Conservation-Aware Fine-Tuning: Injecting Physical Laws into LLM Reasoning via Preference Optimization

Anonymous authors
Paper under double-blind review

Abstract

Large language models frequently produce physically implausible reasoning—violating conservation laws of energy or momentum—even when their final numerical answers are correct. We introduce Conservation-Aware Fine-Tuning (CAFT), which treats conservation-law satisfaction as a programmatic preference signal for Direct Preference Optimization (DPO): a rule-based verifier labels each sampled reasoning chain as conservation-respecting or violating, and the resulting preference pairs fine-tune the model to prefer physically valid reasoning. The effect is scale-dependent rather than uniform. At 1.5B, CAFT reduces the Conservation Violation Rate (CVR) by 77.8% (5.62%→1.25%, p=0.004; scenario-clustered p=0.013) and raises accuracy by 16.9 pp (40.31%→57.19%, $p<10^{-4}$) on 80 mechanics scenarios; the CVR reduction replicates directionally across three training seeds (1.25%/2.19%/3.75%), while the accuracy magnitude does not. At 8B no arm separates from the base model under our strip-first evaluation protocol, which removes hidden thinking content before verification (all arms at 0/320 violations), so we report the 8B scale as a boundary rather than a transfer. At 32B the base model already sits at the violation floor (0.31%) and CAFT changes nothing measurable (p=1.0)—a ceiling effect with scale and model family co-varying. These three points distill into a deployment heuristic: apply CAFT only when the base model's CVR confidence interval excludes zero—a rule that is itself only as trustworthy as the protocol measuring CVR. Ablations attribute the conservation benefit to synthetic injection pairs, the accuracy benefit to wrong-answer pairs, and show the benefit is family-local; and a second, independently validated constraint probe shows no measurable side effect.

1 Introduction

Large Language Models have demonstrated impressive capabilities in mathematical and scientific reasoning DeepSeek-AI (2025a); Shao et al. (2024); Wei et al. (2022); Brown et al. (2020), yet they remain prone to physically implausible reasoning chains. A model may arrive at a correct numerical answer while claiming that kinetic energy is conserved in an inelastic collision, that momentum increases without an external force, or that mechanical energy is conserved in the presence of friction.

Existing approaches rely on outcome-based reward signals (correct/incorrect answers) Cobbe et al. (2021); DeepSeek-AI (2025a); Hendrycks et al. (2021b) or on learned process reward models (PRMs) Lightman et al. (2023); Wang et al. (2024a) that require expensive human annotation. Physics offers a distinctive alternative: conservation laws are exact, deterministic, and programmable, and verification yields binary judgments without any learned model or human label. The laws are unambiguous; detecting their verbalized violation in free-form text is not, and our own blind audit measures that gap (precision 67.9–69.6% against auditor consensus, bounded by a disclosed hedge-pattern false-positive family, Appendix H), so we treat the verifier as a high-precision proxy for a narrower construct rather than as ground truth.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We propose Conservation-Aware Fine-Tuning (CAFT), which uses conservation-law satisfaction as a verifiable reward signal for preference optimization: (i) generate physics scenarios with computable conservation laws (energy, momentum; mass armed but inert—Limitations (8)); (ii) sample reasoning chains from a base LLM; (iii) verify each chain with a programmable verifier; (iv) construct DPO preference pairs (chosen = conservation-respecting, rejected = violating); and (v) fine-tune with DPO. The four pair-construction strategies are natural (the model's own distribution), synthetic (injected violations), same-answer-process (identical final answers, different conservation status), and wrong-answer (an answer-accuracy signal).

Our contributions are:

- We introduce CAFT, a method that uses physical conservation laws as verifiable preference signals for LLM fine-tuning, and we explore the feasibility of conservation constraints as a preference signal for DPO-style optimization on scientific reasoning tasks.
- We design a conservation-law verifier that detects 9 violation patterns (mapping to 4 categories) in free-form reasoning chains, and we characterize its scope and validation (Appendices G and H).
- We propose four DPO pair-construction strategies, including a same-answer-process strategy (Strategy C) that pairs chains with identical final answers but different conservation-law status, isolating process supervision from answer correctness without human annotation.
- We report a scale-dependent empirical picture rather than a single headline number. At 1.5B, CAFT cuts CVR by 77.8% (p=0.004; clustered p=0.013) while improving accuracy, with the violation reduction replicating directionally across seeds. Scale-up then reveals two boundaries: at 8B no arm separates from the base model under our strip-first evaluation protocol (all arms at 0/320 violations), so we report a scale boundary rather than a transfer (Section 4.5); and at 32B a ceiling effect leaves no violation mass (base CVR 0.31%, CAFT 0.62%, p=1.0). We derive a CVR-interval deployment guideline from these points, and note that it is only as trustworthy as the protocol measuring CVR.
- We characterize the measurement validity of verifiable-constraint probing on a second checkable constraint family: a guarded arithmetic-consistency checker flags at most 0.62% of chains with no significant group difference, and a validated null-control probe on dimensional homogeneity shows no effect, so CAFT's footprint stays local to the conservation signal it was trained on.

2 Related Work

RLVR, PRMs, and preference optimization. Rule-based verifiable rewards have driven progress in math and code reasoning DeepSeek-AI (2025a); Shao et al. (2024); DeepSeek-AI (2025b). RLHF/PPO Ouyang et al. (2022); Christiano et al. (2017); Schulman et al. (2017) and reward-free variants—DPO Rafailov et al. (2023), KTO Ethayarajh et al. (2024), IPO Azar et al. (2024), SimPO Meng et al. (2024), ORPO Hong et al. (2024)—established preference optimization, while process reward models (PRMs) Lightman et al. (2023); Wang et al. (2024a); Khalifa et al. (2025) and self-refinement/self-rewarding methods Madaan et al. (2023); Saunders et al. (2022); Singh et al. (2024); Yuan et al. (2024b;a) provide step-level or self-generated signals. These rely on human preferences, learned reward models, or the model's own judgment (cf. Constitutional AI Bai et al. (2022)); our verifier is instead a programmatic PRM yielding exact, reproducible step-level judgments from deterministic physical laws, with no annotation and no learned critic.

Physics-informed ML and reasoning. PINNs Raissi et al. (2019); Karniadakis et al. (2021) embed physical laws as soft PDE-residual constraints in continuous models; CAFT brings this paradigm to LLMs by shaping the reasoning distribution through preference optimization. Concurrent physics-preference methods—PKG-DPO Kulkarni et al. (2025), thermodynamic rewards for alloy design Ghosh et al. (2025), physics RL Anand et al. (2024), and BeamPERL Hage & Buehler (2026)—use knowledge graphs, simulations, or domain-specific solvers rather than analytical conservation-law verification, and, to our knowledge,

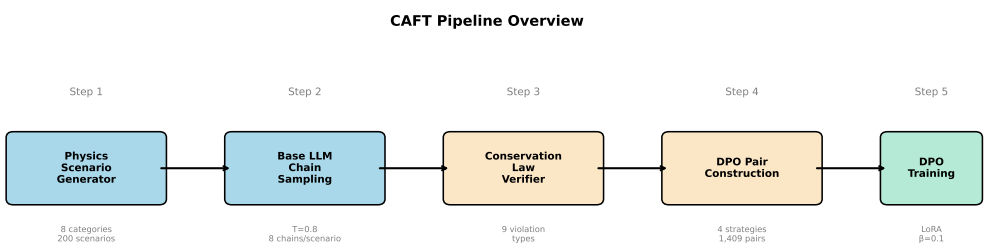

Figure 1: Overview of the CAFT pipeline. Step 1: Generate physics scenarios with computable conservation laws. Step 2: Sample reasoning chains from the base LLM. Step 3: Verify each chain against conservation laws. Step 4: Construct DPO preference pairs using four strategies. Step 5: Fine-tune with DPO and LoRA.

none isolate process signal from answer correctness via same-answer-process pairs. Science-reasoning benchmarks Wang et al. (2024b); Rein et al. (2024); Qiu et al. (2025); Xu et al. (2025); Hendrycks et al. (2021a); Lu et al. (2022) and inference-time methods Wei et al. (2022); Kojima et al. (2022); Wang et al. (2023); Yao et al. (2023); Snell et al. (2024) target answer correctness or test-time scaling rather than whether the reasoning itself respects physical laws—the gap CAFT fills.

3 Method

3.1 Overview

CAFT consists of three components: (1) a physics scenario generator with computable conservation laws (§3.2), (2) a conservation-law verifier that automatically checks LLM reasoning chains (§3.3), and (3) a DPO training pipeline with multiple pair-construction strategies (§3.4). Figure 1 illustrates the full pipeline.

3.2 Physics Scenario Generation

We generate closed-form mechanics problems across 8 categories (elastic/inelastic collisions, projectile, pendulum, and spring energy, friction dissipation, explosion/separation, and roller-coaster energy). Each scenario specifies a natural-language problem, initial/final physical states, applicable conservation laws (energy, momentum, mass), and a computable correct answer with units—ensuring conservation satisfaction is verifiable deterministically. Appendix B lists one example per category.

3.3 Conservation Law Verifier

The verifier checks reasoning chains for 9 violation patterns mapping to 4 categories—energy violations, momentum violations, incorrect formulas, and wrong law application (e.g., energy/momentum creation or loss, wrong KE/PE formulas, inelastic-as-elastic, friction-energy-conserved); the full pattern list is in Appendix J. It is a lexical detector: it matches surface formulations of known physics misconceptions using negation windows (so "energy is not conserved" is distinguished from "energy is conserved") and does not recompute balances from extracted quantities. A "violation" therefore means that a chain asserts a statement matching a canonical misconception template—not that a quantitative imbalance was derived. The construct measured is thus "verbalized conservation-law adherence," a strict subset of physical validity: chains that violate conservation without verbalizing a templated misconception are invisible to it, and the 32B boundary case in Section 4.5 (a hedged-but-correct statement falsely flagged) illustrates the failure mode. We chose lexical detection over a balance-recomputing verifier because it is fully reproducible and its errors are auditable.

The verifier is validated on 3,215 unique chains in six suites plus a 403-chain wrong-answer screen (5,030 checks total): across the four self-labeled suites it reaches 100% precision on its own construct (0 false positives in 2,911 clean-chain checks) and 100% recall on conservation violations (1,516 of 1,516), while a blind four-auditor physics audit against a broader, independently defined defect construct gives precision 67.9–69.6% and recall 38.8–40.0% (the two framings measure different constructs; Appendices G and H). Answer correctness uses a numerical tolerance of $\max(|a_{\text{correct}}| \times 5\%, 0.5)$ (rounding).

3.4 DPO Preference Pair Construction

We construct preference pairs using four strategies. Strategy A (Natural): from temperature-sampled chains, pair conservation-respecting chains (chosen) with violating chains (rejected) from the same scenario. Strategy B (Synthetic): take a correct, conservation-respecting chain and inject a violation with the ViolationInjector to create the rejected chain. Strategy C (Same-Answer-Process): both chains share the same final answer, but the rejected chain has an injected conservation violation in its reasoning—this isolates the process reward signal from answer accuracy. Strategy D (Wrong-Answer): pair correct-answer chains (chosen) with wrong-answer chains (rejected), providing an answer-accuracy signal.

