates from same-topic neighbors:** SciInstruct $4{,}846 \rightarrow \mathbf{134}$ near-duplicates (PhysReason-full $2{,}687 \rightarrow 36$, PhysUniBench-en $1{,}027 \rightarrow 22$, PhyX-mini $703 \rightarrow 46$ dominant); UGPhysics-Train $9 \rightarrow \mathbf{0}$; MMK12 $66 \rightarrow \mathbf{0}$. The close-duplicate share is sharply cosine-driven: 100% at $\cos \geq 0.95$ vs. 1.5% at $\cos \in [0.85, 0.87)$ (Appendix A). **Both released pools are fully Stage-3 clean against all six evals:** PHYSCORP-A (6,432) after dropping 804 from a 7,236 candidate, with $\mathbf{0}/\mathbf{0}$ Stage-3 close-duplicates against all six evals (no S2 candidates surviving the joint S1$\vee$S2 cleanup); PHYSR1CORP (2,268) after dropping 87 MMMU-Pro + 78 PhyX-mini/PhysUniBench-en hits from a 2,433 candidate, with all 19 remaining S2 candidates classified as same-topic neighbors by Stage-3 ($\mathbf{0}/\mathbf{19}$ near-duplicates), agreeing 100% with manual inspection.

Eval ↓ / Train pool →	*Other published pools (we re-audit)*			*This work (cleaned)*	
	UGPhysics-Train (200 sub)	SciInstruct (en_phy_chem; 42 K)	MMK12 (MM-Eureka; 15 K)	PHYSCORP-A (6,432)	PHYSR1CORP (2,268)
PHYSOLYM-A (500)	0 / 0 / **0**	0 / 163 / **8**	0 / 0 / **0**	0 / 0 / **0**	0 / 3 / **0**
PhyX-mini (1,000)	0 / 1 / **0**	0 / 703 / **46**	0 / 0 / **0**	0 / 0 / **0**	0 / 0 / **0**
MMMU-Pro Phys (60)	0 / 0 / **0**	0 / 141 / **7**	0 / 0 / **0**	0 / 0 / **0**	0 / 1 / **0**
OlymBench-Phys (692)	0 / 2 / **0**	0 / 130 / **15**	0 / 0 / **0**	0 / 0 / **0**	0 / 4 / **0**
PhysReason-full (1,200)	0 / 4 / **0**	0 / 2,687 / **36**	0 / 62 / **0**	0 / 0 / **0**	0 / 11 / **0**
PhysUniBench-en (1,022)	0 / 2 / **0**	0 / 1,027 / **22**	0 / 4 / **0**	0 / 0 / **0**	0 / 0 / **0**
Total S2 / S3 real	9 / **0**	**4,846 / 134**	66 / **0**	0 / **0**	**19 / 0**

Format: **Stage-1 / Stage-2 / Stage-3** pair counts (Stage-3 = Haiku-4.5 LLM-judge classifying each Stage-2 candidate as close duplicate vs. same-topic neighbor). SciInstruct's S3-near-dup cells reveal the close-duplicate share is threshold-driven: $17/17$ at $\cos \geq 0.95$, $54/1{,}159$ at $[0.87, 0.90)$, $53/3{,}543$ at $[0.85, 0.87)$ (Appendix A, Table A).

PHYSOLYM-A (Physics Olympiad, Audited) is composed of 200 problems from Kevin Zhou's olympiad handouts, 136 from the Estonian PhO collection, 85 from an IPhO/NBPhO/EuPhO scrape, and 79 from an APhO/USAPhO/INPhO scrape (500 total, **499** novel-source under our four-corpus audit). Native difficulty signals: 27% of records carry Estonian organizer-issued 1–10 difficulty; 38% carry Zhou's pedagogical 1–5 point values; 2% carry Zhou's advanced [A] flag. The three-stage audit (§3.3) certifies **0** Stage-3 near-duplicate overlaps between the audited training pool and PHYSOLYM-A, and **0** overlaps between the novel pool and PhyX 1000q. The single non-novel record is an EuPhO 2020 problem also present in OlympiadBench-Physics at $J{=}0.91$; we disclose this in Appendix A rather than silently drop it. The scoring protocol (LLM-judge with strict/liberal accuracy, κ inter-judge agreement, and the auxiliary held-out splits used during training) is described in §5.

3.3 The Three-Stage Audit Pipeline

The pipeline constructs both the audited training pool and the held-out PHYSOLYM-A eval under the same definition of contamination, applied pairwise across the training pool, four external corpora (PhyX, MMMU-Pro Physics, OlympiadBench-Physics, UGPhysics-Train), and the held-out splits. *Stage 1 (n-gram).* Tokenize each problem statement with a unicode word tokenizer, build the 5-gram shingle set, and flag pairs with Jaccard ≥ 0.4. *Stage 2 (embedding).* Encode each statement with `mxbai-embed-large` (1024-dim, L_2-normalized) and flag pairs with cosine ≥ 0.85. Stage-2 has high recall on close-content pairs, including the rephrasing-class duplicates Stage-1 misses, but its single-threshold operating point also flags same-topic-but-distinct-problem pairs. *Stage 3 (LLM-judge precision filter).* For each Stage-2 candidate, a Haiku-4.5 judge receives both problem statements and classifies the pair as a *close duplicate* (paraphrase or numeric variation of the same problem) or a *same-topic neighbor* (related physics, distinct setup). Only Stage-3 close-duplicate

records are removed from the training pool. Pseudocode is in Algorithm 1; worked examples in Appendix H.5; calibration of the embedder + thresholds in Appendix A.

On the train/test contamination matrix of Table 2, the cosine-bucketed precision pattern (100% close-duplicates at $\cos \geq 0.95$ vs. 1.5% at $\cos \in [0.85, 0.87)$; Appendix A, Table A) confirms the protocol's design hypothesis: embedding cosine alone is recall-dominant and an LLM judge is the appropriate precision filter. **Both released training pools are fully Stage-3 clean against all six public evals (Table 2):** PHYSCORP-A (6,432), built via Stage-1∨Stage-2 audit dropping 804 of a 7,236 candidate, surfaces 0 Stage-2 candidates and hence $\mathbf{0/0}$ Stage-3 close-duplicates by construction; PHYSR1CORP (2,268), additionally dropping 87 MMMU-Pro and 78 PhyX-mini/PhysUniBench near-duplicates from a 2,433-record candidate (Appendix A.1), retains 19 Stage-2 candidates classified as same-topic neighbors by Stage-3 with 100% manual-inspection agreement ($\mathbf{0/19}$ close-duplicates).

Algorithm 1 Three-stage contamination audit.

Require: Train pool T, external corpora $\{E_k\}_{k=1}^{K}$, held-out splits $\{H_j\}_{j=1}^{J}$, normalize fn $\mathrm{norm}(\cdot)$, embedder $\mathrm{enc}(\cdot)$, LLM judge $\textsc{Judge}(\cdot,\cdot) \in \{\text{close-dup}, \text{topic-neighbor}\}$, thresholds $\tau_J{=}0.4$, $\tau_C{=}0.85$

Ensure: Audited pool T' disjoint from $\bigcup_k E_k \cup \bigcup_j H_j$ at the joint thresholds.

Stage 1: 5-gram Jaccard (n-gram audit).

$S_t \leftarrow \{\text{5-gram shingle set of } \mathrm{norm}(t)\}$ for each $t \in T \cup \bigcup E_k \cup \bigcup H_j$

for $t \in T$ **do**

  $J_{\max}(t) \leftarrow \max_{x \in \bigcup E_k \cup \bigcup H_j} |S_t \cap S_x| / |S_t \cup S_x|$

end for

Stage 2: `mxbai-embed-large` cosine (paraphrase recall).

$\mathbf{e}_t \leftarrow \mathrm{enc}(\mathrm{norm}(t)) / \|\mathrm{enc}(\mathrm{norm}(t))\|$ for each t

for $t \in T$ **do**

  $C_{\max}(t) \leftarrow \max_{x \in \bigcup E_k \cup \bigcup H_j} \mathbf{e}_t^{\top} \mathbf{e}_x$

end for

$C(T) \leftarrow \{t : J_{\max}(t) \geq \tau_J \text{ OR } C_{\max}(t) \geq \tau_C\}$ ▷ candidate set, high-recall union

Stage 3: Haiku-4.5 LLM-judge (precision filter). For each $t \in C(T)$ with top-matching $x^*(t) \leftarrow \arg\max_x \mathbf{e}_t^{\top} \mathbf{e}_x$, query JUDGE to classify the pair as a close duplicate (paraphrase / numeric variation of the same problem) or a same-topic neighbor (related physics, distinct setup).

$R(T) \leftarrow \{t \in C(T) : \textsc{Judge}(t, x^*(t)) = \text{close-dup}\}$

$T' \leftarrow T \setminus R(T)$

return T' and per-stage counts $|J_{\max} \geq \tau_J|$, $|C_{\max} \geq \tau_C|$, $|R(T)|$ (Tables 2, 4).

Threshold-sensitive leakage on a researcher-curated baseline (Finding 1). On a 1,679-record sample drawn from PHYSCORP-PRE-AUDIT under conventional 5-gram-Jaccard + within-pool embedding dedup, audited against a 500-record internal analysis eval (distinct from PHYSOLYM-A, constructed post-audit), the joint Stage-1∨Stage-2 audit raises the detected leak rate from 3.3% (Stage-1 alone, all exact matches at $J{=}1.0$) to $\mathbf{8.8}\%$, sweeping 4.7–27.1% as the cosine threshold moves between 0.90 and 0.80 (Appendix A.1, Table 4). The 5.5-pp gap is the rephrasing dark-matter that justifies the audited release as a measurement intervention.

4 Physics-R1: A Multi-Model RL Recipe

Physics-R1 is reported as evidence the audited corpus has training utility under standard rule-based RL, not as an algorithmic contribution. The optimizer is GSPO [Zheng et al., 2025]+DAPO [Yu et al., 2025], unmodified. For each prompt x, sample $K{=}16$ rollouts $\{y_k\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$, score with reward $r(y_k, x)$, form group-normalized advantages and the clipped sequence-level GSPO objective

$$A_k = \frac{r(y_k, x) - \bar{r}}{\sigma_r + \varepsilon}, \qquad w_k(\theta) = \left(\frac{\pi_\theta(y_k \mid x)}{\pi_{\theta_{\text{old}}}(y_k \mid x)} \right)^{1/|y_k|},$$
$$\mathcal{L}_{\text{GSPO}} = -\mathbb{E}\left[\tfrac{1}{K} \sum_k \min\big(w_k A_k,\ \mathrm{clip}(w_k,\ 1{\pm}\epsilon) A_k \big) \right] + \beta_{\text{KL}} D_{\text{KL}}(\pi_\theta \| \pi_{\text{base}}), \tag{1}$$

with $(\bar{r}, \sigma_r)$ the group mean/std, $(\epsilon_{\text{lo}}, \epsilon_{\text{hi}})=(0.20, 0.28)$, $\beta_{\text{KL}}=10^{-3}$, $\pi_{\text{base}}=$ Qwen3-VL-8B-Thinking BASE. Cold-start from base, KL anchor, MM-Eureka [Meng et al., 2025] difficulty curriculum (drop $0/N$ and N/N prompts, $\sim$22% filtered), 12,288-token CoT budget, and held-out PhyX-mini-MC early stopping fix the joint setting (Algorithm 3, Table 10); implementation uses `verl` 0.6.1 [Sheng et al., 2024] on Qwen3-VL-8B-Thinking [Qwen Team, 2025] with FSDP1 sharding (§6).

Two reward shapes: binary (recommended) vs. dense (ablation). Physics rollouts admit physics-native per-step signals—units, conservation, symbolic form—so a denser reward looks free. We compare:

$$\begin{aligned} \text{(binary, recommended)} \quad & r_{\text{bin}}(y, x) = \mathbb{1}\big[\textsc{Match}(\textsc{ExtractBoxed}(y),\, g(x))\big] \in \{0, 1\}, \\ \text{(dense, ablation)} \quad & r_{\text{dense}} = \text{clip}\big(r_{\text{ans}} + r_{\text{fmt}} + r_{\text{dim}} + r_{\text{sym}} + r_{\text{cons}},\, -1,\, 1\big). \end{aligned} \tag{2}$$

where MATCH accepts MCQ-letter equality, $\pm 1\%$ numeric tolerance, or symbolic equivalence (Appendix C.1); the dense components are $r_{\text{ans}} \equiv r_{\text{bin}}$, $r_{\text{fmt}} \in \{0, +0.1\}$ (`\boxed{}` present), $r_{\text{dim}} \in \{0, +0.15\}$ (`sympy.physics.units`), $r_{\text{sym}} \in \{0, +0.20\}$ (`\frac` sympifies), $r_{\text{cons}} \in \{-0.25, 0\}$ (energy/momentum violation; Appendix C.2). Under GSPO with group-normalized advantages and a difficulty curriculum, four properties land binary as variance-optimal and Goodhart-robust (full derivation in Appendix C.1). **(P1) Group normalization absorbs reward magnitude:** A_k is invariant to affine rescaling of r within a group, so dense only matters when it *reorders* rollouts—we measure 14.3% of within-group pairs flipped, 87% inside the all-wrong subgroup. **(P2) Wrong-group reorderings are a Goodhart channel:** rewarding well-formatted-but-wrong above poorly-formatted-but-wrong biases the policy toward LaTeX-format proxies that transfer poorly to the audited held-out eval. **(P3) Variance-optimal advantage:** on a Bernoulli reward, $\text{Var}(A^{\text{bin}})=1$ saturates the K-sample bound; a bounded shaping term $\delta_k \in [0, \Delta]$ inflates σ_r by $O(\Delta^2)$, shrinking $|A^{\text{dense}}_{\text{correct}}|$ below $|A^{\text{bin}}_{\text{correct}}|$. Empirical signature at matched step 60 on the seed-42 ablation (Table 3): binary beats dense by $+8.9/+4.9/+6.4$ pp on PhysReason/OlymBench-Phys-liberal/PHYSOLYM-A-liberal while tied with dense on PUB-OE (-0.7 pp) and trailing dense by at most 0.6 pp on saturated MCQ. We ship binary as the deployable artifact; the per-component drop-out ablation (Table 11) is left to follow-up work.

5 Evaluation

We organize this section in two parts. §5.1 characterizes PHYSOLYM-A as a measurement instrument and grounds Findings 2–3 of §1 (Finding 1, audit-leakage, is in §3.3). §5.2 reports Physics-R1 results to validate PHYSCORP-A as trainable. The Physics-R1 (binary, seed 42) row in Table 3 is the headline single-seed checkpoint; the 3-seed mean row aggregates seed 42 with two additional seeds (seed-17/step-63 and seed-23/step-60) on the audited PHYSR1CORP corpus.

Scoring protocol. All open-ended columns of Table 3 use *problem-level* liberal Sonnet-as-judge accuracy (Appendix D): for multi-sub-part problems on PhysReason and PhysUniBench-OE, `llm_judge_v2_alignment.py` and `llm_judge_v3_pubeo.py` respectively call Sonnet 4.5 once per gold sub-answer with YES/NO, and the problem is judged correct only if every sub-part is correct (AND across sub-parts); OlympiadBench-Physics and PHYSOLYM-A use `judge_olympiad.py`, which makes a single YES/NO call per problem against the full gold solution. The unjudgeable rate on PHYSOLYM-A is 13.9% (gold solutions consisting of grading rubrics, administrative notes, or figure-only references). Three layers bound judge optimism: strict vs. liberal gap on Sonnet (4.7 pp on PHYSOLYM-A); inter-judge Cohen's κ [Cohen, 1960] between two Sonnet seeds; and a 100-problem human-graded random subset (Appendix D). The cross-vendor judge agreement on a 50-problem Sonnet/GPT-4o pair test shows GPT-4o is *more* lenient than Sonnet (16% vs. 8% positive rate), bounding self-grading concern in the opposite direction from naive worry.

Judging concurrency and reproducibility. All Sonnet-judge runs reported in Table 3 are executed at `workers`=2–4 concurrency to stay below Anthropic API rate limits; sub-judgments that exceed the per-call timeout are retried at lower concurrency rather than counted as wrong. Per-cell judge-error counts (typically 0–10 out of 629–1200 records, all $\leq 1\%$) and per-record verdicts are released as `judge_audit.json` in the supplementary archive. The Sonnet 4.5 PhysReason responses (Table 3, †) are regenerated with `max_tokens`=16384 to match the response budget used by

all open-source baselines and Physics-R1; intermediate-length Sonnet responses (mean < 200 chars under default API settings) systematically fail to commit a `\boxed{}` final answer on multi-sub-part problems, which v2_alignment scores as wrong.

5.1 PHYSOLYM-A as a Measurement Instrument

Same-model evaluation reveals a 46-point format-and-novelty gradient. On identical Sonnet 4.5 weights evaluated in the same week, the score sweeps from 79.7% on PhyX (4-way MCQ) down to 50.4% liberal on OlympiadBench-Physics and 33.4% liberal on PHYSOLYM-A—a 46-point gradient on fixed weights, the strongest evidence the paper has for the central claim that physics evaluation is format- and novelty-bound (Finding 3). Three forces drive the drop: format (4-way MCQ vs. open-ended), genre (PhyX is K-12 to early-undergraduate, the bottom two are competition-grade), and contamination-removal (only PHYSOLYM-A is three-stage audited against the Physics-R1 training pool). The PhyX→OlympiadBench-Physics step accounts for $\sim$29 pp of the gradient (dominated by format + genre, since both are public and not contamination-cleaned), and the OlympiadBench-Physics→PHYSOLYM-A step adds $\sim$17 pp on top (dominated by audit and novelty since both are open-ended and competition-grade); a controlled 2×2 (format $\times$ audit on identical items) is left to follow-up work to attribute the residual cleanly. PHYSOLYM-A sits at the bottom of this gradient by construction. The per-physics-category breakdown on OlympiadBench-Physics (electromagnetism hardest at 38.4%, astrophysics easiest at 72.9%) and the saturation-gradient table are in Appendix H.7.

Difficulty-stratified accuracy from organizer-issued labels. The Estonian Physics Olympiad is the only public physics olympiad whose problems carry organizer-issued difficulty labels (1–10) by construction, eliminating self-annotation circularity. On the 131 Estonian problems carrying native annotation, Sonnet 4.5 strict accuracy decays near-monotonically from 62.5% at difficulty 1 to a hard 0% floor at difficulties 3, 6, 8, and 10 (full table: Appendix H.7, Table 12). The trivial end of the Estonian olympiad (62.5% at difficulty 1) already lies below Sonnet's PhyX score (79.7%); the clean zeros at four difficulty bins are the empirical signature of a non-saturating benchmark, the property required for PHYSOLYM-A to serve as a stopping signal during Physics-R1 training.

Cross-lingual ablation: 17-point translation delta on identical problems. The Estonian PhO bilingual subset enables a controlled cross-lingual experiment on 59 paired problems graded against the same gold by the same Sonnet 4.5: $\mathbf{30.5}\%$ strict on Estonian originals vs. $\mathbf{13.6}\%$ on English translations (sign test p=0.011; McNemar p=0.021; paired bootstrap 95% CI $[+5.1, +28.9]$ pp). The per-problem agreement matrix is asymmetric: 13 correct on Estonian but wrong on English, 3 the reverse. For weaker open-source 8B-class models with limited Estonian training the gap is expected to flip in sign (pre-registered as a follow-up direction; Appendix H.4, item (viii)).

5.2 Physics-R1

Recipe. Physics-R1 is the GSPO [Zheng et al., 2025]+DAPO [Yu et al., 2025] recipe of §4 cold-started from Qwen3-VL-8B-Thinking BASE [Qwen Team, 2025] on PHYSR1CORP (§3.1) under MM-Eureka difficulty filtering [Meng et al., 2025] and a binary correctness reward. `PhyX-mini-MC` (1,000-problem audit-clean MCQ subset [Shen et al., 2025]) is held out as the in-training early-stop signal: 4-way MCQ gives a cleaner per-step trajectory than open-ended judging, and it is certified disjoint from PHYSR1CORP under the audit pipeline of §3.3.

Capability across formats and the held-out olympiad column. Table 3 reports Physics-R1 alongside closed-frontier and open-source bases on three answer formats and the held-out olympiad split. The Qwen3-VL-8B-Thinking BASE checkpoint attains 73.7% on PhyX-mini-1k; the 32B sibling is indistinguishable on PhyX (73.8%), so scale alone in the 8B–32B Thinking band does not move PhyX. All Physics-R1 evals use `max_tokens`=16384 to permit full thinking-mode CoT; eval-budget sensitivity, harness-canonical re-evaluation, and per-judge gap analysis are in Appendix H.7.

Physics-R1 (binary, recommended): step 60 closes the audited held-out gap. The binary-reward checkpoint at step 60 lifts the 8B base across all formats (Table 3); the largest lift lands on PHYSOLYM-A liberal ($+18.3$ pp at 3-seed mean), with lifts on saturated public open-ended splits substantially smaller—the contamination signal Finding 1 predicts: where the 8B base is

Table 3: **Capability across MCQ, numerical, open-ended, and held-out olympiad benchmarks.** MCQ random baseline: PhyX-1k/3k 25%. All open-ended columns (PhysReason, PUB-OE, OlymBench-Phys, PHYSOLYM-A) use *problem-level* liberal Sonnet-as-judge accuracy under our v2/v3 judges (Appendix D): every sub-question of a multi-part problem must be judged correct for the problem to count. All Sonnet-judge runs use `workers`=2–4 concurrency for rate-limit safety, with errored sub-judgments retried at lower concurrency. Sonnet PhysReason cell (†) is generated with `max_tokens`=16384 to match the protocol used by Physics-R1 and the open-source baselines. GPT-4o PhyX-1k/3k from Shen et al. [2025]; Gemini PhyX-1k/3k cells (*) measured here. All Physics-R1 evals use `max_tokens`=16384.

Model	MCQ		Open-ended (problem-AND aggregation, liberal Sonnet-judge)			
	PhyX-1k MCQ-exact	PhyX-3k MCQ-exact	PhysReason subpart-AND (v2)	PUB-OE subpart-AND (v3)	OlymBench problem-lvl	PHYSOLYM-A problem-lvl
Closed-source frontier						
Claude Sonnet 4.5	79.7	80.6	**49.1**†	25.4	**50.4**	**33.4**
Gemini 2.5 Pro	75.1*	49.8*	38.8	33.4	37.4	12.2
GPT-4o	70.4	53.6	<u>51.1</u>	31.0	19.7	19.5
*Open-source bases (best / second-best across this block + ours: **bold** / <u>underline</u>)*						
Qwen3-VL-32B-Thinking	73.8	**84.2**	25.1	32.8	**53.9**	13.2
Qwen3-VL-8B-Thinking (base)	73.7	74.4	23.9	35.3	39.3	8.0
InternVL3-8B	46.8	43.1	13.3	23.5	10.4	4.0
This work (subscripts: Δ vs. Qwen3-VL-8B-Thinking base)						
Physics-R1 (dense)	$\mathbf{78.3}_{+4.6}$	$\underline{77.5}_{+3.1}$	$23.3_{-0.6}$	$\mathbf{37.7}_{+2.4}$	$40.5_{+1.2}$	$\underline{19.2}_{+11.2}$
Physics-R1 (binary, seed 42)	$\underline{78.0}_{+4.3}$	$76.9_{+2.5}$	$\underline{32.2}_{+8.3}$	$\underline{37.0}_{+1.7}$	$\underline{45.4}_{+6.1}$	$25.6_{+17.6}$
Physics-R1 (binary, 3-seed mean $\pm\sigma$)‡	$77.8^{+4.1}_{\pm 0.3}$	$76.9^{+2.5}_{\pm 0.3}$	$\mathbf{39.6}^{+15.7}_{\pm 6.4}$	$34.8^{-0.5}_{\pm 3.3}$	$\mathbf{46.2}^{+6.9}_{\pm 1.5}$	$\mathbf{26.3}^{+18.3}_{\pm 1.7}$

† Sonnet PhysReason regenerated at `max_tokens` = 16384, 1,192/1,200 clean records. ‡ 3-seed mean over seeds {42, 17, 23} on the audited PHYSR1CORP (2,268 records, binary reward; seed-42 also the dense-ablation seed). Per-seed (42/17/23): PR 32.2/43.1/43.4; PUB-OE 37.0/36.4/30.9; OlymBench-Phys 45.4/45.3/48.0; PHYSOLYM-A 25.6/25.0/28.2; PhyX-mini 78.0/77.4/77.9; PhyX-3k 76.9/77.2/76.6. PHYSOLYM-A lift decomposes into ∼ 3.5 pp from `\boxed{}`-emission rate (33.8%→64.4% from base to Physics-R1) and ∼ 14.1 pp from conditional accuracy (22.5→36.3); details in Appendix H.7.

already close to ceiling (PUB-OE 35.3, OlymBench-Phys 39.3), there is little headroom; where the eval is novel-source and audited (PHYSOLYM-A liberal 8.0), the post-training lift is large. The numerical/open-ended jumps from step 40 to step 60 are driven by the additional 20 GRPO steps lifting the `\boxed{}`-emission rate from 46% to 87–96%. The 3-seed mean (seed 42 + seed-17/step-63 + seed-23/step-60, all on the audited PHYSR1CORP) is tight across most open-ended columns: per-seed PR $\{32.2, 43.1, 43.4\}$ (mean 39.6 ± 6.4; seed-42 outlier on PR, $\sim$ 11 pp below seeds 17/23 with otherwise comparable performance on other columns), PHYSOLYM-A liberal $\{25.6, 25.0, 28.2\}$ (mean 26.3 ± 1.7), OlymBench-Phys $\{45.4, 45.3, 48.0\}$ (mean 46.2 ± 1.5), PUB-OE $\{37.0, 36.4, 30.9\}$ (mean 34.8 ± 3.3). The audited corpus is trainable and the lift over the 8B base is reproducible across seeds.

Where Physics-R1 helps: failure modes of the base it mitigates. The +18.3-pp 3-seed-mean lift on PHYSOLYM-A liberal corresponds to ∼92 problems flipped from wrong-on-base to correct-on-Physics-R1 (per-seed range 85–101). Hand-inspecting 30 such flips reveals three recurring failure modes of the 8B base, each addressed by a specific recipe lever: (i) *reasoning-without-committing* (long correct CoT, no `\boxed{}` final) — fixed by r_{bin} (§4); (ii) *unit/dimensional shortcuts* (dimensionally-consistent but answer-wrong) — fixed by MM-Eureka curriculum filtering of N/N surface-heuristic prompts; (iii) *multi-image evidence integration* (base attends only to the first panel) — fixed by the cold-start from Qwen3-VL-8B-Thinking BASE under FSDP1, which preserves the visual encoder. Physics-R1 does *not* fix genuine physics-content gaps (graduate-level Tripos-style perturbation theory remains wrong on both). Full transcripts in Appendix H.1.

PHYSOLYM-A grounds the central training-utility claim. The PHYSOLYM-A-liberal column of Table 3 is the cleanest non-saturating capability signal in our 7-axis comparison (Table 1). Sonnet attains 33.4%; Physics-R1 binary at the 3-seed mean reaches $\mathbf{26.3 \pm 1.7}\%$ (per-seed $\{25.6, 25.0, 28.2\}$ across seeds 42/17/23), exceeding every open-source baseline (Qwen3-VL-32B

13.2%, 8B 8.0%, InternVL3 4.0%) and the non-Sonnet closed APIs (GPT-4o 19.5%, Gemini 2.5 Pro 12.2%), trailing only Sonnet by 7.1 pp.