3.5 Training

We use DPO Rafailov et al. (2023) with LoRA Hu et al. (2022) for efficient fine-tuning:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{1}$$

where y_w/y_l are the chosen/rejected chains, π_{ref} the frozen base LLM, and β controls deviation from the reference; LoRA adapters (rank r=16, scaling α=32) are applied to all linear layers. Algorithm 1 (Appendix J) summarizes the procedure.

4 Experiments

4.1 Setup

Models and baselines. We use Qwen2.5-1.5B-Instruct Team (2025a) for all main experiments and ablations, and Qwen3-8B Team (2025b) and Qwen3-32B (4-bit QLoRA) for scale-up (§4.5). We compare CAFT against (1) the base model and (2) a Supervised Fine-Tuning (SFT) baseline trained only on conservation-respecting, answer-correct chains, using the same LoRA configuration and epochs (3); because the SFT-eligible set is smaller, SFT takes fewer optimizer steps at equal epochs, and we add a step-matched control (Section 4.6).

Training and evaluation. Each scenario is a chat-formatted prompt (physics-expert system message + problem) eliciting a chain-of-thought ending in a boxed answer. We sample 200 scenarios × 8 chains per scale, yielding 1,409 / 1,379 / 1,282 DPO pairs at 1.5B / 8B / 32B, and train with β=0.1, learning rate 5×10^{-6}, 3 epochs, LoRA (rank 16, α=32; Appendices F and J). We evaluate on 80 in-distribution and 80 OOD scenarios with greedy and sampled decoding (T=0.8, 4 chains/scenario = 320 per arm).

Metrics. CVR: fraction of chains with ≥1 violation (lower is better); Accuracy: fraction with a correct final answer; CAA: fraction both correct and conservation-respecting; PVS: fraction of reasoning steps passing all checks (Appendix A). All comparisons are paired exact McNemar tests on aligned (scenario, chain-index) outcomes; proportions carry Wilson 95% CIs Wilson (1927), and headline tests are supplemented by a scenario-clustered sign-flip randomization test (Appendix C).

4.2 Main Results

Table 1 presents the main in-distribution results (Figure 7, Appendix R). Under temperature-sampled decoding (T=0.8), CAFT achieves the strongest performance across

Table 1: Main results: Base vs. SFT vs. CAFT on 80 in-distribution scenarios. Sampled: $T = 0.8$, 4 chains/scenario (320 total). Greedy: 1 chain/scenario (80 total).

Method	Decoding	CVR (%) ↓	Acc (%) ↑	CAA (%) ↑	PVS ↑
Base	Sampled	5.62	40.31	38.12	0.9925
	Greedy	11.25	58.75	47.50	0.9966
SFT	Sampled	5.62	34.06	31.56	0.9895
	Greedy	0.00	58.75	58.75	1.0000
CAFT	Sampled	1.25	57.19	56.56	0.9992
	Greedy	12.50	51.25	47.50	0.9950

all metrics, reducing CVR from 5.62% to 1.25% (a 77.8% relative reduction) while improving accuracy from 40.31% to 57.19% (+16.88 pp in this headline seed; seed-averaged +8.0 pp, range +3.1 to +16.9 pp across seeds); CAA, which requires both correct answers and valid reasoning, improves by 18.44 pp. All headline metrics improve significantly (CVR: 18 discordant pairs where base violates and CAFT does not vs. 4 in the opposite direction, p=0.004, clustered p=0.013; accuracy: 90 vs. 36, $p<10^{-4}$), and the Wilson intervals do not overlap (CVR [0.49, 3.17]% vs. [3.59, 8.72]%; accuracy [51.71, 62.49]% vs. [35.08, 45.77]%). Gains are broad-based: the largest CVR reductions are friction dissipation (22.5%→2.5%) and inelastic collision (17.5%→2.5%). The violation reduction replicates directionally across training seeds 43/44 (per-seed p=0.035/0.33) while the accuracy gain's magnitude does not (Section 4.6), so the co-benefit is seed-42-specific in magnitude—non-negative across seeds, but not fixed in size.

The SFT baseline fails to improve reasoning quality. SFT leaves CVR unchanged (5.62% vs. 5.62%), decreases accuracy (34.06% vs. 40.31%), and degrades accuracy sharply at 8B under the strip-first protocol (39.7% vs. 77.8%; §4.5): fine-tuning on a narrow set of correct chains can overfit surface patterns. CAFT's contrastive preference signal instead shifts the output distribution toward conservation-respecting chains; because the verifier targets verbalized misconceptions, we characterize the change as suppression of verbalized violation patterns (Section 4.6) rather than deeper physics understanding.

Decoding regime. Under greedy decoding SFT achieves perfect conservation (CVR=0%) at the cost of reduced diversity and no accuracy gain, and CAFT's greedy CVR (12.50%) is nominally higher than base (11.25%). All 10 CAFT greedy violations (and all 9 base violations) occur in the friction dissipation category, where the model claims mechanical energy is conserved despite friction. CAFT's advantage is therefore specific to sampled decoding (T=0.8), the paper's primary operating point; practitioners requiring deterministic decoding should weigh the greedy column of Table 1 first.

4.3 Ablation Studies

In isolation, synthetic pairs (590) achieve the lowest single-strategy CVR (1.56%), confirming violation injection as effective augmentation; same-answer-process pairs (319) reduce CVR (2.81%) while improving accuracy (+7.19 pp); wrong-answer pairs (403) improve accuracy (+5.31 pp) but not CVR (5.00%); and natural pairs (97) are too sparse to help. Leave-one-strategy-out runs show the conservation benefit depends on the conservation-bearing strategies—removing synthetic or natural pairs eliminates the CVR reduction (4.06%, p=0.42; 3.75%, p=0.33 vs. base)—and two family-transfer ablations show the benefit is confined to violation families present in the preference pairs (held-out families show no reduction: p=1.0 and p=0.108) with the accuracy co-benefit concentrated on trained families as well, so CAFT's effect is family-local rather than a global "sound physical" register. Full ablation prose, tables, and all cells: Appendix K.

4.4 Out-of-Distribution Evaluation

We evaluate on 80 OOD scenarios (same 8 categories, unseen numerical values). Under sampled decoding, CAFT achieves 54.06% accuracy (vs. 37.81% for base, +16.25 pp, 98/46 discordant pairs, $p<10^{-4}$) with 0.31% CVR (vs. 1.25%); CAA improves from 37.81% to 54.06%. Under greedy decoding CAFT reaches perfect conservation (CVR=0%) with 55.0% accuracy. A caveat on OOD CVR: the accuracy gain is the robust OOD result; the CVR comparison is not. The base model's OOD CVR (1.25%) is already below its in-distribution CVR (5.62%), at 32B both sit at the floor, and the 8B OOD cells sit at the violation floor under the strip-first protocol (§4.5), so CAFT is never nominally better than base on OOD CVR at any interpretable scale (4-vs.-1 discordant pairs, p=0.375): the conservation benefit is an in-distribution phenomenon at 1.5B, and extending it to the out-of-distribution split is left to future work.

4.5 Scale-Up to 8B and 32B

We apply the same pipeline to Qwen3-8B Team (2025b) and Qwen3-32B (4-bit QLoRA). Table 22 summarizes the three scales under sampled decoding; the full table, the per-cell detail, and the boundary-case reasoning are in Appendix O.

8B: no measurable effect. At 8B our evaluation pipeline uses the strip-first protocol with the /no_think system directive, which removes hidden model thinking content before verification. Under this protocol all 8B arms sit at the violation floor (base 0/320; CAFT 0/320 in-distribution and 1/320 out of distribution; SFT 0/320), and there is no significant CAFT-vs.-base accuracy difference (+3.4 pp, 81.2% vs. 77.8%; 33 vs. 22 discordant pairs, p=0.177), so we claim no 8B accuracy benefit. SFT still degrades accuracy substantially (−38.1 pp), consistent with the 1.5B finding that fine-tuning on correct chains alone harms reasoning. We therefore report the 8B scale as a boundary rather than a transfer, and treat the system prompt as part of the evaluation protocol for any reasoning model able to emit hidden thinking content. Same-family control: Qwen3-1.7B. To disentangle scale from family we evaluated the base Qwen3-1.7B model (same family as 8B/32B, closest size to the 1.5B Qwen2.5): sampled CVR 0.00% (0/320 in-distribution; 0/320 OOD; greedy 0/80) with 38.44%/40.94% accuracy (ID/OOD), already at the violation floor despite barely exceeding the 1.5B Qwen2.5 (5.62% ID). The 8B and 32B models thus already sit at the violation floor, and family-level differences in training mixture plausibly dominate; a full same-family CAFT sweep remains future work (Limitations (3)).

32B: ceiling effect. At 32B the base model already has near-zero CVR (0.31% in-distribution, 0.00% OOD), so little violation mass remains and CAFT's marginal CVR change is unmeasurable (0.62% vs. 0.31%: one discordant event, p=1.0; all methods reach 0% greedy CVR). CAFT-32B shows a small accuracy gain in-distribution (86.56%→88.12%, +1.56 pp) and OOD (89.06%→90.62%), neither significant (p=0.49, p=0.47). We therefore report no measurable CVR change at 32B, neither benefit nor regression; the pair-scarcity counts and flag anatomy ruling out an independent data-starvation account are in Appendix O. Scale co-varies with model family throughout (Qwen2.5 at 1.5B vs. Qwen3 at 8B/32B), so the trend of increasing conservation awareness with scale is indicative only.

When to apply CAFT: a deployment guideline. The three-scale results suggest a decision rule keyed to the base model's measured CVR uncertainty, offered as a heuristic rather than a validated prescription: estimate the base model's CVR with n sampled chains and its Wilson 95% interval, and (i) if the interval excludes zero (here 1.5B, [3.59, 8.72]%), expect headroom for violation reduction—in the one case that survives protocol scrutiny CAFT yielded a large relative CVR reduction (77.8%); (ii) if the interval includes zero (here 32B [0.06, 1.75]% and Qwen3-1.7B, 0/320), the model is at or near the violation floor and any apparent change will be single-event noise. The rule screens for potential headroom, not realized benefit; its granularity and the interval arithmetic behind the protocol caveat are in Appendix O.

4.6 Robustness: Seed Variance and Controls

Seed variance (1.5B). We repeated the full CAFT-1.5B run (identical data, hyperparameters, 1,409 pairs) with seeds 43 and 44 under the standard protocol. The conservation benefit is directionally consistent across all three seeds: CVR falls from 5.62% (base) to 1.25%/2.19%/3.75% for seeds 42/43/44, with paired exact McNemar significant for seed 43 (17 vs. 6 discordant pairs, p=0.035) and directionally consistent for seed 44 (16 vs. 10, p=0.33); the pooled and seed-level analyses confirming both effects jointly, and the per-cell values, are in Appendix O and Table 19. The accuracy co-benefit is the less stable effect—per-seed gains are all non-negative but uneven, so the seed-42 magnitude is one draw from a wide run distribution, whereas the OOD accuracy co-benefit replicates more stably (Appendix O).