Reward-shape ablation: dense gives a small saturated-MCQ benefit, binary wins on open-ended. Under problem-level liberal Sonnet-judge scoring (seed 42), dense at step 60 slightly leads on saturated MCQ (PhyX-mini 78.3 vs. binary 78.0; PhyX-3k 77.5 vs. 76.9) but trails binary on every non-MCQ split: PhysReason (23.3 vs. **32.2**, +8.9 pp), OlympiadBench-Physics (40.5 vs. **45.4**, +4.9 pp), and PHYSOLYM-A liberal (19.2 vs. **25.6**, +6.4 pp). On PUB-OE, dense and binary are within 0.7 pp (37.7 vs. 37.0). The binary advantage is concentrated on the multi-sub-part numerical-answer column (PR) and the held-out audited olympiad column (PHYSOLYM-A)—the two columns where the recipe contribution should matter most. Dense is a reward-shape ablation; binary is the recommended default. The five-component reward drop-out (Table 11), recipe-flag, and SFT data-scaling ablations are left to follow-up work.

6 Discussion and Limitations

Physics-R1 uses unmodified GSPO+DAPO; the dense reward and recommended baseline (Algorithm 3) are reproducibility artifacts, not method claims. The audit pipeline catches verbatim and lightly-paraphrased duplicates and is empirically robust to both embedder choice (Spearman ρ=0.78 vs. `text-embedding-3-large`; OpenAI candidate set is a strict subset of mxbai's at every threshold tested; Appendix A) and judge choice (Sonnet 4.5 vs. GPT-4o cross-judge κ=0.44 on a 50-problem PHYSOLYM-A subset, with GPT-4o more lenient—self-grading direction is opposite to the feared bias; Appendix D); the Sonnet-as-judge 13.9% unjudgeable rate is a disclosed noise floor. The cross-lingual finding is Sonnet-4.5-specific on n=59 paired items (65.7% MC power, all three tests reject H_0); direction is pre-registered to reverse for cross-lingual-weak models (Appendix H.4, item viii).

7 Conclusion

Three findings—**134** near-duplicates in SciInstruct surfaced only by three-stage audit (Jaccard→cosine→Haiku-4.5 judge), a 17-pp Estonian–English translation delta on identical olympiad problems, and a 46-pp format-and-novelty gradient on fixed Sonnet 4.5 weights—motivate four released artifacts: PHYSCORP-A (6,432-record audited corpus, fully Stage-3 clean against all six public physics evals; Table 2), PHYSR1CORP (2,268-record closed-form RL pool), PHYSOLYM-A (500-problem held-out olympiad eval, 99.8% novel-source), and Physics-R1, a binary-reward GSPO+DAPO recipe that lifts PHYSOLYM-A liberal +18.3 pp over the 8B base at the 3-seed mean (8.0→**26.3** ± **1.7**, still 7.1 pp below Sonnet 4.5; per-seed {25.6, 25.0, 28.2} across seeds {42, 17, 23} on the audited PHYSR1CORP). Audit-pipeline robustness to embedder and judge choice is established in §6 (Appendices A, D). We recommend binary correctness reward as the deployable default (variance-optimal under GSPO with group-normalized advantages, Goodhart-robust against unit/format proxies; §4); reward-component drop-out (Table 11) is left to follow-up work; the 3-seed mean reported in Table 3 ($\sigma \leq$ 3.3 pp on PUB-OE, OlymBench-Phys, and PHYSOLYM-A; σ=6.4 pp on PhysReason driven by a seed-42 outlier) demonstrates that the audited corpus retains training signal across seeds.

Acknowledgments

The author thanks Kevin Zhou for granting permission to redistribute his olympiad handouts under CC BY-NC 4.0, the Estonian Physics Olympiad committee for making their archived problems and solutions publicly available at `https://fyysika.ee/`, the international olympiad committees (IPhO, NBPhO, EuPhO, APhO, USAPhO, INPhO) for the public archives that enabled the novel-source held-out evaluation, and the maintainers of the public physics-VL benchmarks (PhyX, MMMU-Pro, OlympiadBench, UGPhysics, PhysReason, PhysUniBench) whose released training pools and evals enabled the contamination audit reported in this work. Compute support for Physics-R1 training and evaluation was provided by RunPod.

References

Sanchit Ahuja, Varun Gumma, and Sunayana Sitaram. Contamination report for multilingual benchmarks, 2024. Tests 7 LLMs on multilingual benchmarks; finds nearly all are contaminated.

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, et al. Croissant: A metadata format for ML-ready datasets. In *NeurIPS Workshop on Data-centric Machine Learning Research*, 2024. URL https://arxiv.org/abs/2403.19546.

American Association of Physics Teachers. U.s. physics olympiad: Archived problems and solutions. https://www.aapt.org/physicsteam/, 2025.

Anthropic. Claude sonnet 4.5, 2025. URL https://www.anthropic.com/news/claude-sonnet-4-5. Model used as judge and frontier-baseline reference throughout this paper.

Asian Physics Olympiad Committee. Asian physics olympiad: Archived problems and solutions. https://apho2025.fkfi.lt/, 2025.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *ICLR*, 2025. Oral; https://openreview.net/forum?id=Q6a9W6kzv5.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jasper Dekoninck, Mark Niklas Mueller, and Martin Vechev. ConStat: Performance-based contamination detection in large language models. In *NeurIPS*, 2024.

Estonian Physics Olympiad. Estonian physics olympiad: Problem collection 2004–2018. https://www.fyysika.ee/, 2018.

EuPhO Committee. European physics olympiad: Archived problems and solutions. https://eupho.ee/, 2025.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021.

Eric Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, et al. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, et al. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *ACL*, 2024.

Homi Bhabha Centre for Science Education. Indian national physics olympiad: Archived problems and solutions. https://olympiads.hbcse.tifr.res.in/, 2025.

International Physics Olympiad. International physics olympiad: Archived problems and solutions. https://ipho-unofficial.org/, 2025.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread - new fluffy embeddings model, 2024. URL https://www.mixedbread.com/blog/mxbai-embed-large-v1. mxbai-embed-large-v1: 1024-dim sentence embedding model used in our Stage-2 audit.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Tao Zhao, Yu Qiao, and Ping Luo. Mm-eureka: Exploring visual aha-moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

NBPhO Committee. Nordic-baltic physics olympiad: Archived problems and solutions. https://www.nbpho.eu/, 2025.

OpenStax. College Physics 2e and University Physics Volumes 1-3. https://openstax.org/subjects/science, 2024.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam, 2025.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, et al. PHYBench: Holistic evaluation of physical perception and reasoning in LLMs. In *NeurIPS Datasets and Benchmarks Track*, 2025. https://openreview.net/forum?id=brG8FPq1cf.

Qwen Team. Qwen3-VL. https://huggingface.co/Qwen/Qwen3-VL-8B-Thinking, 2025. Vision-language model release.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, et al. A comprehensive survey of contamination detection methods in large language models, 2024.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, et al. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *EMNLP Findings*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Hui Shen, Taiqiang Wu, Qi Han, et al. PhyX: Does your model have the "wits" for physical reasoning? *arXiv preprint arXiv:2505.15929*, 2025.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, et al. HybridFlow: A flexible and efficient RLHF framework (verl). *arXiv preprint arXiv:2409.19256*, 2024.

Aaditya K. Singh, Muhammed Yusuf Kocyigit, Andrew Poulton, et al. Evaluation data contamination in LLMs: how do we measure it and (when) does it matter? *arXiv preprint arXiv:2411.03923*, 2024.

Stack Exchange Inc. Physics stack exchange: Question and answer archive. https://physics.stackexchange.com/, 2024.

George Tsoukalas, Jasper Lee, John Jennings, Yifan Xin, et al. PutnamBench: Evaluating neural theorem-provers on the putnam mathematical competition. *NeurIPS Datasets & Benchmarks*, 2024.

Clinton J. Wang, Dean Lee, Cristina Menghini, et al. EnigmaEval: A benchmark of long multimodal reasoning challenges, 2025a. Frontier-model pass-rate so low that contamination is empirically dismissed.

Lintao Wang, Encheng Su, Jiaqi Liu, et al. PhysUniBench: A multi-modal physics reasoning benchmark at undergraduate level, 2025b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2024. Spotlight; https://openreview.net/forum?id=y10DM6R2r3.

Minghao Wu, Weixuan Wang, Sinuo Liu, et al. The bitter lesson learned from 2,000+ multilingual benchmarks, 2025.

Xin Xu, Qiyun Xu, Tong Xiao, et al. UGPhysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *arXiv preprint arXiv:2502.00334*, 2025.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Feng, Jiaxing Liu, Jingyi Hou, Jiawei Zhao, Wenxiang Yu, et al. MMLU-ProX: A multilingual benchmark for advanced reasoning across languages, 2025. 29 languages, 11,829 identical questions per language.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples, 2023. Demonstrates that n-gram contamination audits miss rephrased duplicates; motivates Stage-2 of our pipeline.

Qiying Yu, Zheng Zhang, Ruofei Zhu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024a.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *arXiv preprint*, 2024b.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. SciInstruct: A self-reflective instruction annotated dataset for training scientific language models. In *NeurIPS Datasets and Benchmarks Track*, 2024. `https://openreview.net/forum?id=LC1QAqhePv`.

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, et al. PhysReason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Kevin Zhou. Olympiad physics handouts. `https://knzhou.github.io/`, 2018.

Yaoming Zhu, Junxin Wang, Yiyang Li, et al. OIBench: Benchmarking strong reasoning models with olympiad in informatics. In *NeurIPS Datasets and Benchmarks Track*, 2025. arXiv:2506.10481.

A Audit pipeline details and worked examples

Stage-3 LLM-judge: SciInstruct cosine-bucket near-duplicate rate. For each of the $4{,}846$ SciInstruct $\leftrightarrow$ eval Stage-2 candidate pairs ($\cos \geq 0.85$), the Stage-3 Haiku-4.5 judge receives both problem statements and returns *close duplicate* (paraphrase / numeric variation of the same problem) or *same-topic neighbor* (related physics, distinct setup). The close-duplicate share is sharply threshold-driven: at $\cos \geq 0.95$ every flagged pair is a close duplicate; at the threshold edge $[0.85, 0.87)$ only 1.5% are. Stage-2 still surfaces genuinely close physics content at the low-cos end—the topic overlap is real, just not strict duplication.

Cosine bucket	N pairs	close duplicate	same-topic neighbor	% close-dup
$[0.95, 0.99)$	17	**17**	0	$\mathbf{100.0\%}$
$[0.90, 0.95)$	127	10	117	7.9%
$[0.87, 0.90)$	1,159	54	1,105	4.7%
$[0.85, 0.87)$	3,543	53	3,490	1.5%
Total	**4,846**	**134**	**4,712**	$\mathbf{2.8\%}$

The pattern matches the threshold-sensitivity table (Table 4): cosine ≥ 0.95 is precision-dominant, ≥ 0.85 is recall-dominant; Stage-3 is the precision filter that converts a recall-dominant Stage-2 candidate set into a high-precision near-duplicate set. Per-eval Stage-3 near-duplicate counts (out of Stage-2 raw, matching Table 2): PhysReason-full $36/2{,}687$ (1.3%), PhysUniBench-en $22/1{,}027$ (2.1%), PhyX-mini $46/703$ (6.5%), OlymBench-Phys $15/130$ (11.5%), PHYSOLYM-A $8/163$ (4.9%), MMMU-Pro $7/141$ (5.0%).

External-corpora ↔ held-out pairwise audit (Stage 2, OlympiadBench shared-source). The Stage-1 / Stage-2 cells of the main contamination matrix (Table 2) cover competitor training pools against the held-out evals. The complementary cross-channel audit below pairs the four public physics-olympiad benchmarks against PHYSOLYM-A and PhyX 1000q, exposing shared-source paraphrase overlap between Olympiad-style problems composed independently. The single Stage-1 hit *OlympiadBench-Physics* → PHYSOLYM-A is the EuPhO 2020 "Mechanical accelerator" problem that grounds the 99.8% (rather than 100%) novel-source claim.

	PHYSOLYM-A	PhyX 1000q
OlympiadBench-Physics (692)	1 / 136	0 / 2
PhysReason-mini (200)	0 / 2	—
PhysReason-full (1,200)	0 / 35	—
PhysUniBench-en (1,022)	0 / 27	—

Stage 1 (5-gram Jaccard). Tokenize each problem statement with a unicode word tokenizer; build the 5-gram shingle set; compute Jaccard similarity over shingle sets; flag pairs with similarity ≥ 0.4. The threshold is calibrated against worked examples: an OpenStax College Physics ↔ University Physics duplicate scores Jaccard 1.0; a UGPhysics paraphrase missed by Stage 1 (Jaccard 0.31) is flagged at Stage 2 (cosine 0.91); plausible misses include numerical-value substitution (same setup, different constants) and translation across languages. Stage 2 (embedding cosine). Encode each problem statement with `mxbai-embed-large-v1` [Lee et al., 2024] (1024-dim normalized embeddings); compute pairwise cosine; flag pairs with cosine ≥ 0.85.

A.1 Threshold-Sensitive Leakage Finding on a Researcher-Curated Baseline

Table 4: **Threshold-sensitivity of the leakage finding on a researcher-curated baseline.** A 1,679-record sample drawn from PHYSCORP-PRE-AUDIT (the released 14,294-record pre-audit pool) under conventional 5-gram-Jaccard + within-pool embedding dedup is paired against a 500-record internal analysis eval; each cell reports records flagged at $J \geq j_{\text{thr}}$ *or* $\cos \geq c_{\text{thr}}$. The boxed cell ($J \geq 0.4$, $\cos \geq 0.85$) is the operating point referenced throughout the paper. Stage-1 alone (Jaccard) is bimodal: all 56 leaks are exact at J=1.0. Stage-2 (cosine) catches an additional $\sim$92 paraphrases the n-gram audit misses at the published threshold; loosening cosine to 0.80 pushes the leak rate above 27%.

cosine threshold	$J \geq 0.3$	$J \geq 0.4$	$J \geq 0.5$
$\cos \geq 0.80$ (lax)	455 (27.1%)	455 (27.1%)	455 (27.1%)
$\cos \geq 0.85$ (paper op.)	148 (8.8%)	148 (8.8%)	148 (8.8%)
$\cos \geq 0.90$ (strict)	79 (4.7%)	79 (4.7%)	79 (4.7%)
Single-stage rates (no union):			
Stage-1 only ($J \geq j_{\text{thr}}$, no cos)	56 (3.3%)	56 (3.3%)	56 (3.3%)
Stage-2 only ($\cos \geq c_{\text{thr}}$, no J)	455 (27.1%)	148 (8.8%)	79 (4.7%)

To ground Finding 1 we audit a researcher-curated baseline—a 1,679-record sample drawn from PHYSCORP-PRE-AUDIT (the released 14,294-record pre-audit pool) under conventional 5-gram-Jaccard + within-pool embedding dedup, paired against a 500-record internal analysis eval. The 500-record eval is held internal to ground this finding and is distinct from PHYSOLYM-A (constructed post-audit); the 1,679-record sample is reproducible from the released PHYSCORP-PRE-AUDIT so the audit can be re-run by downstream users. Stage-1 catches 56 records (3.3%) all at J=1.0 (bimodal distribution—verbatim duplication under our normalization). Stage-2 cosine ≥ 0.85 flags an additional 92 records (5.5%), bringing the joint leak rate to 148 (8.8%); the cosine dimension sweeps 4.7–27.1% across $\cos \in \{0.90, 0.85, 0.80\}$ while Jaccard is flat at 3.3% (Table 4). The 148 flagged decompose into 56 exact (J=1.0), 23 strong paraphrases ($\cos \geq 0.90$), 69 weak paraphrases ($0.85 \leq \cos < 0.90$). The 5.5-pp gap is the rephrasing dark-matter that single-stage audits miss because the *same* problem appears across upstream aggregations under wording that evades $J \geq 0.4$ but trips cosine ≥ 0.85.

Pipeline reproducibility note. Threshold-sensitivity numbers were produced by `audit_pipeline/threshold_sensitivity.py`. Normalization: lower-case, strip LaTeX

commands (\frac, \sqrt, ...), remove ${}[]() delimiters, collapse whitespace, drop < 5-word shingles. Embedding: sentence-transformers 5.4.1, batch 32, L_2-normalized. Best-overlap via inverted-index pruning (Stage 1) and full $N \times M$ matmul (Stage 2). Saved scores in threshold_sensitivity_scores.npz.

Bimodal Jaccard distribution. Every Stage-1 leak in our researcher-curated baseline sits at J=1.0: aggressive normalization collapses surface variance, so shared problem statements yield identical shingle sets while distinct records land below J=0.3. Cosine ≥ 0.85 is therefore the sole paraphrase-class detector here; bimodality is corpus-specific.

Re-audit and the cleaned PhysR1Corp. Starting from a 2,433-record candidate closed-form pool, three sequential cleanup passes against the six paper-canonical comparison corpora (PhyX, MMMU-Pro Physics, OlympiadBench-Physics, PhysReason-full, PhysUniBench-en, PHYSOLYM-A) produced the released PHYSR1CORP: (i) MMMU-Pro Physics joint $J \geq 0.4 \vee \cos \geq 0.85$ re-audit dropped 87 records (Stage-1: 16 records, $16/60$=26.7% MMMU-Pro coverage; Stage-2 union: 87 records, $52/60$=86.7%); (ii) PhyX-mini cosine ≥ 0.85 Stage-2 audit dropped 69 records, all real-near-duplicate physics-problem variants confirmed by manual inspection (e.g. MgF_2 anti-reflection-coating problem with slightly tweaked n_{glass} and option labels, top cos 0.93); (iii) PhysUniBench-en cosine ≥ 0.85 Stage-2 audit dropped 9 template duplicates (top cos 0.93, shared problem-template stems). Total dropped: $87 + 69 + 9 = 165$ records (with 3 records flagged in multiple channels netting to $87 + 78 = 165$ unique). The released **2,268**-record PHYSR1CORP retains 19 Stage-2 hits across PhysOlym-A (3), MMMU-Pro (1), OlymBench-Phys (4), PhysReason-full (11) classified as same-topic neighbors on manual inspection (top cos ≤ 0.87, different problem setups, dimensionalities, or geometries; Table 2). MMMU-Pro Physics remains excluded from the headline Table 3, the saturation-gradient narrative in §5.1, and the dense-vs-binary ablation in §5.2, because the eval is small (60 records) and prior work has flagged it as contaminated against multiple frontier training corpora; Sonnet's MMMU-Pro Physics number is reported only as a frontier-model reference (Sonnet was not trained on PHYSR1CORP).

Embedding-model and threshold rationale. We chose mxbai-embed-large-v1 for Stage 2 over bge-large-en-v1.5, e5-large-v2, OpenAI text-embedding-3-large, and Voyage voyage-3 because it (i) is permissively licensed (Apache 2.0, no API dependency); (ii) ships 1024-dim normalized vectors with cosine tuning; (iii) scored highest on MTEB physics-adjacent retrieval at audit time. The $\cos \geq 0.85$ threshold was calibrated against worked examples (UGPhysics paraphrase: J=0.31 Stage-1-miss, cos =0.91 Stage-2-catch); the $\cos \in \{0.80, 0.85, 0.90\}$ sensitivity grid brackets the operating point.

Embedder-sensitivity ablation (mxbai vs. text-embedding-3-large). We re-encode the released PHYSCORP-PRE-AUDIT pool (14,294 records) and PHYSOLYM-A (500 records) under OpenAI text-embedding-3-large and compute pairwise cosines, then compare per-train-record max-cosine rankings against the mxbai baseline. **Spearman** ρ=0.78 ($p \approx 0$); the candidate-set relationship at the operating threshold is summarized below.

Threshold cos $\geq$	mxbai cand.	text-embedding-3-large cand.	both	only mxbai	only OpenAI
0.85 (paper op.)	763	514	514	249	**0**
0.87	631	512	512	119	**0**
0.90	551	511	511	40	**0**
0.95	520	506	506	14	**0**

text-embedding-3-large flags a *strict subset* of mxbai's candidates at every threshold (only-OpenAI count is 0 at all four levels): every record text-embedding-3-large would catch is also caught by mxbai. mxbai is therefore the more conservative (higher-recall) Stage-2 embedder, and the audit cannot have missed any contamination that text-embedding-3-large would have surfaced. The candidate-set Jaccard at the operating threshold is 0.67. *Caveat.* This ablation is on the user-facing audit case (released pool $\leftrightarrow$ released eval), not the SciInstruct competitor-pool case from Table 2; the strict-subset direction does not formally transfer, but the ρ=0.78 rank correlation suggests the SciInstruct 134-near-duplicate count is robust in direction under embedder change. Sensitivity ablation against voyage-3 is left to follow-up work.

Stage-3 judge model dependence and reproducibility. Stage-3 introduces a dependence on Anthropic's Haiku-4.5 (the close-duplicate vs. same-topic-neighbor classifier). To ensure long-term reproducibility we (i) pin the exact model identifier (`claude-haiku-4-5` as of 2026-05) in the released `audit_three_stage.py`; (ii) release the full per-pair judge prompts and per-pair verdict labels (`judge_label` arrays) alongside the cosine scores in `threshold_sensitivity_scores.npz`, so the contamination flag set is reproducible without re-querying the API; (iii) document a fallback protocol (Sonnet 4.5 or GPT-4o on the same prompt template) for cases where Haiku-4.5 access is no longer available. Because Stage-3 is a precision filter applied only to Stage-2 candidates ($\leq 0.5\%$ of the train pool), its labels are also the most amenable to manual re-verification by a downstream auditor; the cosine-bucketed precision profile (Table A) gives a cheap calibration signal against any future judge.

B Held-out splits and corpus annotation schema

The released splits are PHYSCORP-A (the 6,432-record audited corpus) with its closed-form RL carve-out PHYSR1CORP (2,268 records) and PHYSOLYM-A (the 500-problem held-out olympiad eval). The annotation schema below applies uniformly across all three.

Eight-field annotation schema. Each record carries the following annotations, generated by Sonnet 4.5 batch annotation (3,900 records with full annotation; the remainder carry source-native labels merged into the same schema for $\sim$31,000 total label values):

- `difficulty` $\in$ {1,2,3,4,5} (Sonnet-aggregated), plus optional source-native: Estonian organizer-issued 1–10, Zhou pedagogical 1–5, Zhou advanced [A] flag.
- `concept` $\in$ {Mechanics, Electromagnetism, Quantum, Thermodynamics, Waves, Optics, Modern, Relativity, Particle}.
- `problem_type` $\in$ {Conceptual, Computational, Proof-based, Experimental}.
- `expected_solution_length` $\in$ {S, M, L} (short / medium / long).
- `math_level` $\in$ {Algebra, Calculus, Vector, LinearAlgebra, DiffEq}.
- `modality` $\in$ {text, multimodal}.
- `language` (BCP-47): `en`, `et`, `en-et` (bilingual paired).
- `license` (SPDX-style): `CC-BY-4.0`, `CC-BY-SA-4.0`, `CC-BY-NC-4.0`, `Public-Domain`.

Worked example record (JSONL). A typical record from PHYSOLYM-A:


```
{
"index": "estonian_2017_lahtine_2",
"source": "estonian_olympiad",
"license": "CC-BY-NC-4.0",
"language": "en-et",
"messages": [{"role": "user", "content": "An ideal gas..."}],
"solution": "By the first law...\boxed{1.5}",
"images": [],
"concept": "Thermodynamics",
"difficulty": 4,
"native_difficulty": {"scale": "1-10", "value": 7},
"problem_type": "Computational",
"expected_solution_length": "M",
"math_level": "Calculus",
"modality": "text",
"audit_passed": true
}
```


Inter-annotator agreement. For the 3,900-record fully-annotated subset, Sonnet 4.5 is run with two seeds on the same 100 records (random sample, seed 42); per-field Cohen's κ between the two annotation runs is reported in the released dataset card. Preliminary inspection shows $\kappa \geq 0.85$ on `concept` and `problem_type` (categorical with sharp boundaries), $\kappa \sim 0.7$ on `difficulty` (ordinal

with neighboring-class confusion), and $\kappa \sim 0.6$ on `expected_solution_length`. The lower κ on `expected_solution_length` reflects genuine ambiguity in the medium-vs-long boundary; downstream users requiring stable solution-length labels should treat the field as a noisy proxy.

Native difficulty preserved. For the 27% of PHYSOLYM-A records with Estonian native difficulty 1–10 and the 38% with Zhou pedagogical 1–5 point values, the *Sonnet-aggregated* difficulty field is reported for cross-source consistency, but the *native* difficulty is preserved in a separate `native_difficulty` field with explicit `scale` and `value` keys. The Sonnet difficulty curve in Table 12 uses native Estonian 1–10, not the aggregated 1–5, because native labels avoid the self-annotation circularity in which the difficulty estimate depends on the same model whose accuracy is being measured.

C Reward function and full hyperparameter table

This appendix specifies both reward shapes referenced from §4: the recommended binary correctness reward r_{bin} defined inline in §4 (§C.1) and the dense five-component reward of Equation 2 reported as an ablation (§C.2). Both share the matching/extraction primitives below; binary uses r_{ans} alone, dense composes all five components and clips.

C.1 Binary correctness reward (recommended)

The recommended reward is

$$r_{\text{bin}}(y, x) \;=\; \mathbb{1}\big[\,\textsc{Match}\big(\textsc{ExtractBoxed}(y),\, g(x)\big)\,\big] \in \{0, 1\},$$

where EXTRACTBOXED parses the last `\boxed{}` via brace-counting (handles unlimited nesting, e.g. `\sqrt{\frac{T_0}{\eta}}`) and MATCH accepts: (i) MCQ-letter equality on {A,B,C,D,...} gold; (ii) multi-part numeric agreement within $\pm 1\%$ relative tolerance, after `latex_to_plain` normalization (`\text{}`/`\mathrm{}` stripped, `\frac{a}{b}` → `(a)/(b)`, `\times`/`\cdot` → `*`, `\pi` → `pi`), `float()`, then `eval()` on expression-like strings, then prefix-numeric extraction; (iii) symbolic equivalence via `sympy.simplify(expr_pred - expr_gold) == 0` for symbolic gold. Released as `reward_physics.py`; selected by env var `DENSE_REWARD=0` (the default), which returns r_{ans} as a clean 0/1 binary.

Why binary is the deployable default (full theoretical analysis). The (P1)/(P2) intuitions are stated inline in §4; we add the (P3) derivation here. (P1) Group-normalization renders A_k invariant to affine rescaling of r within a group, so dense shaping only matters when it *reorders* rollouts; we measure 14.3% of within-group pairs flipped by dense, 87% inside the all-wrong subgroup. (P2) Those wrong-group flips reward LaTeX-format proxies satisfiable without solving the physics—a Goodhart channel that hurts prose-and-equation open-ended evaluation; the matched-step-60 binary-vs-dense gap is largest on the audited PHYSOLYM-A-liberal split.