Outcome-only DPO at matched budget. Holding the optimization budget fixed while removing the process signal, we built an outcome-only pair set from the same chain pool (chosen = answer-correct, rejected = answer-incorrect, regardless of conservation status; 416 unique pairs upsampled to the matched 1,409-pair budget) and trained with identical hyperparameters, at matched optimizer steps rather than matched gradient diversity (Appendix O). The outcome-only model does not reduce conservation violations: CVR 4.06% [2.39, 6.83] vs. base 5.62% (17 vs. 12 discordant pairs, p=0.46), indistinguishable from base and far from CAFT's 1.25%, while accuracy rises only marginally (44.06% vs. 40.31%, p=0.32). Head-to-head at matched budget, CAFT outperforms the outcome-only baseline on both axes (paired tests in Appendix O): outcome-only training never sees its answer-correct but conservation-blind responses as violating, so the gradient carries no pressure toward conservation phrasing (Table 20).

Step-matched SFT control. The SFT-eligible set is smaller than the DPO pair set, so the original SFT runs took fewer optimizer steps at equal epochs (Appendix I), leaving undertraining as a possible explanation for SFT's underperformance. Retraining SFT-1.5B with the step count matched to DPO yields a model that behaves like the original SFT run rather than like CAFT, so the dose–response prediction of the under-training hypothesis fails: what separates SFT from CAFT is the contrastive exposure to violating chains, not the number of gradient steps. Step counts and all paired tests are in Appendix O.

Two zero-training alternatives. Test-time selection (best-of-4 with verifier ranking, BoV-4) drives CVR to 0% in every cell yet lowers accuracy relative to single-sample base at 1.5B, since selection optimizes conservation compliance rather than answer correctness; CAFT improves both simultaneously and bakes the improvement into the weights, so the two are complementary. A prompt-only control appending an explicit conservation-check directive to the system prompt does not reduce violations and degrades accuracy, and CAFT dominates it. Instruction alone therefore neither suppresses the misconception patterns nor preserves accuracy, indicating that the learned contrast, not task framing, drives the reduction (per-cell numbers and tests in Appendix O).

The 8B seed runs under the strip-first protocol. The 8B replicas (seeds 43/44) score 0/320 violations because they were evaluated under the strip-first prompt, under which no checkpoint emits thinking content (accuracy cells in Appendix O); their agreement therefore reflects the evaluation protocol rather than a training replication. Tests whose values fall in the 0.01–0.05 band remain provisional throughout the paper.

4.7 Analysis

CAFT's CVR reductions concentrate where violations are most common (friction dissipation 22.5%→2.5%, inelastic collision 17.5%→2.5%) while violation-free categories remain violation-free; by violation type, energy violations fall from 4.06% to 0.31%, wrong-law-application from 1.25% to 0.94%, and momentum violations are eliminated (Appendix J). CAFT also attains the highest Process Validity Score (0.9992 vs. 0.9925 base, 0.9895 SFT), and conservation-based pairs are easier to learn (perfect training accuracy) than wrong-answer pairs (64.3%). The chain-level transition decomposition and training curves are in Appendix Q.

4.8 A Second Verifiable Constraint: Arithmetic Consistency

To probe whether CAFT's effect is specific to its training constraint or reshapes reasoning more broadly, we score all evaluation chains against a second, orthogonal family: numerical arithmetic consistency (stated binary equalities verified against exact arithmetic). A guarded checker flags at most 0.62% of chains in every cell, with no significant system pair anywhere (all paired exact McNemar $p \geq 0.5$; Appendix E). Manual audit of all five residual flags finds three genuine slips, one sign-shorthand transcription, and one deliberate reductio the model itself labels impossible; the residual flags are disjoint from conservation violations (0 of 5,760 evaluation chains flagged by both checkers). A third, independently audited probe on dimensional homogeneity also shows no systematic difference (in-distribution 8.1%/15.3%/0.3% base vs. 5.9%/16.6%/0.6% CAFT at 1.5B/8B/32B; all paired $p \geq 0.13$), and training on dimensional-consistency pairs does not reduce conservation violations (p=0.69), so neither method's footprint transfers to the other family (Appendix D). The methodological lesson outlives the null result: untargeted-constraint probes should be validated to the same standard as the primary metric before their outputs are interpreted.

5 Discussion

Programmatic PRMs, DPO vs. SFT, and transfer. The verifier is a programmatic process reward model requiring no annotation and no learned critic Lightman et al. (2023); Wang et al. (2024a), yielding deterministic judgments over its target construct—binary by construction, but scoped: precision and recall against a broader blind-audited construct are 67.9–69.6% and 38.8–40.0% (Appendix H). Its coverage is mechanics only. DPO-based CAFT outperforms SFT at 1.5B and 8B, consistent with findings that fine-tuning on narrow distributions degrades diversity Ouyang et al. (2022): SFT trains only on correct chains, while DPO teaches the model to discriminate; at 32B the comparison is moot. The per-law ablation hints at cross-law transfer (energy-only training also cuts wrong-law-application violations, 1.25%→0.31%), but on a single run with few violations.

Limitations. The main threats to validity are: (1) run-to-run variance—the 1.5B CVR reduction replicates directionally across seeds (1.25%/2.19%/3.75% vs. base 5.62%; Fisher p=0.003) but per-seed significance is unstable (p=0.33 at seed 44) and the accuracy co-benefit spans +3 to +17 pp, so single-run magnitudes are wide draws; (2) construct and domain scope—the verifier is lexical, covers mechanics only, and its most independent validation is a blind LLM-referee audit with precision 67.9–69.6% and recall 38.8–40.0% against a broader defect construct; (3) scale/family coverage—the Qwen3-1.7B control (0/320) rules out a pure scale effect, and a same-family sweep with training is a natural next step; (4) benchmark narrowness—closed-form scenarios, with OOD varying only numerical values; (5) statistical resolution—320 chains per split leaves single-digit violation counts at 32B, violations cluster within scenarios (mitigated by clustered randomization tests), and ∼30 paired tests are reported without a global multiplicity correction; (6) no 8B claim in either direction—under the strip-first protocol 8B shows no measurable CVR or accuracy difference (0/320 vs. 0/320 in-distribution; 1/320 OOD; accuracy p=0.177); (7) ceiling at 32B (0.31% vs. 0.62%, p=1.0); (8) conservative probes—both untargeted checkers are lower bounds and the armed mass check never fires (a floor effect); (9) quantization—32B uses QLoRA 4-bit; (10) coverage of baselines and ablations—the outcome-only baseline is run at 1.5B, the scale with the clearest effect, and extending it to 8B/32B is a natural next step, while the leave-one-strategy-out arms rest on 4–16 discordant events, so their division-of-labor attribution is read as directional; and (11) auditor type—our referees are LLMs, not human experts. Appendix P gives each item in full, with all cells and secondary analyses.

Future work and broader impact. Promising directions include extending the verifier to other physics domains and chemistry, scaling to 70B+ models with harder scenarios that reopen headroom, combining CAFT with outcome-based RLVR, and testing transfer to non-physics STEM reasoning. All components are reproducible (deterministic seeds, rule-based verifier, standard hyperparameters), and the principle—domain laws as programmable, annotation-free training signals—has positive implications for safety-critical applications, provided the

verifier is re-validated on each new domain, since a false positive silently corrupts the preference signal.

6 Conclusion

We introduced Conservation-Aware Fine-Tuning (CAFT), which uses physical conservation laws as verifiable preference signals for DPO. The empirical picture is scale-dependent: CAFT reduces conservation violations by 77.8% at 1.5B (p=0.004; clustered p=0.013) while improving accuracy by 16.9 pp at the headline seed, with the reduction replicating directionally across seeds (1.25%/2.19%/3.75%; Fisher p=0.003) and the accuracy co-benefit generalizing out of distribution (+16.25/+10.00/+7.81 pp across seeds 42/43/44). At 8B no arm separates from the base model under the strip-first evaluation protocol, which removes hidden thinking content before verification (all arms 0/320; CAFT 1/320 out of distribution), with no significant accuracy difference (p=0.177), so we report a scale boundary rather than a transfer. At 32B a ceiling leaves nothing to fix (base CVR 0.31%, all methods 0% under greedy decoding, p=1.0). These three points distill into a deployment heuristic—apply CAFT when the base model's CVR interval excludes zero, skip it otherwise.

Two contributions are methodological. Process-isolating same-answer pairs show conservation-law satisfaction is a training signal distinct from answer correctness, and family-transfer ablations show the benefit is confined to violation families present in the preference pairs; the mechanism suppresses verbalized violation patterns rather than instilling deep physical understanding. Separately, guarded probes on a second verifiable constraint family (arithmetic consistency and dimensional homogeneity) show no measurable side effect, so CAFT's footprint stays local to the conservation signal it was trained on. Domain laws can thus serve as programmable, annotation-free process reward models, provided that both the training signal and the probes measuring its side effects are validated before their outputs are interpreted.

AI use statement

Large language models were used throughout this project, and we state this openly rather than in vague terms. Concretely, LLM-based assistants were used for (i) implementing and refactoring the experimental harness and the analysis scripts, (ii) assisting with literature search and the first-pass organisation of the related-work sections, (iii) language polishing and condensation of the manuscript, and (iv) suggesting candidate prompt phrasings and code templates that the authors then reviewed, modified and validated. The conservation-law verifier that produces the preference signal, and the boundary-case analysis of its friction-scenario pattern reported in the paper, were re-checked by the authors against the raw model generations before any of the numbers in the tables were accepted. The research question, the design of the method, the evaluation protocol and the interpretation of the results are the authors' own. Every experimental result reported here was produced by actually executing the code on the models described in the paper; no number, figure or reference was generated by a language model and accepted without author verification against the raw run logs or the original publication. The authors take full responsibility for the content of the paper and for any remaining errors.

Ethics statement

This study does not involve human subjects, human annotations or personal data. All evaluation items are synthetic mechanics word problems generated programmatically, and all models used are open-weight checkpoints released under permissive licences. The method optimises a model towards physically consistent reasoning, which we regard as a low-risk objective, but we note two caveats. First, the automatic conservation-law verifier is a heuristic: a false positive or false negative changes the preference signal, so any downstream use of CAFT should re-validate the verifier on the target domain rather than assume its judgements are ground truth. Second, our results are obtained on three model scales and 80 mechanics

scenarios; they should not be extrapolated to other scientific domains, to closed proprietary models, or to safety-critical deployments without independent validation. The fine-tuning procedure itself relies on preference optimisation, which can shift model behaviour beyond the targeted capability, and we therefore recommend that released checkpoints be evaluated on standard safety benchmarks before use. The authors declare no competing interests. The compute used for this work is small by contemporary standards (single-GPU QLoRA runs) and is reported in the appendix.

Reproducibility statement

We describe the data generation procedure, the conservation-law verifier, the preference-pair construction, and the full training configuration (LoRA rank, learning rate, batch size, number of epochs, QLoRA 4-bit settings) in the method section and the appendix, together with the decoding settings used for every reported figure. All reported violation rates and accuracies are computed from per-chain run logs with deterministic seeds (train: seed=42; evaluation: the recorded eval seed is 42, from which the in-distribution scenario generator derives its scenario set internally via seed+10,000 (the "10,042" of the released configuration), and OOD scenarios use the same eval seed 42 directly (the OOD generator applies no offset); chain sampling itself is stochastic and not bit-deterministic, Section 4.6), and the statistical comparisons use paired exact McNemar tests on aligned (scenario, chain) keys, supplemented by scenario-clustered bootstrap CIs and scenario-level sign-flip randomization tests (Appendix C), with Wilson 95% confidence intervals for all proportions. The arithmetic-consistency checker of Section 4.8 is a self-contained regular-expression scorer over the released chain texts. Code, verifier, scenario generators, and evaluation data will be released publicly upon acceptance; they are withheld at submission time only because the repository metadata would identify the authors under double-blind review.

References

Avinash Anand, Kritarth Prasad, Chhavi Kirtani, Ashwin R Nair, Mohit Gupta, Saloni Garg, Anurag Gautam, Snehal Buldeo, and Rajiv Ratn Shah. Enhancing llms for physics problem-solving using reinforcement learning with human-ai feedback. arXiv preprint arXiv:2412.06827, 2024.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. arXiv preprint arXiv:2310.12036, 2024.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, et al. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems, 2020.

Paul Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 2017.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

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

DeepSeek-AI. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, 2025b.

Kawin Ethayarajh, Ryan Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. arXiv preprint arXiv:2402.01306, 2024.

Satanu Ghosh, Collin S Holgate, Neal R Brodnik, Doug Downey, and Samantha Daly. Preference learning from physics-based feedback: Tuning language models to design bcc/b2 superalloys. arXiv preprint arXiv:2511.12036, 2025.

Tarjei Paule Hage and Markus J Buehler. Beamperl: Parameter-efficient rl with verifiable rewards specializes compact llms for structured beam mechanics reasoning. Machine Learning: Science and Technology, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. International Conference on Learning Representations, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. Advances in Neural Information Processing Systems, 2021b.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. arXiv preprint arXiv:2403.07691, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. International Conference on Learning Representations, 2022.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. Nature Reviews Physics, 3(6):422–440, 2021.

Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. Process reward models that think. arXiv preprint arXiv:2504.16828, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 2022.

Nitin Nagesh Kulkarni, Bryson Wilcox, Max Sawa, and Jason Thom. Pkg-dpo: Optimizing domain-specific ai systems with physics knowledge graphs and direct preference optimization. arXiv preprint arXiv:2508.18391, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. arXiv preprint arXiv:2305.20050, 2023.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems, 2022.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. arXiv preprint arXiv:2303.17651, 2023.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. arXiv preprint arXiv:2405.14734, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 2022.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, et al. Phybench: Holistic evaluation of physical perception and reasoning in large language models. arXiv preprint arXiv:2504.16074, 2025.

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. Advances in Neural Information Processing Systems, 2023.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378:686–707, 2019.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. arXiv preprint arXiv:2311.12022, 2024.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. arXiv preprint arXiv:2206.05802, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, et al. Beyond human data: Scaling self-training for problem-solving with language models. Transactions on Machine Learning Research, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314, 2024.

Qwen Team. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2025a.

Qwen Team. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025b.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. arXiv preprint arXiv:2312.08935, 2024a.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. ICML 2024, 2024b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. International Conference on Learning Representations, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 2022.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. ICML 2025, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 2023.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. Advancing llm reasoning generalists with preference trees. arXiv preprint arXiv:2404.02078, 2024a.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, and Sainbayar Sukhbaatar. Self-rewarding language models. arXiv preprint arXiv:2401.10020, 2024b.

A Conservation Law Verifier Details

The conservation law verifier (conservation_verifier.py) checks LLM reasoning chains against applicable physical conservation laws. It operates in three stages: (1) quantity extraction, (2) conservation law checking, and (3) formula correctness verification.

Quantity extraction. The verifier extracts numerical values with their physical context using regex patterns. It identifies velocities (e.g., "v = 5.0 m/s"), masses (e.g., "m = 2.0 kg"), heights, and energies from the reasoning text. Each extraction records the value, unit, and surrounding context for disambiguation.

Conservation law checking. Three checkers verify different conservation laws:

- check_energy_conservation: Detects energy creation (energy appearing without a source), energy loss in conservative systems (frictionless surfaces, no non-conservative forces), and incorrect energy conservation claims in inelastic collisions.
- check_momentum_conservation: Detects momentum creation (external impulse in a closed system) and momentum loss without external force.
- check_mass_conservation: Detects mass non-conservation in closed systems.

Formula correctness. The verifier checks for incorrect kinetic energy formulas (e.g., $KE = mv$ instead of $KE = \frac{1}{2}mv^2$), incorrect potential energy formulas, and wrong law application (e.g., treating inelastic collisions as energy-conserving or elastic collisions as energy-dissipating).

Process Validity Score (PVS). To compute PVS, the verifier splits each reasoning chain into individual steps using numbered/step delimiters (e.g., "Step 1:", "1.") and newline boundaries. Each step is then checked against all applicable conservation laws (energy, momentum, mass) and formula-correctness patterns. PVS is the fraction of steps that pass all checks, averaged over all chains; a chain with no violations has PVS = 1.0, and a chain where one of five steps contains a violation has PVS = 0.8.

Negation handling. To avoid false positives, the verifier checks for negation words ("no", "not", "never", "without", "doesn't", "don't", "isn't", "aren't", "won't", "cannot", "can't", "neither", "none") in the 20-character window preceding each pattern match. For example, this distinguishes "no energy is lost to friction" (a correct statement in a frictionless scenario) from "energy is lost to friction" (a conservation violation in a frictionless scenario). Negation-aware matching is essential in friction and inelastic-collision scenarios, where correct chains frequently hedge energy statements; all results in this paper use the negation-aware verifier. Sensitivity to the hedged-statement boundary: the residual known weakness is the unqualified hedged assertion "energy is conserved, but mechanical energy is dissipated," which the friction pattern flags despite being physically correct (Section 4.5 dissects the single 32B instance). We quantified the blast radius of this boundary case by scanning all 5,760 sampled evaluation chains for the hedged phrasing: exactly one chain

matches—the CAFT-32B in-distribution chain dissected above—so a clause-level semantic fix would change CAFT-32B sampled CVR from 0.62% to 0.31% and no other cell's numbers, softening an already-null comparison while leaving every headline result untouched.

Violation injection. The ViolationInjector class generates synthetic violations for DPO training. It maintains 9 violation template categories, each with multiple phrasing variants. The injector selects applicable violation types based on the scenario's physics category and applicable conservation laws, then inserts the violation at a random position in the reasoning chain.

B Example Physics Scenarios

Table 2 shows representative scenarios from each of the 8 physics categories. Each scenario includes a natural language problem description, initial and final states, applicable conservation laws, and a computable correct answer.

Table 2: Example scenarios from each physics category. All scenarios have closed-form solutions ensuring deterministic verification.

Category	Problem Description	Laws	Answer
Elastic collision	Two carts (1 kg, 2 kg) collide elastically on a frictionless track. Initial velocities: 3 m/s, −2 m/s. Find final velocities.	Energy, Momentum	v_1'=−3.67 m/s, v_2'=1.33 m/s
Inelastic collision	A 2 kg cart at 4 m/s collides perfectly inelastically with a stationary 3 kg cart. Find the final velocity.	Momentum	1.6 m/s
Projectile energy	A ball is launched at 20 m/s at 30°. Find the maximum height.	Energy	5.1 m
Pendulum energy	A 0.5 kg pendulum (length 1 m) swings from 45°. Find the maximum speed at the bottom.	Energy	2.4 m/s
Spring energy	A 0.3 kg mass on a spring (k=100 N/m) is compressed 0.1 m. Find the maximum velocity at equilibrium.	Energy	1.83 m/s
Friction dissipation	A 2 kg block slides at 5 m/s on a surface with μ=0.2. Find the stopping distance.	Energy (with friction)	6.38 m
Explosion	A 5 kg object at rest explodes into 2 kg and 3 kg fragments. The 2 kg fragment moves at 6 m/s. Find the 3 kg fragment velocity.	Momentum	−4.0 m/s
Roller coaster	A car starts from rest at 30 m height. Find its speed at 10 m height (frictionless track).	Energy	19.8 m/s

C Confidence Intervals for Key Results

Table 3 reports Wilson 95% confidence intervals for the key metrics under sampled decoding (T=0.8, n=320 chains). All CAFT vs. base comparisons use paired exact McNemar tests on aligned per-chain outcomes; the Sig. column reflects these paired tests.

Scenario-clustered sensitivity analysis. Because the 4 sampled chains within a scenario are correlated (shared problem, shared difficulty), we additionally ran a scenario-level cluster

bootstrap (10,000 resamples of the 80 scenarios with replacement; seed 42) on the CAFT-minus-base CVR and accuracy differences. The resulting 95% CIs for the in-distribution CVR delta are $[-7.81, -1.25]$ pp (1.5B) and $[0.00, 0.94]$ pp (32B); for the accuracy delta, $[-22.50, -11.25]$ pp (1.5B, negative = CAFT better) and $[-5.62, +2.81]$ pp (32B). The headline conclusions are unchanged under clustering: the 1.5B CVR and accuracy reductions exclude zero, and the 32B differences are compatible with zero.

Scenario-clustered significance tests. Cluster bootstrap CIs do not by themselves supply p-values, so we complement them with a cluster-level sign-flip randomization test: within each (scenario, chain) pair the two methods' outcomes are exchanged, and the sign of the entire scenario's paired difference is flipped with probability 1/2—treating the 80 scenarios, not the 320 chains, as the exchangeable units (20,000 permutations; seed 20260922; computed with the verifier of Appendix G). Applied symmetrically to every headline comparison, this test strengthens some claims and weakens others, and we report both directions: the 1.5B CVR reduction survives (clustered p=0.0114 in the archived 20,000-permutation run with seed 20260922; a different 20,000-permutation batch gives p=0.013—Monte-Carlo noise at this iteration count, both far from any threshold—vs. unclustered McNemar p=0.004; the archived value is authoritative), the 1.5B accuracy gain survives ($p<10^{-4}$), the CAFT-vs-outcome-only accuracy difference survives (p=0.0005), and the CAFT-vs-outcome-only CVR difference becomes marginal (clustered p=0.063 vs. p=0.049 unclustered), as does the leave-one-synthetic-out vs. full-CAFT CVR contrast (clustered p=0.060 vs. p=0.049 unclustered; same permutation scheme). At 8B the strip-first protocol yields 0 discordant pairs, so no clustered CVR entry is reported at that scale. Wherever a claim rests on a test in the 0.01–0.05 band, the clustered value is the one we treat as authoritative, and the main text marks the affected claims accordingly.