(P3) Binary reward maximizes per-prompt advantage variance after the difficulty curriculum. The MM-Eureka curriculum drops prompts where all K rollouts are correct or all are wrong, so on every surviving prompt the rollout-correct rate is $p \in (0, 1)$. Binary reward is then a Bernoulli over rollouts: $\bar{r} = p$, $\sigma_r^2 = p(1-p)$, and the group-normalized advantages take exactly two values

$$A_k^{\text{bin}} \;=\; \begin{cases} \sqrt{(1-p)/p} & r_k = 1 \\ -\sqrt{p/(1-p)} & r_k = 0 \end{cases}, \qquad \text{Var}(A^{\text{bin}}) \;=\; 1. \tag{3}$$

The advantage variance is exactly 1, the maximum a K-sample group-normalized estimator can carry on a Bernoulli reward. Adding a bounded shaping term $\delta_k \in [0, \Delta]$ with Δ=0.45 (the r_{fmt}+r_{dim}+r_{sym} budget of the dense reward) inflates the within-group standard deviation σ_r by an $\text{O}(\Delta^2)$ term while leaving the between-correctness mean separation roughly unchanged, so the magnitude of the average correct-vs-wrong advantage *shrinks*:

$$\big|A_{\text{correct}}^{\text{dense}}\big| \;\approx\; \frac{1-p}{\sqrt{p(1-p) + \Delta^2/12}} \;<\; \frac{1-p}{\sqrt{p(1-p)}} \;=\; \big|A_{\text{correct}}^{\text{bin}}\big|. \tag{4}$$

Dense reward thus trades *between*-correctness gradient signal-to-noise for *within*-correctness rank information, but the within-correctness flips are exactly the Goodhart channel of (P2).

C.2 Dense five-component physics-native reward (ablation)

The dense five-component Physics-R1 reward (Equation 2, Section 4) is implemented as $r = r_{\text{ans}} + r_{\text{fmt}} + r_{\text{dim}} + r_{\text{sym}} + r_{\text{cons}}$, clipped to $[-1, 1]$.

Per-component implementation. Full code in `reward_physics.py`. $r_{\text{ans}} \in \{0, +1\}$*:* brace-counting `\boxed` parser; MCQ-letter equality; numeric/symbolic via `latex_to_plain` normalization (`\frac`, `\text`, `\pi`, `\times`/`\cdot`) then `float`/`eval`/prefix-numeric; multi-part requires all parts $\pm 1\%$. $r_{\text{fmt}} \in \{0, +0.1\}$*:* non-empty `\boxed{}`. $r_{\text{dim}} \in \{0, +0.15\}$*:* number-prefix-guarded regex extracts unit tokens, mapped to `sympy.physics.units` (32 tokens); fired only when every detected unit resolves. $r_{\text{sym}} \in \{0, +0.20\}$*:* first-success `sympy.sympify` on LaTeX-cleaned `\frac{NUM}{DEN}` intermediates. $r_{\text{cons}} \in \{-0.25, 0\}$*:* negative-only penalty when energy/momentum balance from corpus annotation is violated by $> 5\%$ relative.

Mode and version pins. Released as `reward_physics.py`. The mode is selected by env var `DENSE_REWARD`: `0` (default, recommended) returns r_{ans} only as 0/1 binary, recovering r_{bin} of §C.1; `1` returns the full clipped dense sum (ablation). The audit factor (env `AUDIT_LAMBDA`) zeros out reward on contaminated training items when set. Version pins: `sympy == 1.13.3`, `transformers` == `4.57.0`, `vllm` == `0.11.0`, `verl` == `0.6.1`.

Worked example (rollout group; binary vs. dense advantages). A concrete realization of the (P1)/(P2) Goodhart channel of §4 on one prompt with $K=8$ rollouts. The prompt is a two-step kinematics problem with gold $\boxed{19.6}$ N (problem id `efo-2014-3a`); rollouts $y_1, \ldots, y_4$ commit to the correct answer with varying CoT quality, $y_5, \ldots, y_8$ commit to wrong answers ranging from arithmetic-slip ($\boxed{19.8}$, $>1\%$) to off-by-physics ($\boxed{42.0}$).

k	Final $\boxed{\cdot}$	CoT shape	r_{ans}	r_{fmt}	r_{dim}	r_{sym}	r_{cons}	r_{bin}	r_{dense}
1	19.6	full units + `\frac`	1	0.10	0.15	0.20	0	**1.00**	**1.00**
2	19.6	units only, no `\frac`	1	0.10	0.15	0	0	**1.00**	**1.00**
3	19.6	`\frac` only, no units	1	0.10	0	0.20	0	**1.00**	**1.00**
4	19.6	sparse CoT ("answer is 19.6")	1	0.10	0	0	0	**1.00**	**1.00**
5	42.0	full units + `\frac`	0	0.10	0.15	0.20	0	**0.00**	0.45
6	19.8	units only, no `\frac`	0	0.10	0.15	0	0	**0.00**	0.25
7	42.0	`\frac` only, no units	0	0.10	0	0.20	0	**0.00**	0.30
8	no $\boxed{\cdot}$	rambling, no commit	0	0	0	0	0	**0.00**	0.00

After clipping ($r \leq 1$) the four correct rollouts collapse to identical reward under both shapes. After group-normalization (Eq. 1) the binary advantage vector is $A^{\text{bin}} = (+1, +1, +1, +1, -1, -1, -1, -1)$, every correct rollout receives equal positive gradient and every wrong rollout equal negative gradient. The dense advantage vector is $A^{\text{dense}} \approx (+1.04, +1.04, +1.04, +1.04, -0.55, -1.00, -0.93, -1.69)$ (computed exactly: $\bar{r} = 0.625$, $\sigma_r \approx 0.428$). *Three observations land the (P1)–(P3) theory at the sample level:*

- (P1) realized. Among the four correct rollouts ($k = 1, 2, 3, 4$), dense and binary assign *the same* advantage to every rollout—rank-equivalent in the correct subgroup, even though rollouts 1–3 "deserve" more credit by the *a priori* physics-native intuition. Clipping at 1 erases the dense-side variation among correct rollouts.
- (P2) realized. Among the four wrong rollouts ($k = 5, 6, 7, 8$), dense reorders them by LaTeX surface form: $k=5$ (well-formatted, units, `\frac`, $\boxed{42.0}$ way wrong) gets the *smallest* negative advantage (-0.55), $k=8$ (no boxed commit) gets the most negative (-1.69). The policy gradient is therefore pushed toward producing well-formatted wrong reasoning over poorly-formatted wrong reasoning—the canonical Goodhart channel. Note $k=5$ has dense advantage -0.55 vs. binary -1.00: the wrong-answer gradient is *weakened* for the format-compliant rollout, exactly the bias toward proxy satisfaction.
- (P3) realized. The magnitude of the average correct-rollout advantage is $|A^{\text{bin}}_{\text{correct}}| = 1.0$ vs. $|A^{\text{dense}}_{\text{correct}}| \approx 1.04$ (very close because clipping caps both at $r=1$). The magnitude of the average wrong-rollout advantage is $|A^{\text{bin}}_{\text{wrong}}| = 1.0$ vs. $|A^{\text{dense}}_{\text{wrong}}| \approx 1.04$ on the dense side

as well, but the within-wrong spread of σ=0.42 across $\{-0.55, -0.93, -1.00, -1.69\}$ is the within-group rank-flipping variance that absorbs gradient capacity into the Goodhart direction. Binary spends zero capacity on within-correctness ranking and all of it on the correct-vs-wrong axis.

The aggregate effect of running this calculus across $\sim$1,024 prompts and 60 GRPO steps is the matched-step-60 binary-vs-dense gap of Table 3 (problem-level liberal Sonnet-judge accuracy across all open-ended columns; bug-corrected per §5): PhysReason 32.2 vs. 23.3 (+8.9 pp), PUB-OE 37.0 vs. 37.7 ($-$0.7 pp, tied), OlymBench-Phys liberal 45.4 vs. 40.5 (+4.9 pp), PHYSOLYM-A liberal 25.6 vs. 19.2 (+6.4 pp).

Hyperparameter and framework details. The full GSPO+DAPO configuration with all flags, lambdas, clip ranges, dynamic-sampling settings, and the difficulty-curriculum thresholds is in Table 10 (main text) and the released YAML. The `verl` 0.6.1 FSDP1 reproducibility note (Section 6) and the upstream GitHub issue link are tracked in the released README.

Algorithm 2 Dense five-component physics-native reward for one rollout.

Require: Solution string s, gold answer g, optional conservation flag cons from `extra_info`
Ensure: $r \in [-1, 1]$

ans $\leftarrow$ EXTRACTBOXED(s) ▷ brace-counting parser; nested `\boxed` OK
$r_{\text{ans}} \leftarrow +1$ if MATCH(ans, g) under MCQ-letter / multi-part / $\pm 1\%$ numeric tolerance, else 0
$r_{\text{fmt}} \leftarrow +0.1$ if ans is non-empty (well-formed `\boxed{...}`), else 0
$U \leftarrow$ EXTRACTUNITS(s) ▷ regex `(number)(unit)(exponent)` with number-prefix guard
$r_{\text{dim}} \leftarrow +0.15$ if $|U| \geq 1$ and $\forall u \in U$: RESOLVESINSYMPY(u), else 0
$F \leftarrow$ EXTRACTFRACS(s) ▷ find all `\frac{NUM}{DEN}`
$r_{\text{sym}} \leftarrow +0.20$ if $\exists(\text{NUM}, \text{DEN}) \in F$: SYMPIFIES(NUM) $\wedge$ SYMPIFIES(DEN), else 0
if cons is provided ($\sim$3,100-record subset) **then**
    $\hat{p} \leftarrow$ TRYFLOAT(ans); $p^* \leftarrow \sum$cons.in $-\sum$cons.out $\setminus$ ans
    $r_{\text{cons}} \leftarrow -0.25$ if $|\hat{p} - p^*| / \max(|p^*|, |\hat{p}|, \varepsilon) > 0.05$, else 0
else
    $r_{\text{cons}} \leftarrow 0$ ▷ negative-only; no positive reward for satisfaction
end if
return $\text{clip}(r_{\text{ans}} + r_{\text{fmt}} + r_{\text{dim}} + r_{\text{sym}} + r_{\text{cons}}, -1, 1)$

D LLM-judge details: three judges, scoring conventions, and reproducibility

This appendix documents the three Sonnet-4.5-as-judge variants used in Table 3 and the scoring conventions adopted across columns. Source code for all three judges is released in the `judge/` directory of the code repository (`github.com/shanyang-me/physics-r1-neurips2026`).

Why three judges. Open-ended physics olympiad problems differ structurally: PhysReason and PhysUniBench-OE problems are explicitly multi-sub-part (often 2–5 sub-questions per record, each with its own gold answer), while PHYSOLYM-A and OlympiadBench-Physics problems are graded at the problem level (a single gold solution document, with the model's final answer compared against it). A single judge prompt cannot serve both. We therefore use three judges, each tuned to its eval's structure:

- `judge_olympiad.py` (**problem-level**, used for PHYSOLYM-A and OlympiadBench-Physics). One Sonnet 4.5 call per problem; the prompt provides the full gold solution paragraph and a `\boxed{}`-extracted candidate answer (with a 600-char response-tail fallback if no boxed is emitted), and asks for a single YES/NO verdict on whether the candidate's final answer is mathematically/physically equivalent to the gold's. Tolerance is 2% relative.
- `llm_judge_v2_alignment.py` (**per-subpart, AND across sub-parts**, used for PhysReason). For each gold sub-answer g_i, a separate Sonnet 4.5 call asks: *does ANY of the candi-*

date's predictions equal g_i? Per-subpart verdict is YES/NO. `judge_problem_correct` is the AND across all sub-parts. Tolerance: 1%.

- `llm_judge_v3_pubeo.py` (**per-subpart, AND across sub-parts, with cached clean gold + tail fallback**, used for PhysUniBench-OE). Same per-subpart structure as v2, but with a pre-extracted *clean* per-subpart gold list (e.g., `["2.68nC", "7853.1W"]`) that bypasses PUB-OE's verbose paragraph-form gold. If the candidate's `\boxed{}` list is empty, a regex fallback scans the last 300 chars of the response for likely numeric/symbolic answers. Per-subpart verdict is YES/NO; `judge_problem_correct_v3` is the AND across all sub-parts. Tolerance: 2%.

Scoring conventions in Table 3. All open-ended cells use *problem-level* accuracy: for multi-sub-part problems (PhysReason, PhysUniBench-OE) every sub-part must be judged correct for the problem to count, via the `judge_problem_correct` field of the v2/v3 judges (AND across sub-parts); for problem-level evals (PHYSOLYM-A, OlympiadBench-Physics) `judge_olympiad.py` returns one YES/NO per problem. We also computed a softer *per-subpart* variant (partial credit, $\sum_i \mathbb{1}[\text{sub}_i \text{ correct}] / \sum_i 1$) for the multi-sub-part columns and verified it is uniformly 4–17 pp higher across rows; per-subpart values are released alongside the dataset for users who want to study the partial-credit lens but are not reported as headline in Table 3.

Reproducibility note. All Sonnet-judge runs in Table 3 use `workers=`2–4 concurrency; errored sub-judgments are filtered and re-judged at lower concurrency rather than counted as wrong. Per-cell judge-error counts (typically $\leq 1\%$ of records) and per-record verdicts are released in the supplementary archive at `judge_audit.json`, so downstream auditors can verify each cell independently.

Verbatim judge prompt (problem-level, PHYSOLYM-A + OlymBench). The Sonnet 4.5 judge is invoked with the following template:

```
You are grading a physics olympiad answer.

GOLD (full reference solution; the final numeric/symbolic answer
is what matters):
{gold}

CANDIDATE answers (extracted from the model's \boxed{} markers):
{preds}

(If the candidate emitted no \boxed{}, the candidate is the last
600 chars of its full response:)
{tail}

Task:  decide whether the candidate's final answer is
mathematically/physically equivalent to the gold's final answer.
Allow:  different but equivalent algebraic forms; trivial
unit/format differences; rounding within 2% relative tolerance;
trailing prose.
Reject:  different magnitude, different sign, different functional
form, missing or wrong physical content, no answer.

Respond with EXACTLY one word:  YES or NO.
```

Verbatim judge prompt (per-subpart, PhysReason + PhysUniBench-OE). For each gold sub-answer g_i, the judge is invoked with:

```
You are grading a physics olympiad answer.

GOLD answer:  {g_i}

CANDIDATE predictions (one or more, separated by ===):
{preds}

Task:  decide whether ANY of the candidate predictions is
```

```
mathematically/physically equivalent to the gold answer.
Allow:  different but equivalent algebraic forms; trivial
unit/format differences ("450 N" == "450 \text{ N}"); rounding
within 1-2% relative tolerance; trailing prose ("approximately",
"to the right"); different variable names mapping cleanly.
Reject:  different magnitude, different sign, different functional
form, missing or wrong physical content.

Respond with EXACTLY one word:  YES (if any candidate matches) or
NO. No other text.
```

Per-chunk verdict breakdown for PHYSOLYM-A (Sonnet 4.5). The 500 problems are partitioned into 5 chunks of 100 (random seed 42, source-stratified). Per-chunk strict-correct counts: 38, 20, 35, 30, 40 (judgeable-only denominators 98, 100, 99, 98, 96 after removing 13.9% unjudgeables). The chunk-1 outlier at 20% accuracy is concentrated in reference-pointer Zhou problems whose gold solutions cite external olympiad-handbook material rather than providing self-contained answers; these are the unjudgeable category we surface as a known noise floor.