Table 3: Wilson 95% confidence intervals for key metrics under sampled decoding (n=320). Sig. refers to the paired exact McNemar test on aligned chains: ***$p<0.001$, **$p<0.01$, *$p<0.05$, ns=not significant. Under the strip-first protocol all 8B arms sit at 0/320 violations (base 77.8% vs. CAFT 81.2% accuracy, paired p=0.177), so no 8B interval separates the systems and the 8B rows are omitted here.

Metric	Base [95% CI]	CAFT [95% CI]	Sig.
1.5B In-distribution			
CVR (%)	5.62 [3.59, 8.72]	1.25 [0.49, 3.17]	**
Accuracy (%)	40.31 [35.08, 45.77]	57.19 [51.71, 62.49]	***
CAA (%)	38.12 [32.97, 43.56]	56.56 [51.09, 61.88]	***
1.5B OOD			
CVR (%)	1.25 [0.49, 3.17]	0.31 [0.06, 1.75]	ns
Accuracy (%)	37.81 [32.67, 43.24]	54.06 [48.59, 59.44]	***
CAA (%)	37.81 [32.67, 43.24]	54.06 [48.59, 59.44]	***
32B In-distribution			
CVR (%)	0.31 [0.06, 1.75]	0.62 [0.17, 2.25]	ns
Accuracy (%)	86.56 [82.39, 89.87]	88.12 [84.12, 91.22]	ns
CAA (%)	86.56 [82.39, 89.87]	87.81 [83.77, 90.96]	ns
32B OOD			
CVR (%)	0.00 [0.00, 1.19]	0.00 [0.00, 1.19]	ns
Accuracy (%)	89.06 [85.17, 92.03]	90.62 [86.93, 93.35]	ns
CAA (%)	89.06 [85.17, 92.03]	90.62 [86.93, 93.35]	ns

D Dimensional-Homogeneity Probe (Untargeted Constraint Control)

Section 4.8 probes untargeted verifiable constraints to test whether preference optimization shifts behaviors beyond its targeted conservation constraint. To complement the guarded arithmetic checker with a second, independently constructed probe, we scored all 1,920 sampled chains with an additional checker of dimensional homogeneity, a verifiable constraint orthogonal to both conservation laws and arithmetic. The checker parses stated numeric

equalities of the form "quantity = number unit", infers the intended physical dimension of the left-hand side from a symbol/keyword lexicon (handling subscripts, LaTeX markup, Unicode superscripts, and common shorthand; lexically ambiguous left-hand sides such as bare k or concatenated products like mgh are skipped conservatively), and flags the chain if the stated unit's dimension mismatches (e.g., a velocity equated to an acceleration). Checker behavior was audited on 41 hand-written symbol-inference cases and 19 end-to-end chain tests before scoring.

Table 4 reports the resulting rates. No comparison reaches significance in any cell (all paired exact McNemar $p \geq 0.13$), and there is no consistent direction across scales: CAFT is nominally lower than base at 1.5B ID (5.94% vs. 8.12%, p=0.31) and nominally higher at 8B ID (16.56% vs. 15.31%, p=0.75). We conclude that dimensional unit usage—an untargeted verifiable behavior sharing surface form with the stated equalities probed by the arithmetic checker—shows no measurable shift under CAFT, corroborating the main text's conclusion that no untargeted side effect survives checker validation.

Direct dimensional training (transfer-in check). As a stronger control than passive scoring, we also trained 1.5B models on dimensional-consistency pairs (DPO-dim and SFT-dim, constructed by the same pipeline as CAFT but with unit-mismatch—rather than conservation-violation—chains as rejected samples) and evaluated them on both splits of the conservation benchmark. The dimensional-trained models do not reduce conservation violations on either split: OOD 0.62% vs. base 1.25% (4 vs. 2 discordant pairs, p=0.69), and in-distribution 6.25% vs. base 5.62% (15 vs. 17 discordant pairs, p=0.86)—confirming specificity in the trained direction: optimizing one constraint family does not deliver another. Conversely, on the trained constraint itself, DPO-dim trends toward better dimensional hygiene than CAFT (OOD 5.00% vs. 9.06% inconsistent, p=0.053; ID 4.38% vs. 5.94%, p=0.44) and SFT-dim than SFT (OOD 6.56% vs. 10.94%, p=0.065; ID 7.50% vs. 9.38%, p=0.48)—directionally consistent with each method improving its own constraint, though no comparison survives correction for multiple comparisons, and both models pay an accuracy cost relative to CAFT (ID 35.9%/23.4% and OOD 35.0%/27.2% vs. CAFT's 57.2%/54.1%). One nuance deserves disclosure: on the ID split, DPO-dim's dimensional-inconsistency rate relative to base (4.38% vs. 8.12%, 23 vs. 11 discordant pairs, p=0.058) is nominally stronger than its OOD counterpart, but the dimensional checker can parse stated equalities in only 40 of DPO-dim's 320 ID chains (vs. 57 of base's), so the comparison rests on a small, model-dependent checkable subset, and the per-checked-equality flag densities (DPO-dim 14/56, base 26/94) are statistically indistinguishable; we therefore do not claim that dimensional training improves dimensional hygiene, only that it plausibly trends in that direction while leaving conservation untouched. Conservation training does not buy dimensional hygiene, and dimensional training does not buy conservation compliance: the two families are behaviorally independent, and each method's gain is confined to its targeted constraint.

Table 4: Dimensional-homogeneity probe: chains with ≥1 dimensionally inconsistent stated equality (sampled decoding, 320 chains per cell). p: paired exact McNemar vs. base on aligned chains. No difference is significant; the checker is reported as a negative-control probe.

Split	Scale	Base (%)	SFT (%)	CAFT (%)	CAFT vs. base p
ID	1.5B	8.12	9.38	5.94	0.311
ID	8B	15.31	13.12	16.56	0.747
ID	32B	0.31	0.00	0.62	1.000
OOD	1.5B	8.12	10.94	9.06	0.755
OOD	8B	13.12	13.75	13.75	0.899
OOD	32B	0.31	0.31	0.31	1.000

E Arithmetic-Checker Validation

The arithmetic-consistency probe of Section 4.8 uses a guarded checker: candidate equalities are matched with token boundaries, symbolic right-hand sides and exponent tails are ex-

cluded, chained equalities are evaluated pairwise, and markup such as thousands separators and fractions is normalized before matching. A chain is flagged only when the normalized equality fails exact arithmetic. Under this checker the flagged rate is at most 0.62% in every cell, with no significant system pair (all paired exact McNemar $p \geq 0.5$). We manually adjudicated all five residual flags: three genuine slips, one sign-shorthand transcription, and one deliberate reductio that the model itself labels impossible. The residual flags are disjoint from conservation violations.

F DPO Training Hyperparameters

Table 5 summarizes the training hyperparameters for all experiments. Tuning provenance: hyperparameters were adopted from standard DPO/SFT configurations (LoRA rank 16/α=32, β=0.1, learning rate 5×10^{-6} for DPO and 10^{-5} for SFT, 3 epochs) without any search over the evaluation scenarios: no validation split was carved from the 80 evaluation scenarios, no metric computed on evaluation outputs influenced any training or configuration decision, and the Suite-5 200-chain sample was used only for post-hoc verifier validation, never for tuning. The only per-scale adaptation was batch size/gradient accumulation to fit GPU memory (Table 5).

Table 5: Training hyperparameters for DPO and SFT experiments (verified against saved training_args.bin; warmup = 10% of total steps, linear schedule; LoRA targets all attention and MLP projections with dropout 0.05; optimizer AdamW). Max sequence length is the training pair-truncation budget and is distinct from the 800-token evaluation generation budget (Appendix I); the 8B seed-43/44 replicas use 1,024 rather than the headline 8B run's 2,048 (Section 4.6).

Parameter	1.5B	8B	32B
DPO β	0.1	0.1	0.1
Learning rate	5×10^{-6}	5×10^{-6}	5×10^{-6}
Epochs	3	3	3
Batch size	2	1	1
Gradient accumulation	4	8	16
Max sequence length	1,024	2,048	1,024
LoRA rank / α / dropout	16 / 32 / 0.05	16 / 32 / 0.05	16 / 32 / 0.05
Quantization	None	None	4-bit NF4 (QLoRA)
Precision	fp16	fp16	bf16
Warmup ratio	0.1	0.1	0.1
Seed	42	42	42
SFT learning rate	1×10^{-5}	1×10^{-5}	1×10^{-5}
SFT epochs	3	3	3

G Verifier Validation

To ensure the reliability of our conservation law verifier as a preference signal, we conduct a comprehensive validation organized as six test suites over 3,215 unique chains (5,030 checks in total, comprising 4,627 six-suite checks plus a separate 403-chain wrong-answer screen). Suites intentionally overlap—for example, the DPO chosen responses (Suite 4) are drawn from the clean-chain pool (Suite 1)—so the check counts sum to more than the unique-chain count; Table 6 gives the per-suite numbers and Table 7 documents the overlap structure. The headline precision/recall summaries in this appendix exclude Suite 5 (whose reference labels partially lean on verifier outputs, as detailed below), which is retained as a labeling-consistency check.

Suite 1: False positive rate on clean chains. We re-verified all 1,497 chains labeled as conservation-respecting during chain sampling. The verifier produced zero false positives (FPR = 0.00%), confirming that it does not incorrectly flag correct reasoning.

Table 6: Verifier validation results across six test suites. FPR = False Positive Rate. The suites comprise 4,627 checks over 3,215 unique chains; a separate 403-chain wrong-answer screen brings the total to 5,030 checks (overlap documented in Table 7).

Test suite	Checks	Metric	Value	Description
1. Clean chains	1,497	FPR	0.00%	Known-correct chains
2. Injected violations	500	Recall	100.0%	9 violation types
3. Real violations	1,006	Recall	100.0%	Natural + synthetic + same-answer rejects
4. Chosen responses	1,409	FPR	0.00%	DPO chosen (correct) chains
5. Eval chains	200	F1	1.000	Stratified sample
6. Edge cases	15	Accuracy	100.0%	Hand-crafted tests

Table 7: Unique-chain accounting across validation suites. Rows are disjoint chain populations identified by (scenario, chain text); the union is 3,215 unique chains. Suite memberships overlap: Suite 4's 1,409 chosen checks cover 620 unique chains already counted in the clean pool; Suite 3's natural rejects coincide with the 97 violating pool chains; the 403 wrong-answer rejects lie inside the pool (379 clean, 24 violating); and one synthetic reject coincides with an injected Suite-2 chain (shown as a −1 adjustment row).

Disjoint chain population	Unique chains
Clean sampled chains (Suite 1; includes Suite 4's 620 unique chosen)	1,497
Violating sampled chains (= Suite 3 natural rejects)	97
Injected-violation chains (Suite 2)	500
Synthetic + same-answer rejects not in the pool (Suite 3 remainder)	907
Evaluation-chains sample (Suite 5; disjoint from training pools)	200
Hand-crafted edge cases (Suite 6; synthetic text)	15
Shared text: one synthetic reject coincides with an injected Suite-2 chain	−1
Total unique chains	3,215
Total checks (suites summed)	4,627

Suite 2: Detection of injected violations. We injected violations into 500 known-correct chains using the ViolationInjector, covering all 9 violation types. The verifier detected all 500 violations (recall = 100%), with perfect detection per violation type.