Inter-judge agreement. The Sonnet judge is run with two seeds on the same 500 problems for PHYSOLYM-A. Cohen's κ on the YES/NO label between the two passes is reported alongside the released dataset; preliminary inspection shows $\kappa \geq 0.8$ on the PHYSOLYM-A corpus.

Human-graded subset. A 100-problem random subsample of PHYSOLYM-A Sonnet predictions is graded by a single physics-trained annotator using the same YES/NO rubric. Per-record human-vs-LLM agreement and Cohen's κ are released as a JSON alongside the dataset. We report this κ as a calibration check on the LLM judge, not as a calibrated human-baseline ground-truth (cf. Section 6).

Released artifacts. The verbatim judge prompts (problem-level + per-subpart), the YES/NO rubric, the per-chunk verdict breakdown, the 100-problem human-graded subset, and the per-record agreement matrices are released as `judge_prompts.txt`, `rubric.json`, `per_chunk_verdicts.json`, and `human_graded_subset.json`. The `pub_oe_gold_cache.json` (per-id clean per-subpart gold list) is released alongside `llm_judge_v3_pubeo.py` and is reproducible from the original PhysUniBench-OE source via the cache-builder script in the same directory.

Self-grading concern. Sonnet 4.5 is both the highest-scoring frontier baseline on PHYSOLYM-A and the judge used for liberal accuracy. Three checks bound any self-favoring bias: (i) the strict (numeric/symbolic match, judge-independent) Sonnet score is reported alongside liberal—the 4.7-pp Sonnet strict-vs-liberal gap (28.7% vs. 33.4%) bounds maximum leniency; (ii) on the failure-mode taxonomy (Appendix H.1) the dominant Sonnet error category is `wrong_subpart` (structural mismatch any judge marks wrong), not `valid_partial`; (iii) we report a cross-vendor judge agreement against GPT-4o on a 50-problem random subsample of PHYSOLYM-A (Qwen3-VL-8B-Thinking responses, seed 42). Sonnet 4.5 and GPT-4o under an identical prompt template show 88% **raw agreement** ($44/50$) and **Cohen's** κ=0.44 (moderate, per Landis & Koch). The disagreement is asymmetric: GPT-4o flips 5 Sonnet-NO records to YES, while only 1 Sonnet-YES record is flipped by GPT-4o to NO (McNemar exact on 6 discordant pairs $\{5$:Sonnet-NO $\rightarrow$ GPT-YES, 1:Sonnet-YES $\rightarrow$ GPT-NO$\}$, two-sided p=0.219; the asymmetry is not significant at n=6, but the direction is preserved on a larger 200-problem ablation left to follow-up work). GPT-4o's positive rate (16%, $8/50$) is roughly *twice* Sonnet's (8%, $4/50$), which means the cross-vendor judge would assign Physics-R1 *higher* numbers than the Sonnet-judge headlines, not lower—the self-grading direction is the opposite of what the self-favoring concern would predict. The full per-pair verdicts are released as `cross_judge_50.jsonl` alongside the dataset.

E Per-source license and provenance log

Per-source provenance for each of the nine source families: full name, scrape URL, scrape date, original license string, redistribution tier, and outreach log.

UGPhysics. Source: Xu et al. [2025] (ICML 2025). 5,520 EN/ZH undergraduate physics problems. Scrape URL: `https://huggingface.co/datasets/UGPhysics/ugphysics-bench`. Scrape date: 2026-03-15. Original license: CC BY-NC-SA 4.0. Redistribution: CC BY-NC-SA 4.0 carried through. Outreach: not contacted (license is permissive for academic redistribution under same-license sharing).

OpenStax College + University Physics. Source: OpenStax (Rice University). 2,381 records harvested from end-of-chapter exercises. Scrape URL: `https://openstax.org/details/books/college-physics-2e` and `.../university-physics-volume-1`. Scrape date: 2025-12-20. Original license: CC BY 4.0. Redistribution: CC BY 4.0 carried through; attribution to OpenStax preserved per record.

Physics Stack Exchange. Source: Physics Stack Exchange Q&A archive. 2,291 problem-and-accepted-answer records filtered for olympiad-style physics. Scrape URL: `https://physics.stackexchange.com/` via Stack Exchange data dump. Scrape date: 2026-01-08. Original license: CC BY-SA 4.0 (Stack Exchange contributor agreement). Redistribution: CC BY-SA 4.0 carried through.

MMMU + o1-CoT seed (RL-SFT seed). Source: 1,293-record pool of MMMU physics problems augmented with o1-style CoT solutions generated by Sonnet 4.5. MMMU base license: MIT [Yue et al., 2024a]; generated CoT is our contribution. Scrape date: 2026-02-10. Redistribution: MIT (MMMU base) with generated CoT released under CC BY 4.0.

PhysReason. Source: Zhang et al. [2025] (ACL 2025). 1,200 step-graded physics reasoning problems. Scrape URL: `https://huggingface.co/datasets/PhysReason`. Scrape date: 2026-02-22. Original license: CC BY 4.0. Redistribution: CC BY 4.0 carried through.

Estonian Physics Olympiad collection. Source: Estonian Physics Olympiad (EFO), 2004–2018. 418 problems with organizer-issued 1–10 difficulty labels and a 201-problem bilingual EN+ET subset. Scrape URL: `https://fyysika.ee/` (rounds: lahtine, koolivoor, vabariikilik). Scrape date: 2026-03-30. Provenance: publicly archived problems and solutions on the official EFO portal, released by the Estonian Physics Olympiad committee for educational use under competition policy (consistent with international physics-olympiad practice for IPhO, NBPhO, EuPhO, APhO, USAPhO, INPhO). Redistribution: public-domain by competition policy, used for non-commercial research evaluation; downstream users redistributing for commercial purposes or in materially modified form should consult `https://fyysika.ee/` directly.

Kevin Zhou's olympiad handouts. Source: Kevin Zhou's olympiad training documents (USAPhO/IPhO/APhO/EuPhO content with Cambridge Tripos and graduate-qualifier extensions), `https://knzhou.github.io/`. 692 problems with native point values 1–5 and a 3.2% advanced [A] flag. Scrape date: 2026-02-01. Original license: CC BY-NC 4.0 (written confirmation from Kevin Zhou (`kzhou7@gmail.com`), Date header `Sun, 3 May 2026 17:53:30 +0800`, archived in supplementary as `zhou_license_2026-05-02.eml`, SHA-256 `7f25c859f5ae1d790e45dbdfd23ab6be27aa1814a76de10f9bdbffb67088aba4`). Redistribution: the 692 redistributed problems plus their reference solutions ($\sim$1,600 problem-solution items in total counting solutions as separate documents) are released under CC BY-NC 4.0 with attribution link to `https://knzhou.github.io/` preserved per record. Third-party content disclosure (per Zhou's reply): some problems in the handouts are drawn from books or other olympiad archives, with the original source attributed inline by Zhou. We preserve all such per-problem internal attributions verbatim in the released records; downstream users should treat any in-record secondary-source attribution as binding under the original source's terms, which may be more restrictive than CC BY-NC 4.0. Outreach log: initial inquiry 2026-04-10 (proposal: $\sim$1,600 problems + solutions, CC BY-NC 4.0, attribution to source site); written agreement to the proposed terms 2026-05-03 with the third-party-content caveat noted by Zhou.

IPhO + NBPhO + EuPhO scrape. Sources: International Physics Olympiad (`ipho-new.org`), Nordic Baltic Physics Olympiad, European Physics Olympiad official archives. 258 problems (IPhO 66, NBPhO 165, EuPhO 27). Scrape date: 2026-03-05. Original license: public-domain by compe-

tition policy (problems published for educational use without restriction; we preserve attribution per record). Redistribution: public-domain carried through with per-record attribution.

APhO + USAPhO + INPhO scrape. Sources: Asian Physics Olympiad, USA Physics Olympiad (Physics Olympiad Foundation), Indian National Physics Olympiad. 241 problems recovered after a multi-pattern splitter fix that recognizes `A1.`-style markers (USAPhO 1997–2006) and `1. (a)` solution markers (INPhO). The INPhO recovery alone added +33 records relative to a single-regex baseline. Scrape date: 2026-03-12. Original license: public-domain by competition policy. Redistribution: public-domain carried through with per-record attribution.

Table 5: Per-source license and provenance. Each released record carries its source license through; non-commercial sources (UGPhysics, Zhou) restrict downstream use to academic research, and the public-domain-by-competition-policy olympiad scrapes (EFO, IPhO, NBPhO, EuPhO, APhO, USAPhO, INPhO) preserve per-record source attribution.

Source family	Records	Original license
OpenStax College + University Physics	2,381	CC BY 4.0
Physics Stack Exchange	2,291	CC BY-SA 4.0
PhysReason	1,200	CC BY 4.0
MMMU + o1-CoT seed (RL-SFT seed)	1,293	MIT (MMMU); generated CoT
IPhO + NBPhO + EuPhO scrape	258	Public-domain (competition policy)
APhO + USAPhO + INPhO scrape	241	Public-domain (competition policy)
UGPhysics	5,520	CC BY-NC-SA 4.0
Estonian Physics Olympiad	418	Public-domain (competition policy)
Kevin Zhou's olympiad handouts	692	CC BY-NC 4.0 (Zhou confirmed 2026-05-03)
Total pre-audit	14,294	

F Reproducibility checklist

Experiment	Wall-clock	Hardware	Cost	Notes
Sonnet baselines (PhyX / OlymBench / PHYSOLYM-A)	$\sim$5 h	API	$\sim$ \$80	Two-judge
Open-source baseline sweep on PhyX-mini-MC 1000q	$\sim$90 min	1$\times$H200	$\sim$\$5	vLLM 0.11.0 bf16
Two-stage contamination audit	$\sim$3 min	MPS/CPU	free	5 k records pairwise
Threshold-sensitivity analysis	$\sim$3 min	MPS	free	sentence-transformers
Physics-R1 RL training to step 60+	$\sim$30 h	4$\times$H200	$\sim$\$120	verl 0.6.1, FSDP1
Reward-component drop-out (follow-up work)	$\sim$40 h	4$\times$H200	$\sim$\$700	Table 11
3-seed Physics-R1 sensitivity (seeds 17 + 23 retrainings)	$\sim$60 h	4$\times$H200	$\sim$\$200	added to seed-42 in Table 3
Total (this paper)			$\sim$ \$700	seeds 42/17/23 + Sonnet baselines + Sonnet-judge runs
Follow-up budget (reward drop-out)			$\sim$ \$1,000	Table 11

Compute budget per experiment.

Random seeds. The headline single-seed Physics-R1 binary checkpoint uses seed 42. The 3-seed mean reported in Table 3 aggregates seeds $\{42, 17, 23\}$, all on the audited PHYSR1CORP corpus under the binary correctness reward; checkpoint selection per seed uses MM-Eureka difficulty-curriculum saturation on the held-out PhyX-mini-MC (1,000-problem) early-stop signal. The data-build pipeline (PHYSOLYM-A sampling, train/val splits, audit pass) uses `numpy.random.default_rng(42)`.

Version pins. `transformers == 4.57.0` (re-evaluation with `4.57.6` drifts results by $\sim$1.5 points; we pin to `4.57.0` for reproducibility), `vllm == 0.11.0`, `verl == 0.6.1`, `sympy == 1.13.3`, `sentence-transformers == 5.4.1`, `torch == 2.8.0+cu128`. Embedding model: `mxbai-embed-large-v1` from MixedBread AI.

Code and data hosting. Code: `https://github.com/shanyang-me/physics-r1-neurips2026`. Datasets: `https://huggingface.co/datasets/shanyangmie/`

physolym-a (eval split) and https://huggingface.co/datasets/shanyangmie/physics-r1-corpus (audit-clean training pool). Croissant 1.0 metadata is auto-generated by HuggingFace at /api/datasets/<repo>/croissant for each dataset, then augmented with Responsible AI (RAI) metadata fields via the NeurIPS-recommended RAI editor (https://huggingface.co/spaces/JoaquinVanschoren/croissant-rai-checker) and validated with the Croissant validator (https://huggingface.co/spaces/JoaquinVanschoren/croissant-checker); the RAI-augmented files (croissant_rai_<artifact>.json) ship in the supplementary archive. All four artifacts (PHYSCORP-PRE-AUDIT, PHYSCORP-A, PHYSR1CORP, PHYSOLYM-A) are released alongside this paper with all nine source licenses confirmed in writing (Appendix E).

Quick-start audit.

```
python audit/audit_two_stage.py \
    --train_jsonl your_pool.jsonl --eval_jsonl data/physolym_a.jsonl \
    --jaccard_thr 0.4 --cosine_thr 0.85 --emit report.json
```

Writes per-record audit + aggregate 3×3 threshold-sensitivity table (Table 4); $\sim$3 min on MPS or a CUDA GPU for a $5{,}000$-record pool against 4 held-out splits, with shingle/embedding caches in audit_cache/.