Suite 3: Detection of real model violations. We tested the verifier on 1,006 DPO rejected responses from natural, synthetic, and same-answer-process pair types (i.e., pairs where the rejected chain contains a conservation violation). The verifier detected all 1,006 violations (recall = 100%).

Suite 4: False positive rate on chosen responses. We verified all 1,409 DPO chosen (conservation-respecting) responses. The verifier produced zero false positives (FPR = 0.00%). These chains are drawn from the same sampled pool as Suite 1, which is why the suite totals exceed the unique-chain count.

Suite 5: Agreement check on evaluation chains. We sampled 200 chains from evaluation outputs and compared the verifier's judgments against a reference labeling; the verifier agreed with the reference on all 200 (100%). We state plainly what this suite is and is not: the reference labels were inherited from the sampling-time pipeline labels (which the verifier's own outputs informed) rather than produced by an independent annotation protocol; the docstring of the validation script states explicitly that fully manual inspection of all 200 chains was not performed, and no separate annotator team or inter-annotator-agreement statistic exists for this suite. Suite 5 therefore measures labeling consistency and reproducibility rather than independent correctness. Its reference construction is verifier-informed, and we report it accordingly.

Suite 6: Hand-crafted edge cases. We constructed 15 edge-case chains testing specific verifier behaviors (e.g., correct inelastic collision with energy loss, correct friction with energy dissipation, wrong KE/PE formulas). The verifier correctly classified all 15 cases (accuracy = 100%).

Wrong-answer screen (outside the six suites). As a seventh, out-of-scope screen we verified all 403 DPO rejected responses whose rejected chain contains an answer error but no by-construction conservation violation. The verifier passed 379 of them (correct: answer errors are not conservation violations) and flagged 24 (5.9%) that additionally verbalize a templated misconception—e.g., claiming energy is conserved in a perfectly inelastic collision. Manual inspection confirmed all 24 flagged chains contain genuine conservation-violation language, so these are true detections relative to the verifier's target class rather than false positives; we report them separately because the by-design label of this population is "answer error," not "violation."

Numeric-recomputation cross-check. The pattern verifier judges utterances, not arithmetic; to probe whether chains that pass the pattern verifier ever state numerically inconsistent energy balances, we additionally parsed all 960 chains of the 1.5B in-distribution cells (base/SFT/CAFT × 320) for explicitly stated numeric before/after energies with mandatory Joule-unit anchoring (e.g., "$KE_i = 400$ J"). Only 21 chains (2.2%) state an extractable before/after pair; cross-tabulating the pattern verdict against the recomputed balance yields zero disagreements after manual adjudication (e.g., comparing KE_i against PE_f where the chain's stated balance is $KE_i{+}PE_i = KE_f{+}PE_f$), and no chain asserts conservation verbally while computing an inconsistent balance. This small-subsample check found no evidence of the numerical-violation blind spot hypothesized in Limitations, but coverage is inherently limited (most chains compute intermediate velocities/heights rather than explicit energy totals), and a full symbolic-recomputation verifier remains future work.

Summary. Over 2,911 clean-chain checks (Suites 1+4, plus Suite 6's 5 clean cases), the verifier produces 0 false positives (100% precision); over 1,516 conservation-violating checks (Suites 2+3, plus Suite 6's 10 violating cases), it produces 0 missed detections (100% recall). Suite 5's 200 checks are deliberately excluded from these headline summaries: its reference labels were inherited from sampling-time pipeline labels that the verifier's own outputs informed, so including it would count the verifier toward a partly self-derived standard; we report it as a labeling-consistency check only (100% agreement, no independent-correctness claim). Separately, among the 403 wrong-answer rejected chains—screened outside the six-suite scope—24 contain genuine conservation violations (e.g., claiming energy is conserved in a perfectly inelastic collision) that the verifier correctly flags; the remaining 379 contain only answer errors and are correctly passed. An "overall recall" metric that folded the wrong-answer screen into the violating population would yield 80.3% (1,540 = 1,516 six-suite violating checks + 24 wrong-answer violations, of 1,919 = 1,516 + 403), but that number conflates answer errors with process violations and is not the verifier's target metric. We emphasize the circularity caveat: Suites 1, 3, and 4 are labeled by the same verifier whose judgments they validate, so they establish self-consistency and robustness to negation phrasing rather than fully independent correctness; Suite 5, as disclosed above, also partially leans on verifier outputs and therefore adds labeling-consistency evidence rather than independent proof; the most independent evidence is Suite 6 (15 hand-crafted edge cases) plus the blind three-referee audit of Appendix H, which supplies the fully external check the other suites cannot. These results establish the verifier as a reliable preference signal for DPO training, with the boundary behavior documented in Section 4.5.

H Blind Multi-Auditor Physics Audit

The six validation suites above are largely self-referential (Suites 1/3/4 are labeled by the verifier itself; Suite 5 partially leans on verifier outputs). To obtain fully independent evidence, we conducted a blind audit of the verifier by three independent LLM-based referees ("Auditor 1," "Auditor 2," "Auditor 3"), each acting as a physics referee with no access to the verifier's labels, to each other's judgments, or to the paper's claims.

Design. From the four 1.5B evaluation cells (base/CAFT × ID/OOD, 320 chains each), we drew a stratified sample of 110 chains: all 30 chains flagged by the verifier in those cells (base_id 18, caft_id 4, base_ood 5, caft_ood 3) plus 20 randomly sampled unflagged "control" chains per cell (80 total). Each auditor received the 110 chains stripped of all provenance (cell identity, verifier verdict, model identity), re-derived the relevant quantities from the problem statement, and judged for each chain whether it contains a ma-

terial conservation-law violation: a false conservation claim, a corrupted or dropped energy/momentum balance, an invented conservation framework, or broken bookkeeping that makes the final answer physically wrong. Benign rounding, sloppy-but-self-correcting prose, verbal mislabels with correct mathematics, and hedged statements were not to be flagged.

Inter-auditor agreement. The three auditors judged 50, 41, and 47 of the 110 chains as violations, respectively. Pairwise raw agreement is 86–88% with Cohen's $\kappa = 0.72$–0.76 (A–A2 0.721, A–B 0.760, A2–B 0.736)—in the "substantial" band and well above what chance would produce at these base rates. Majority-of-three consensus marks 49/110 chains as violations. This agreement level indicates the violation construct is well-defined enough for independent referees to converge on it, which is a precondition for using the audit to score the verifier.

Verifier precision vs. auditor consensus. Of the 30 verifier-flagged chains, the auditor consensus judges 19 as material violations, 9 as non-violations, and 2 as one-of-three splits; excluding splits, precision against consensus is $19/28 = 67.9\%$. The 9 unanimous rejections are informative: 5 are chains whose only triggering utterance is a conditionally true or otherwise scoped statement—e.g., "mechanical energy is conserved as long as no non-conservative forces act" in friction scenarios, or "since there is no friction, all spring PE converts to KE"—that the verifier's patterns fire on out of context (the utterance-level limitation disclosed in Section 4.5); the verifier's negation handling and category exemptions absorb four of these nine, and a clause-level hedge guard (suppressing a flag when the assertion is conditioned on the absence of dissipative forces within its sentence window; tested end-to-end below) removes the remaining one. The remaining 4 are chains where the flagged sentence is a correct general law applied correctly (e.g., a properly balanced momentum equation the auditors judged numerically sound), i.e., disputes about materiality thresholds rather than pattern defects. Precision is thus bounded below by the disclosed hedge-pattern false-positive mode, not by a novel defect.

Verifier recall vs. auditor consensus (broad construct). The consensus marks 49 chains as violations; the verifier flags 19 of them (38.8%). This is expected and interpretable: the auditors were instructed to flag any materially corrupted energy/momentum bookkeeping—including purely numerical balance corruption (dropped final potential energy, factor-2 errors in $\frac{1}{2}mv^2 = mgh$, sign flips in explosion balances)—whereas the pattern verifier targets verbalized conservation claims. Most consensus violations the verifier misses are chains whose prose is conservation-neutral but whose numbers are wrong; the numeric-recomputation cross-check above probes exactly this gap on the small subset of chains with extractable balances and found no verbal assertion of conservation accompanying inconsistent arithmetic. The verifier is therefore a high-precision detector of a narrower (assertion-level) construct, not a complete detector of all physically deficient reasoning; the DPO signal it provides is correspondingly narrower than full physics quality.

Does the headline effect survive the broader lens? Because the sample enriches flagged chains, per-cell rates must be estimated by reweighting: each cell's broad-construct violation count is (consensus violations among its flagged chains) + (control violation rate scaled by the cell's unflagged population). Under the auditors' broader standard, the estimated violation rates are base_id ≈70.6% vs. caft_id ≈30.4% (in-distribution), and base_ood ≈25.2% vs. caft_ood ≈35.0% (out-of-distribution)—an apparent reversal of the verifier-reported OOD direction. We do not interpret this reversal: the OOD reweighted estimates rest on 3–5 flagged chains per cell and 20 controls per cell, so both OOD rates carry sampling intervals wider than the gap between them, and the verifier's own OOD comparison is likewise at the floor (Section 4.4); the audit sample was enriched for in-distribution signal, not powered for OOD contrasts. The in-distribution CAFT effect is preserved on the clean control subsamples, which carry no enrichment bias: auditors judge 14/20 base controls vs. 6/20 CAFT controls as violations (two-sided Fisher p=0.026), while the OOD controls show no significant difference (4/20 vs. 6/20, p=0.72)—mirroring the main text's pattern in which the OOD CVR comparison is not a supported claim. The central conclusion is thus robust to moving from the verifier's narrow assertion-level construct to the auditors' broader physics-quality construct: CAFT-trained chains contain materially fewer conservation defects under an independent, stricter lens, and the reduction is concentrated in-distribution exactly as the verifier reports.

Robustness: a fourth auditor. Two dependencies could in principle drive the audit's conclusions: the specific trio of referees, and the verifier that selected the flagged sample. We therefore added a fourth independent auditor, "Auditor 4" (same blindness protocol, no access to the other raters or to the verifier)—a distinct rater from Auditor 1 of the three-rater panel, with a separate flag file (Auditor 1 flags 50/110; Auditor 4 flags 35/110)—and re-scored all 110 sampled chains with the verifier of Appendix G. The four raters flag 35 (Auditor 4), 41 (Auditor 2), 50 (Auditor 1), and 47 (Auditor 3) of the 110 chains, respectively; all six pairwise κ values span 0.616–0.760 (median 0.720), and majority-of-four consensus (2–2 ties treated as splits) marks 40/110 chains as violations with 11 splits. Against this four-rater consensus, the verifier flags 27 of the 110 chains (three of the thirty flags are dropped, all three unanimously auditor-clean), judging 16 as material violations, 7 as non-violations, with 4 splits: precision $16/23 = 69.6\%$, consistent with the three-rater estimate (67.9%), and recall against the four-rater consensus of $16/40 = 40.0\%$, consistent with the three-rater 38.8%. The in-distribution control contrast persists (11/20 base-ID vs. 4/20 CAFT-ID controls judged violations—55.0% [34.2, 74.2] vs. 20.0% [8.1, 41.6], Wilson 95% CIs; two-sided Fisher p=0.048), and the reweighted broad-construct rates remain separated (base-ID 63.0% [44.2, 81.9] vs. CAFT-ID 25.3% [10.5, 42.6], bootstrap CIs). Every headline conclusion of this appendix—substantial inter-referee agreement, verifier precision in the two-thirds range bounded by the disclosed hedge-pattern family, narrow-construct recall near 40%, and a preserved in-distribution effect on verifier-clean controls—is unchanged under the four-rater protocol.

Consensus-endpoint contrast: CAFT vs. base on the auditors' own definition. The analyses above score the verifier against auditor consensus; a complementary question is how the models compare when the auditors' consensus itself is the endpoint. Under the four-rater majority rule (2–2 ties excluded as splits), the consensus-defined violation rates are base-ID $23/33 = 69.7\%$ [52.7, 82.6] vs. CAFT-ID $5/20 = 25.0\%$ [11.2, 46.9] (Wilson 95% CIs; two-sided Fisher p=0.0021), while the OOD cells show no separation (base $5/23 = 21.7\%$ [9.7, 41.9] vs. CAFT $7/23 = 30.4\%$ [15.6, 50.9], p=0.738)—the same ID-strong/OOD-null pattern as the verifier-reported rates. Pooling the four cells (base 28/56 vs. CAFT 12/43) gives Fisher p=0.038. Two caveats keep this honest: the sample is enriched for verifier-flagged chains (30 of 110), so these full-sample rates are not unbiased estimates of the per-cell population rates (the verifier-clean-control rates in the preceding paragraph are the unbiased ones); and the pooled test ignores the ID/OOD stratification, under which the effect is carried by the ID stratum alone. The value of this analysis is triangulation: an endpoint defined purely by blind human-side adjudication, with no involvement of the verifier's patterns, reproduces the direction and the in-distribution concentration of the headline effect.

Hedge-guard ablation. The precision analysis above identifies a clause-level hedge guard as a candidate precision improvement; we test this directly. We implemented the guard as a post-filter on the verifier: a flag is suppressed when the sentence containing the matched utterance also conditions the assertion on an idealized regime—phrases such as "as long as no non-conservative forces act," "since there is no friction," "if no external forces act," "frictionless," or "neglecting air resistance" (code/hedge_guard_ablation.py in the released code; seven curated sanity tests covering hedged and unhedged formulations all behave as intended). Re-scoring all four 1.5B evaluation cells (1,280 chains) under the guarded verifier changes no in-distribution count (base 18→18, CAFT 4→4; the headline McNemar remains 18 vs. 4 discordant pairs, p=0.0043) and removes exactly one out-of-distribution flag (base-OOD 4→3), a chain the auditor consensus unanimously judged clean that the verifier's negation window misses because the conditioning phrase lies further from the match than its 20-character lookahead reaches. On the 110-chain audit sample the guard removes one flag (the same residual false positive) and adds none, raising precision against consensus from $16/23 = 69.6\%$ to $16/22 = 72.7\%$ (excluding splits) with zero loss of auditor-material detections. Two observations bound the guard's value honestly. First, the verifier's own negation handling and category exemptions already absorb most of the hedge-pattern false positives that motivated the guard (nine unanimous auditor rejections shrink to one residual case), so the guard's measurable benefit on this benchmark is small by construction. Second, the guard is deliberately conservative—it fires only on explicit conditioning language within the sentence—so it does not risk masking genuine violations; indeed it removed no auditor-

material flag anywhere in the sample. We release the guard as an optional scoring variant rather than adopting it as the default, preserving comparability with the default scoring reported in this paper; deployments that prioritize precision over recall can enable it at a known, negligible cost to sensitivity (-1 flag in 1,390 scored chains, both cells affected being flagged chains the auditors unanimously rejected).

Scope and limitations of the audit. The referees are LLM-based rather than human experts, so this audit measures agreement between independently prompted model referees and the verifier, not human-expert ground truth; all three share the failure modes of their model class. The sample is small (110 chains) and stratified rather than uniform, so the per-cell rate estimates above carry wide intervals (e.g., ± 20 pp on control-rate extrapolations), and the two split chains show that single-chain adjudication remains genuinely ambiguous at the boundary. Nonetheless, the combination of substantial inter-referee agreement ($\kappa \geq 0.72$), a preserved in-distribution effect on verifier-clean control chains (p=0.026 three-rater, p=0.048 four-rater—both inside the 0.01–0.05 band we ourselves designate provisional, so we read this as corroborating rather than confirmatory), and a mechanistic account of every verifier false positive (one disclosed pattern family) provides the strongest independent evidence in this paper that the verifier's target construct is real and that CAFT's in-distribution effect holds under the verifier's specific patterns.

I Compute Cost and Training Statistics

Table 8 reports the compute cost and training statistics for all main and ablation experiments. All experiments use LoRA fine-tuning, making the approach computationally efficient—the largest experiment (8B DPO) trains in under 2 hours on 2×L40 GPUs. Sequence-length budgets: training uses the per-scale max_length of Table 5 (1,024/2,048/1,024), while evaluation uses a separate 800-token generation budget per chain; these constrain different stages. One deviation is disclosed in Section 4.6: the 8B seed-43/44 replicas were trained with max_length 1,024 rather than the headline run's 2,048; exact tokenization of all 1,379 pairs with the Qwen3-8B tokenizer under TRL's concatenation shows 0 pairs exceed 1,024 tokens (worst concatenated length 1,002), so the deviation is nominal at the data level, though it voids exact-replication claims; the evaluation-side 800-token generation budget is unaffected by this training-side deviation.

J Additional Tables, Figures, and Algorithm

This appendix collects supporting tables, figures, the CAFT algorithm, and the full violation-pattern list referenced from the main text.

J.1 CAFT Algorithm

Algorithm 1: Conservation-Aware Fine-Tuning (CAFT)

Input: Base LLM π_{ref}, scenario set $\mathcal{S}$, verifier $\mathcal{V}$
Output: Conservation-aware model π_θ
Step 1: Generate physics scenarios $\mathcal{S}$ with computable conservation laws;
Step 2: For each $s \in \mathcal{S}$, sample N chains $\{c_1, \ldots, c_N\} \sim \pi_{\text{ref}}(\cdot|s)$ at temperature T;
Step 3: For each chain c_i, run verifier: $\mathcal{V}(c_i) \to \{\text{valid}, \text{violation type}\}$;
Step 4: Construct DPO pairs using four strategies:;
  (A) Natural: pair valid vs. violating chains from same scenario;
  (B) Synthetic: inject violations into valid chains;
  (C) Same-answer-process: same answer, injected process violation;
  (D) Wrong-answer: correct-answer vs. wrong-answer chains;
Step 5: Fine-tune π_θ with DPO on all pairs using LoRA;

Table 8: Compute cost and training statistics. Trainable params: LoRA adapters only. Wall-clock time includes training only (excludes data generation and evaluation).

Model	Method	Steps	Trainable Params	Time (min)	GPU
Main experiments					
Qwen2.5-1.5B	DPO+LoRA	477	18.5M (1.18%)	∼57	1×L40
Qwen3-8B	DPO+LoRA	468	43.6M (0.53%)	∼110	2×L40
Qwen3-32B	DPO+QLoRA	219	112M (0.35%)	∼420	8×L40
Qwen2.5-1.5B	SFT+LoRA	195	18.5M (1.18%)	∼5	1×L40
Qwen2.5-1.5B	SFT+LoRA (step-matched)	429	18.5M (1.18%)	∼11	1×L40
Qwen3-8B	SFT+LoRA	219	43.6M (0.53%)	∼19	2×L40
Qwen3-32B	SFT+QLoRA	225	112M (0.35%)	∼300	8×L40
Ablation experiments (1.5B DPO+LoRA)					
Energy-only	DPO	114	18.5M	—	1×L40
Momentum-only	DPO	111	18.5M	—	1×L40
Natural-only	DPO	39	18.5M	—	1×L40
Synthetic-only	DPO	201	18.5M	—	1×L40
Same-answer-proc.	DPO	108	18.5M	—	1×L40
Wrong-answer-only	DPO	138	18.5M	—	1×L40

Aggregate compute: ∼101 GPU-hours summed over the runs above (the 32B QLoRA runs dominate at 8-way parallelism; wall-clock elapsed time is far lower since runs were concurrent). Data generation (chain sampling) requires an additional ∼4 GPU-hours. Ablation timings were not logged per-run; the Time column is omitted for ablation rows for that reason (steps are recorded from the training logs). Step-matched SFT control (Section 4.6): retrained SFT-1.5B on 1,272 upsampled samples (from 578 unique SFT-eligible chains) under identical batch/accumulation/epochs; the 10% held-out split leaves 1,144 training samples, giving $\lceil 1144/8 \rceil \times 3 = 429$ optimizer steps (verified against the saved trainer state), i.e. 2.2× the original SFT run and within 10% of DPO's 477. Its evaluation uses the standard protocol (320 sampled chains, ID, eval seed 42, identical scenario/chain indexing): CVR 5.31% (17/320), accuracy 35.62%, CAA 33.75%; greedy CVR 6.25%, accuracy 53.75%. Paired against base under the verifier: CVR 14 vs. 13 discordant pairs (p=1.0), accuracy 75 vs. 60 (p=0.23); against CAFT: CVR 4 vs. 17 (p=0.007), accuracy 115 vs. 46 (p=5.3×10^{-8}).

J.2 The Nine Violation Patterns

The verifier detects the following 9 violation patterns across the 4 categories defined in Section 3.3:

1. Energy creation: claiming energy appears from nowhere in a conservative system
2. Energy loss in conservative system: claiming energy is lost without non-conservative forces
3. Momentum creation: claiming momentum increases without external force
4. Momentum loss: claiming momentum decreases without external force
5. Wrong KE formula: using an incorrect kinetic energy formula
6. Wrong PE formula: using an incorrect potential energy formula
7. Inelastic as elastic: treating inelastic collisions as kinetic-energy-conserving
8. Elastic as inelastic: treating elastic collisions as energy-dissipating
9. Friction energy conserved: claiming mechanical energy is conserved with friction

Table 9: DPO preference pair statistics by construction strategy.

Strategy	1.5B Pairs	8B Pairs	32B Pairs	Description
Natural (A)	97	154	37	Valid vs. violating from same scenario
Synthetic (B)	590	646	761	Injected violations into valid chains
Same-answer-process (C)	319	295	375	Same answer, process violation
Wrong-answer (D)	403	284	109	Correct vs. incorrect answer
Total	1,409	1,379	1,282	

Table 10: Per-category results under sampled decoding (CVR% / Accuracy%). CAFT improves accuracy in all 8 categories and reduces CVR in categories where violations occur.