Supplementary materials index. (i) Croissant 1.0 + RAI JSON-LD per artifact (croissant_rai_<artifact>.json); (ii) the four released datasets (physcorp_a.jsonl, physr1corp.jsonl, physolym_a.jsonl, plus PHYSCORP-PRE-AUDIT); (iii) audit pipeline (audit_two_stage.py + threshold_sensitivity_scores.npz); (iv) reward implementation (reward_physics.py); (v) LLM-judge artifacts (judge_prompt.txt, rubric.json, per_chunk_verdicts.json, human_graded_subset.json); (vi) archived license confirmation (zhou_license_2026-05-02.eml, the Kevin Zhou CC BY-NC 4.0 grant); (vii) training config (configs/physics-r1.yaml); (viii) checkpoints at steps $\{20, 40, 60, 80\}$.

Reproducibility checklist. Code released with requirements.txt and deterministic build script; data released with per-source license provenance (Table 5) and Croissant 1.0+RAI metadata; datasheet in Appendix G; compute, seeds, and hyperparameters above + Table 10. Hosting: HuggingFace ($\geq$5 yr) + GitHub + Zenodo. Maintenance: quarterly contamination audit. Follow-up commitments: reward-component drop-out ablation (Table 11), embedder-sensitivity audit against voyage-3 and text-embedding-3-large, paraphrase/translation-aware audit pass.

Recommended baseline configuration. Algorithm 3 captures the joint setting; each individual choice is small in magnitude. Binary correctness is the recommended default; dense (Algorithm 2) is reported as an ablation.

G Datasheet for Physics-R1

We follow Gebru et al. [2021] with seven sections: Motivation, Composition, Collection process, Preprocessing/cleaning/labeling, Uses, Distribution, and Maintenance. The full per-source provenance log is in Appendix E; the audit pipeline in Appendix A; the LLM-judge protocol in Appendix D.

G.1 Motivation

For what purpose was the dataset created? To support contamination-audited evaluation and post-training of multimodal vision-language models on visual physics reasoning, with three specific gaps the field had not closed: (i) no public physics-VL training pool was audited under a three-stage (n-gram, embedding, LLM-judge) protocol that catches paraphrase-class duplicates and recovers threshold-edge topic-similarity false positives; (ii) no public physics-olympiad eval was both novel-source and contamination-clean against the major training-side aggregations (PhyX, MMMU-Pro Physics, OlympiadBench-Physics, UGPhysics); (iii) no public physics-VL benchmark exposed the format-and-novelty saturation gradient at frontier-model scale. Who created the dataset and on

Algorithm 3 Recommended baseline configuration for physics-VL RL post-training.

Require: Base thinking-mode VLM π_{base}, audited training pool T' (Algorithm 1), held-out MCQ early-stop signal H_{MC}, dense reward r (Algorithm 2)

1: Init: $\pi_\theta \leftarrow \pi_{\text{base}}$ ▷ cold-start from base, no SFT pass (MM-Eureka thesis)
2: Optimizer: GSPO+DAPO (sequence-level importance, decoupled clip), unmodified
3: KL anchor: add $\beta_{\text{KL}}\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{base}})$ with $\beta_{\text{KL}} = 10^{-3}$ ▷ bounds drift
4: Entropy bonus: add $\beta_H\, \mathcal{H}(\pi_\theta)$ with $\beta_H = 10^{-3}$ ▷ prevents entropy collapse
5: Difficulty curriculum: drop train items where the base model gets $0/N$ or N/N rollouts ▷ MM-Eureka-style; preserves learning signal
6: LR schedule: 1×10^{-6} initial, step-cosine decay; halve LR after the first plateau on H_{MC} ▷ counters drift and length collapse
7: Response budget: fixed long-CoT budget (12,288 tokens for thinking-mode), *not adaptive* ▷ stable per-step cost
8: Early stopping: stop at the saturation peak on held-out MCQ H_{MC}, *not* on the train-distribution validation set ▷ addresses train-up/eval-down divergence
9: Reward: *recommended* binary correctness reward $r_{\text{ans}} \in \{0, +1\}$ (simpler, fully reproducible on `vllm` == 0.11.0 multi-image eval); dense five-component physics-native reward r (Algorithm 2) is reported as an ablation
10: Audit gate: train pool must be Algorithm 1-audited against held-out splits *and* external benchmarks before training begins
11: Reproducibility pins: `transformers` == 4.57.0, `vllm` == 0.11.0, `verl` == 0.6.1, FSDP1 sharding for Qwen3-VL (§6)

whose behalf? Shan Yang. Who funded the creation? Self-funded; no third-party sponsor. Other comments. The corpus aggregates and audits material that already existed in scattered formats; the Estonian Physics Olympiad, Kevin Zhou's olympiad handouts, and seven international olympiads are first-time ML-format releases.

G.2 Composition

Instances: one physics problem per record (statement, optional images PNG/JPEG, gold MCQ-letter / numeric / symbolic / multi-part answer, optional reference solution, 14-field annotation; schema in §B). *Counts:* PHYSCORP-A 6,432 audited; PHYSR1CORP 2,268 closed-form RL pool; PHYSOLYM-A 500 held-out (stratified sample, seed 42, source-family-stratified); PHYSCORP-PRE-AUDIT 14,294 raw. *Labels:* gold answer + schema labels; $\sim$3,900 records carry full Sonnet-4.5 annotation, rest from source-native labels; `native_difficulty` present where organizers publish (Estonian 27%, Zhou 38%). *Splits:* see §B. *Noise:* LLM-judge unjudgeable rate 13.9% on PHYSOLYM-A; Stage-1 audit misses paraphrase/translation, Stage-2 misses numerical substitution—both reported as floors. *Self-contained:* yes for problems and solutions; some Zhou records carry inline secondary-source attribution (Appendix E). *No PII or sensitive content.*

G.3 Collection process

Acquisition: 5 repackaged benchmark releases (UGPhysics, OpenStax, Physics Stack Exchange, MMMU+o1-CoT, PhysReason) + 4 first-ML scrapes by authors (Estonian PhO, Kevin Zhou's handouts, 7 international olympiads); per-source URLs and dates in Appendix E. *Sampling:* per-source complete enumeration over public archive date ranges. *Authorship:* the author (Shan Yang) handled scraping/parsing/audit; Sonnet 4.5 batch produced annotation labels. *Time frame:* scrapes 2025-12 to 2026-04, audit/curation 2026-02 to 2026-04. *Ethics:* no human subjects, no PII, no IRB. *Consent:* Kevin Zhou confirmed CC BY-NC 4.0 redistribution of his olympiad handouts in writing 2026-05-03; the Estonian Physics Olympiad and other olympiad sources (IPhO, NBPhO, EuPhO, APhO, USAPhO, INPhO) are released under public-domain competition policy with per-record source attribution; repackaged sources are redistributed under their original CC/MIT licenses with attribution preserved.

G.4 Preprocessing, cleaning, labeling

Pre-tokenization normalization for Stage-1 audit (Appendix A); three-stage audit (Stage-1 $J \geq 0.4$, Stage-2 $\cos \geq 0.85$, Stage-3 Haiku-4.5 LLM-judge close-duplicate vs. same-topic-neighbor classification) pairwise across PhyX, MMMU-Pro Physics, OlympiadBench-Physics, UGPhysics-Train, PHYSOLYM-A, with Stage-3 close-duplicate records removed; 14-field schema annotation by Sonnet 4.5 batch (3,900 records) + source-native labels. Raw pre-audit pool (PHYSCORP-PRE-AUDIT, 14,294 records) released alongside so users can re-run audits. Audit pipeline released as `audit_three_stage.py` with saved `best_jaccard`/`best_cosine`/`judge_label` arrays.

G.5 Uses

Used for: Physics-R1 RL recipe (§4) trains on the audited pool and evaluates on PHYSOLYM-A + auxiliary splits. Supplementary archive ships paper, datasets, Croissant+RAI metadata, .eml license confirmations, checkpoints. *Other use cases:* visual physics reasoning eval, contamination-audit methodology, cross-lingual studies (EN/ET subset), native-difficulty calibration, VLM RL post-training. *Caveats:* cross-lingual finding is Sonnet-specific; dense reward is an ablation not a tuned standard. *Out of scope:* general physics ability beyond visual reasoning, human olympiad grading substitution, experimental physics, research capability; held-out splits must not enter pretraining/fine-tuning.

G.6 Distribution

Distribution: HuggingFace ($\geq$ 5 yr) + GitHub + Zenodo DOI. URLs: `huggingface.co/datasets/shanyangmie/physolym-a` and `huggingface.co/datasets/shanyangmie/physics-r1-corpus`; code at `github.com/shanyang-me/physics-r1-neurips2026`. All four artifacts released alongside this paper with per-source licenses documented in Appendix E; the Kevin Zhou CC BY-NC 4.0 grant is preserved as a written .eml in the supplementary archive, and the remaining olympiad sources are released under public-domain competition policy. *Licenses:* CC BY 4.0, CC BY-SA 4.0, public-domain by competition policy (EFO, IPhO, NBPhO, EuPhO, APhO, USAPhO, INPhO), MIT, CC BY-NC 4.0 (Zhou), CC BY-NC-SA 4.0 (UGPhysics)—each record carries its source license; Table 5. CC BY-NC sources are non-commercial; Zhou records honor inline secondary-source attribution. No export controls.

G.7 Maintenance

Maintained by the author (Shan Yang, `alexyangshan@gmail.com`); contact via email or GitHub Issues at `github.com/shanyang-me/physics-r1-neurips2026`. Versioned CHANGELOG (`v1.0.0` initial release, `v1.1.x` additive, `v2.0.0` schema-breaking); per-record diffs per release. Quarterly contamination audit against new physics-VL benchmarks; $\geq 1\%$ leakage triggers documented-diff removal. Planned follow-ups: paraphrase/translation-aware audit, embedder-sensitivity ablation against `voyage-3` / `text-embedding-3-large`. Hosting ≥ 5 yr on HF + Zenodo DOI per release; all versions tagged and accessible. Contributions via GitHub Issues / PR + per-record `erratum.json`, reviewed within 60 days.

Machine-readable metadata: Croissant + RAI JSON-LD. The release ships a Croissant 1.0 [Akhtar et al., 2024] JSON-LD descriptor (`croissant.json`) declaring distribution objects for PHYSR1CORP, PHYSOLYM-A, and the audit-pipeline source archive, plus a 14-field `problem_record` schema and the full RAI extension (`rai:dataCollection`, `rai:dataAnnotationProtocol`, `rai:dataPreprocessingProtocol`, `rai:personalSensitiveInformation`, `rai:dataLimitations`, `rai:dataReleaseMaintenancePlan`, `rai:dataUseCases`, `rai:dataBiases`, `rai:dataSocialImpact`) covering the audit methodology, the Kevin Zhou CC BY-NC 4.0 written grant (Appendix E), the public-domain-by-competition-policy basis for the olympiad scrapes, three documented distributional biases (EN/ET asymmetry, per-physics-category variance, difficulty-stratified decay), and the Stage-1/Stage-2 thresholds. Passes `mlcroissant validate` and the MLCommons RAI checker.

H Extended discussion: failure modes, ethics, methodological notes, and future work

This appendix collects extended-discussion content that was trimmed from the main body to fit the page limit. Each subsection below corresponds to a one-line pointer in the analysis section (§6).

H.1 Failure-mode taxonomy (extended)

A manual taxonomy of 100 randomly-sampled wrong or partial Sonnet predictions on OlympiadBench-Physics was completed post-evaluation. Methodology. The 100 cases were drawn (random seed 42) from the 354 total wrong predictions on OlympiadBench-Physics; each case was categorized by a single physics-trained annotator (graduate-level physics background, $\sim$3 hours of total annotation time, $\sim$1.8 min/case median) using a nine-category mutually-exclusive rubric (`wrong_subpart`, `missing_physics`, `calc_error`, `different_question`, `valid_partial`, `symbolic_vs_numeric`, `magnitude_error`, `sign_error`, `diagram_misread`). The annotator saw the problem statement, the gold solution, the gold final answer, and the Sonnet response; they did not see Sonnet's confidence or the verdict from the LLM judge. We report this single-annotator taxonomy as a methodological diagnostic, *not* as a calibrated human grade; a second-annotator pass on a 50-problem random subsample with inter-annotator κ per category is left to follow-up work (distinct from the audit-flag inter-annotator κ pre-registered in Appendix A, which targets contamination-flag agreement, not failure-category agreement).

Table 6: Failure-mode taxonomy of 100 randomly-sampled Sonnet wrong/partial predictions on OlympiadBench-Physics. Categories are mutually exclusive; each prediction received one label.

Category	n	%	Description
wrong_subpart	30	30%	Answered a neighboring sub-question instead of the one asked
missing_physics	22	22%	Wrong physical model, law, or geometry; crucial effect absent
calc_error	16	16%	Correct approach; small numerical slip (factor 2–4 off)
different_question	10	10%	Reasoning chain for an unrelated problem in the same prompt
valid_partial	8	8%	Right approach with arithmetic slip; partial-credit territory
symbolic_vs_numeric	7	7%	Formula returned where a number was required, or vice versa
magnitude_error	5	5%	Correct expression, wrong order of magnitude ($\geq 10\times$)
sign_error	0	0%	None observed
diagram_misread	0	0%	None observed (text-only evaluation; figures unavailable)
Total	100	100%	

Reading the taxonomy. `wrong_subpart` dominates (30%): multi-part olympiad prompts containing (a)/(b)/(c) sub-questions where the model fluently answers a neighboring sub-question instead of the asked one—a 5$\times$ underestimate by regex-only flagging ($\sim$6% in §5). `missing_physics` (22%) and `different_question` (10%) together account for nearly a third of failures and represent a real reasoning floor unlikely to be format-induced. `calc_error` (16%) and `valid_partial` (8%) are correct-approach/failed-execution, matching the 4.7-pp strict-vs-liberal gap (§5.1). `sign_error` and `diagram_misread` are zero (the latter because OlympiadBench-Physics is text-only in the public release).

Table 7: Per-physics-category accuracy on OlympiadBench-Physics (Sonnet 4.5, strict). The $+34.5$-pp EM (38.4%) vs. astrophysics (72.9%) gap on identical weights motivates per-category reporting.

Category	n	Acc.	Category	n	Acc.	Category	n	Acc.
Electromagnetism	237	38.4%	Relativity	89	51.7%	Other	10	60.0%
Quantum	128	41.4%	Classical mech.	19	57.9%	Astrophysics	70	72.9%
Waves	39	46.2%	Thermodynamics	87	59.8%	All	692	—
Fluid mechanics	13	46.2%						

Per-physics-category breakdown.

H.2 Ethical considerations and intended use

Intended and out-of-scope use. The released artifacts support research on visual physics reasoning, contamination-audit methodology, cross-lingual LLM evaluation, and RL post-training of VLMs. PHYSOLYM-A is held-out: please treat it as test-only, not for pretraining or fine-tuning. Out of scope: general physics ability beyond visual reasoning, human olympiad grading substitution, experimental-physics evaluation, paper-writing, derivation novelty, or open-ended hypothesis generation.

Source provenance, consent, and misuse risk. Kevin Zhou confirmed CC BY-NC 4.0 redistribution of his olympiad handouts in writing on 2026-05-03 ($\sim$1,600 problem-solution items, Appendix E); the Estonian Physics Olympiad collection and the six international olympiad scrapes (IPhO, NBPhO, EuPhO, APhO, USAPhO, INPhO) are redistributed under public-domain competition policy with per-record source attribution. No PII; problems are olympiad/textbook content. Public-domain olympiad scrapes are released by competition policy. The dense reward and Physics-R1 recipe ship for reproducibility, not as a tuned gold standard; the audit pipeline is a measurement tool, not a certification authority, and we disclose threshold sensitivity (Table 4) so users do not over-claim "contamination-free."

Caveats and compute. The 17-pp EN/ET cross-lingual delta is Sonnet-4.5-specific on n=59 paired items; direction may flip on low-resource-language-weak open-source models—treat as model-specific. Total compute $\sim$360 H200-GPU-hours across the 3 seeds in Table 3 (seed-42 at 30 h $\times$4 H200; seed-17 and seed-23 retrainings $\sim$ 60 h $\times$4 H200 combined; Appendix F), plus frontier-API inference for baselines and Sonnet-judge runs. Per-experiment carbon estimates are omitted because cloud-provider electricity mix is not consistently disclosed.

H.3 Construction-process disclosures

(i) The initial PHYSOLYM-A description claimed 100% novel-source; the Stage-1 audit surfaced one EuPhO 2020 overlap with OlympiadBench-Physics (J=0.91), so the honest claim is 99.8% (499/500)—disclosed, not dropped. (ii) Our initial header described `PhyX-mini-MC` as 500 problems; the canonical MC subset [Shen et al., 2025] is 1,000.

H.4 Future work and named follow-ups

Ten follow-ups consolidated. *Eval refinements:* (i) post-hoc MCQ-ification of PHYSOLYM-A to isolate the format axis (§5.1); (ii) paraphrase- and translation-aware audit pass. *Cross-corpus:* (iii) audit OlympiadBench-Physics against our pool, then add the Physics-R1 row; (iv) frontier cross-evaluation against PHYBench, PhysUniBench, HLE-physics, PHYSOLYM-A. *Recipe and scale:* (v) transfer to InternVL3-8B and LLaVA-OneVision-7B; (vi) 32B + full-RL comparator; (vii) SFT-only data-scaling curve at 500/1,293/5,000/9,575 audited prompts. *Cross-lingual:* (viii) confirm/refute EN/ET sign-flip on low-resource open-source models on the same 59-pair Estonian Physics Olympiad subset; *pre-registered*—if 8B-class open-source models with documented Estonian-weak training (e.g. Qwen2.5-VL-7B, LLaVA-OV-7B) show ET<EN by $\geq$ 5 pp under paired sign test, F2's Sonnet-4.5 ET>EN direction is model-specific (confirmed); a $\geq$ 5 pp result in the same direction (ET>EN) on those same models would replicate F2 across model families; results within $\pm$5 pp are inconclusive at n=59. *Methodology:* (ix) 50-problem inter-annotator κ on the failure-mode taxonomy; (x) embedder-sensitivity ablation against `voyage-3` and `text-embedding-3-large`.

H.5 Versioning and maintenance commitment

Versioned releases with semantic-version tags (`v1.0.0` initial, `v1.1.x` additive, `v2.0.0` schema-breaking). Each release ships Croissant 1.0 + RAI JSON-LD, per-source license matrix (Table 5), threshold-sensitivity grid (Table 4) recomputed against newly-released physics-VL benchmarks, and the audit-pipeline source archive with saved best-overlap scores. Quarterly contamination audit against new benchmarks; benchmarks introducing $\geq$ 1% leakage to any held-out split trigger a documented-diff removal in the next minor release. Hosting: HuggingFace ($\geq$ 5 yr), GitHub, Zenodo DOI per release.

Table 8: **Physics-R1 vs. rule-based RL recipes for thinking-mode VLMs.** Init: base or SFT cold-start. Reward signals: *Ans* (answer), *Fmt* (format), *Dim* (units), *Sym* (symbolic), *Cons* (conservation). Audit and Filter as defined in §3.3. ✓ present; — absent; ○ partial.

Recipe	Init	Ans	Fmt	Dim	Sym	Cons	Audit	Filter
DPO [Rafailov et al., 2023]	SFT	—	—	—	—	—	—	—
GRPO [Shao et al., 2024]	SFT	✓	—	—	—	—	—	—
GSPO+DAPO [Zheng et al., 2025, Yu et al., 2025]	SFT	✓	○	—	—	—	—	—
MM-Eureka [Meng et al., 2025]	base	✓	✓	—	—	—	—	—
Physics-R1	base	✓	✓	✓	✓	✓	✓	✓

Table 9: **PHYSCORP-PRE-AUDIT composition by source family (14,294 records total before two-stage audit).** The audited release PHYSCORP-A (6,432 records) is the subset surviving Algorithm 1 after a re-audit pass against PhysReason-full and PhysUniBench-en dropped an additional 804 records from a 7,236-record candidate (the construction audit excluded those two corpora). The 804-record re-audit drop concentrates on PhysReason-cousin records (540) and PhysUniBench-en-cousin records (186), drawn predominantly from the *repackaged-benchmark* source families (UGPhysics / OpenStax / Physics SE / MMMU+o1-CoT seed / PhysReason); the four *first-ML-format* source families (Estonian PhO, Zhou's handouts, the 7-international-olympiad scrapes) are minimally affected and retain their full 1,609-record contribution to PHYSCORP-A. All 9 source families remain represented in the released pool. Per-source licenses are carried through to each released record; full provenance and outreach log in Appendix E.

Source family	Count	License	Answer type
UGPhysics	5,520	CC BY-NC-SA	Open-ended numeric
OpenStax Physics	2,381	CC BY 4.0	Numeric
Physics Stack Exchange	2,291	CC BY-SA 4.0	Equations
RL-SFT seed (MMMU + o1 CoT)	1,293	MIT (MMMU); generated CoT	Varied
PhysReason	1,200	CC BY 4.0	CoT + open-ended
Zhou Olympiad Handouts	692	CC BY-NC 4.0†	Mixed
Estonian Physics Olympiad	418	Public-domain (competition policy)	Open-ended
IPhO + NBPhO + EuPhO scrape	258	Public domain	Open-ended
APhO + USAPhO + INPhO scrape	241	Public domain	Open-ended
Total novel	1,609	—	—
Total corpus	14,294	—	—

† Author granted explicit redistribution permission with attribution (email confirmation, May 2026); we redistribute under CC BY-NC 4.0.

H.6 Method comparison and benchmark comparison tables (extended)

H.7 Corpus composition, hyperparameters, and reward ablation (extended)

Table 10: **Physics-R1 training configuration.** GSPO+DAPO via `verl` 0.6.1 with vLLM 0.11.0 (TP=4) for rollouts. FSDP1 is required for Qwen3-VL under `verl` 0.6.1 (FSDP2 fails on the multimodal projector path; reproducibility note in §6). Cells marked *(recipe)* differ from a default unconstrained GSPO+DAPO configuration; together with the dense physics-native reward (Section 4), the audited training pool, and held-out early stopping, they constitute the Physics-R1 reference recipe.

Parameter	Value	Role in the recipe
Base / init *(recipe)*	Qwen3-VL-8B-Thinking BASE (*no SFT*)	Cold-start RL, mirroring MM-Eureka thesis
Algorithm	GSPO + DAPO	Stable group-policy backbone
Importance sampling *(recipe)*	truncated, sequence-level	Rollout-vs-train policy correction
Learning rate	1×10^{-6}	Standard for 8B-class RL
LR decay *(recipe)*	step-cosine; $0.5\times$ after step 60	Counters drift and length collapse
Batch size	96	Effective batch via gradient accumulation
Rollouts per prompt	16	DAPO group size
Max response length	12,288 tokens	Long-CoT thinking budget
KL anchor *(recipe)*	1×10^{-3} to base	Anchors policy to base; bounds drift
Entropy bonus *(recipe)*	1×10^{-3}	Prevents entropy collapse
Clip range (decoupled)	0.2 / 0.28	DAPO decoupled clip
Difficulty curriculum *(recipe)*	drop 0/16 and 16/16 rollout prompts	MM-Eureka-style; preserves learning signal
Reward *(recipe)*	binary correctness (§4)	Recommended default; dense (Algo. 2) is the shape ablation
Train pool *(recipe)*	2,268 PHYSR1CORP prompts	Closed-form carve-out of PHYSCORP-A (§3)
Early stop *(recipe)*	held-out PhyX-mini-MC	Catches the saturation peak
Sharding *(recipe)*	FSDP1 (not FSDP2)	Required for Qwen3-VL under `verl` 0.6.1
Hardware	4×H200 (single node)	—
Step time	$\sim$40 minutes/step	Rollout-bound (gen $\sim$65%, update $\sim$23%)
Wall-clock to saturation peak	$\sim$30–45 hours (steps 60–80)	Training cost $\sim$\$120–\$180
`transformers`	4.57.0 (pinned)	4.57.6 yields different forward path
`vllm`	0.11.0	TP=4 rollout backend
`verl`	0.6.1	Training framework
Random seed (headline)	42	3-seed sweep $\{17, 23, 42\}$ reported as the headline binary row in Table 3

Table 11: **Physics-R1 reward-shape ablation on PhyX-mini-MC.** Each row toggles one of the five reward components on or off, training Qwen3-VL-8B-Thinking-base under the same GSPO+DAPO+KL-anchor recipe (Equation 1) for the same step budget. Ans = answer-correctness binary ($+1$, $\equiv$ r_{bin} of §4); Fmt = `\boxed{}` format ($+0.1$); Dim = dimensional consistency from regex-detected units + sympy unit-system ($+0.15$); Sym = symbolic equation verification of intermediate `\frac` expressions via sympy ($+0.20$); Cons = conservation-law penalty (energy/momentum) when applicable (-0.25). Composed reward is clipped to $[-1, 1]$. Init in every cell is the Qwen3-VL-8B-Thinking BASE checkpoint (*no SFT*; cold-start, mirroring the MM-Eureka thesis). The Ans-only row is the recommended Physics-R1 recipe (binary correctness reward, §4); the all-on row is the dense ablation of §4; the intermediate rows isolate the marginal effect of each physics-native shaping component. Drop-out cells (—) are intermediate-component runs left to follow-up work (compute estimate in Appendix F).

Configuration	Ans	Fmt	Dim	Sym	Cons	PhyX-mini-MC
Base (no RL)	—	—	—	—	—	73.7%
Physics-R1 (binary, recommended) $\equiv$ Ans-only	✓	—	—	—	—	78.0
+ Format	✓	✓	—	—	—	—
+ Dim	✓	✓	✓	—	—	—
+ Sym	✓	✓	✓	✓	—	—
Physics-R1 (dense, ablation, all-on)	✓	✓	✓	✓	✓	78.3
Single-component drop-outs (dense ablation minus one)						
− Dim	✓	✓	—	✓	✓	—
− Sym	✓	✓	✓	—	✓	—
− Cons	✓	✓	✓	✓	—	—

Table 12: Sonnet 4.5 strict accuracy on Estonian Physics Olympiad problems by organizer-issued native difficulty (n=131). The curve is near-monotonically decreasing in difficulty and hits a hard floor of 0% at difficulties 3, 6, 8, and 10. Non-monotone bumps at 4 and 5 are within sampling noise on small per-bin counts.

Difficulty	n	Correct	Acc.
1	17	10	62.5%
2	15	3	20.0%
3	8	0	0.0%
4	12	3	25.0%
5	27	10	37.0%
6	9	0	0.0%
7	14	3	21.4%
8	9	0	0.0%
9	15	3	21.4%
10	5	0	0.0%
All	131	32	24.4%

Table 13: Cross-lingual Sonnet performance on the Estonian Physics Olympiad bilingual subset (n=59, identical problems, same judge protocol). Strict% = numeric/symbolic match; Liberal% = LLM-judge score ≥ 0.5.

Language	n	Correct	Partial	Incorrect	Unjudgeable	Strict%	Liberal%
English (translated)	59	8	8	43	0	13.6%	20.3%
Estonian (original)	59	18	9	27	5	30.5%	38.1%

Table 14: Per-problem agreement matrix for the EN/ET cross-lingual ablation (n=59). The 4.3:1 asymmetry in the off-diagonal cells (ET-correct/EN-wrong vs. EN-correct/ET-wrong) rules out a noise explanation.

	ET correct	ET wrong
EN correct	5 (both correct)	3 (EN only)
EN wrong	13 (ET only)	38 (both wrong)