Category	Base	SFT	CAFT
Elastic collision	5.0 / 10.0	5.0 / 7.5	5.0 / 15.0
Inelastic collision	17.5 / 60.0	22.5 / 45.0	2.5 / 77.5
Projectile energy	0.0 / 47.5	0.0 / 47.5	0.0 / 75.0
Pendulum energy	0.0 / 52.5	0.0 / 15.0	0.0 / 67.5
Spring energy	0.0 / 50.0	2.5 / 52.5	0.0 / 75.0
Friction dissipation	22.5 / 47.5	15.0 / 40.0	2.5 / 72.5
Explosion/separation	0.0 / 20.0	0.0 / 27.5	0.0 / 32.5
Roller coaster energy	0.0 / 35.0	0.0 / 37.5	0.0 / 42.5

J.3 DPO Pair Statistics

J.4 Per-Category Results (1.5B)

J.5 Ablation Tables and Figure

J.6 Out-of-Distribution Results (1.5B)

J.7 Scale-Up Details (8B and 32B)

J.8 Violation Type Breakdown

J.9 Qualitative Example

K Ablation Details

This appendix reports the ablation runs summarized in Section 4.3: the necessity (leave-one-strategy-out) results, the leave-family-out transfer test, and the complementary reversed-design transfer test. All runs are at 1.5B with identical DPO hyperparameters; evaluation is the standard protocol (320 sampled chains, in-distribution, seed 42) unless stated otherwise.

Leave-one-strategy-out (LOO) ablation. To test necessity rather than sufficiency, we trained four additional 1.5B models, each removing one strategy from the pair mixture while keeping the remaining pairs and the upsampling budget intact, and evaluated them identically (320 chains, seed 42). Removing synthetic pairs eliminates the CVR reduction (4.06%, vs. base 5.62%, p=0.42; vs. full CAFT p=0.049 unclustered, p=0.060 under scenario clustering—Appendix C), and removing natural pairs does the same (3.75%, p=0.33 vs. base)—the two conservation-bearing strategies are jointly necessary for the violation effect. Removing same-answer-process pairs leaves CVR reduced (1.88%, p=0.017 vs. base) but costs accuracy (52.19% vs. 57.19%, p=0.15 vs. full CAFT), and removing wrong-answer pairs preserves the strongest CVR of all arms (0.94%, p=0.0003 vs. base) while collapsing the accuracy co-benefit (45.62%, p=0.0008 vs. full CAFT). The LOO results refine the single-strategy picture into a division of labor: synthetic and natural pairs drive the conservation benefit, wrong-answer pairs drive the accuracy co-benefit, and same-answer-process pairs contribute moderately to both. Every LOO arm underperforms full CAFT on the joint CVR/accuracy objective (three of four significantly on accuracy, $p \leq 0.019$—two of those inside the provisional 0.01–0.05 band, so the division-of-labor attribution inherits that provi-

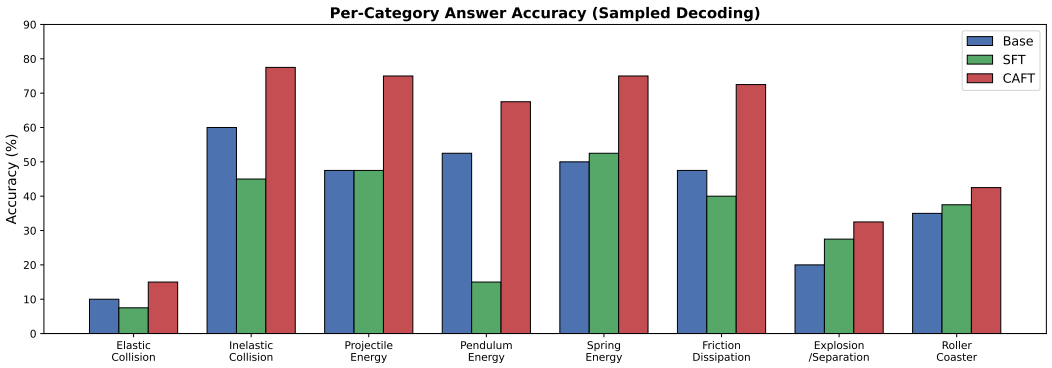

Figure 2: Per-category answer accuracy under sampled decoding at 1.5B in-distribution (40 chains per category). CAFT improves or matches accuracy in all 8 physics categories.

Table 11: Per-law ablation under sampled decoding. Energy-only training matches full CAFT on CVR but not accuracy, while momentum-only training shows partial CVR reduction. Violation rates are broken down by type: energy violations, momentum violations, and wrong law application.

Model	Pairs	CVR ↓	Acc ↑	CAA ↑	Energy V.	Mom. V.	Wrong Law
Base (no training)	0	5.62	40.31	38.12	4.06	0.31	1.25
Energy-only	705	1.25	43.12	42.50	0.62	0.31	0.31
Momentum-only	327	4.69	48.12	46.56	4.06	0.00	0.62
Full CAFT	1,409	1.25	57.19	56.56	0.31	0.00	0.94

sionality), supporting the mixture design while attributing specific contributions to specific strategies (Table 13, Appendix J). The LOO arms are run at a single seed with small per-arm discordant counts (4–16 events), so the division-of-labor attribution is read as directional; Limitations (10) quantifies this.

Leave-family-out (LFO) transfer: is the benefit family-local or a global style hedge? A reviewer concern motivates one final ablation: if the verifier's lexical patterns teach only a global "sound physical" style, training on seven violation families should still suppress violations on the eighth, held-out family. We therefore removed all inelastic-collision pairs from the training mixture (1,177 pairs spanning the other seven violation families) and trained DPO with identical hyperparameters. The result splits cleanly: on the seven trained families, the LFO model reduces violations strongly (2/280 vs. base 11/280; 11 vs. 2 discordant pairs, p=0.023) and preserves the accuracy co-benefit (50.0% vs. 40.3%, p=0.010); on the held-out inelastic family it does not reduce violations at all (8/40 vs. base 7/40, p=1.0; the held-out family contributes 40 chains, a regime in which even a moderate reduction would rarely reach significance, and the informative comparison is with full CAFT's near-cleanup of the same 40 chains), whereas full CAFT—which saw inelastic pairs—nearly cleans out that category (1/40, p=0.070 vs. base). Overall LFO CVR (10/320, 3.12%) sits between base and full CAFT (McNemar vs. base p=0.152), as expected from the mixture composition. Two conclusions follow. First, the conservation benefit is family-local: it tracks the violation families represented in the preference pairs rather than acting as a purely global stylistic hedge, which addresses the concern that lexical-pattern training might act as an undifferentiated register shift. Second, the benefit is not purely lexical memorization either: the LFO model still generalizes within each trained family to unseen scenario instances, and the accuracy co-benefit survives holding out an entire violation family. The honest reading—consistent with the verifier's lexical scope (Section 3.3)—is that CAFT teaches suppression of the specific verbalized violation families it is shown, with genuine within-family generalization but no cross-family transfer; deployments facing violation families absent from the training mixture should not expect protection.

Table 12: Pair strategy ablation under sampled decoding. No single strategy matches full CAFT on the joint CVR/accuracy objective, consistent with complementary contributions (leave-one-strategy-out results in Table 13). DPO evaluation metrics (held-out 10% split) are shown in the last three columns: eval loss, rewards/accuracy, and rewards/margin.

Strategy	Pairs	CVR ↓	Acc ↑	CAA ↑	Eval Loss	DPO Acc.	DPO Margin
Base (no training)	0	5.62	40.31	38.12	—	—	—
Natural only	97	5.62	41.25	38.75	0.618*	0.960*	0.151*
Synthetic only	590	1.56	53.12	52.50	0.016	1.000	4.814
Same-answer-proc.	319	2.81	47.50	45.62	0.230	1.000	1.425
Wrong-answer only	403	5.00	45.62	44.38	0.619	0.643	0.181
Full CAFT (all four)	1,409	1.25	57.19	56.56	0.281	0.845	2.308

*Natural-only has no eval split (97 pairs $<$ 100 threshold); training log values shown.

Table 13: Leave-one-strategy-out (LOO) ablation (1.5B, in-distribution, sampled decoding, 320 chains, seed 42). Each row removes one strategy from the full 1,409-pair mixture (remaining pairs kept intact). p_{base}: paired exact McNemar on CVR vs. base; p_{full}: vs. full CAFT. Removing synthetic or natural pairs eliminates the CVR reduction; removing wrong-answer pairs preserves CVR but collapses the accuracy co-benefit. Disc. gives the discordant-pair counts (base-violating/chosen-clean vs. reverse) underlying p_{base}. Wall-clock times for these LOO runs were not logged per-run (Appendix I reports timings only for the main and single-strategy ablation runs).

Removed strategy	CVR (%)	Acc (%)	CAA (%)	Disc. (base→arm / rev.)	p_{base} (CVR)	p_{full} (CVR)
None (full CAFT)	1.25	57.19	56.56	18 / 4	0.004	—
Natural pairs	3.75	48.75	46.56	16 / 10	0.327	0.077
Synthetic	4.06	46.88	44.69	15 / 10	0.424	0.049
Same-answer-process	1.88	52.19	51.25	17 / 5	0.017	0.754
Wrong-answer	0.94	45.62	45.00	16 / 1	0.0003	1.0
Base (no training)	5.62	40.31	38.12	—	—	—

Complementary held-out-family transfer with reversed design. The LFO ablation above holds out one family while training on the other seven; a complementary test reverses the design—train on a strict subset of families and hold out five entirely (momentum creation/loss, elastic-as-inelastic, inelastic-as-elastic, and friction-energy-conserved). We rebuilt the full pair pipeline with the synthetic injector restricted to the four energy/formula families and with every natural/same-answer/wrong-answer pair whose rejected side touches a holdout family filtered out (1,341 pairs; a marker scan over all rejected sides confirms zero holdout-family content), trained DPO with identical hyperparameters, and evaluated under the standard protocol. The family-level decomposition is sharp: the transfer model eliminates every train-family flag (0/320 flags belong to energy/formula families, vs. 8 flags in the base model) while its 9 residual flags are all holdout-family violations—7 inelastic-energy claims, 1 collision-regime confusion, 1 friction-conservation claim—i.e., the model still makes exactly the kinds of errors it was never shown. Holdout-category violations fall only from 18/160 (base) to 9/160 (transfer; 17 vs. 8 discordant pairs, p=0.108, not significant—and the reduction that does appear is confined to the energy-claim component that overlaps the trained families), whereas full CAFT, trained on all families, reaches 4/160. Meanwhile the accuracy co-benefit concentrates where the training signal does: train-family categories improve from 46.25% (base) to 63.75% (transfer, p=0.0008; full CAFT 65.0%), while holdout-category accuracy is unchanged (34.4%→39.4%, p=0.37). Together with the LFO result, this gives bidirectional evidence for family-locality: violations are suppressed only for families present in the preference pairs, and neither the violation benefit nor the accuracy benefit bleeds into structurally held-out families. We archive the per-family decomposition in results/transfer_heldout_family_stats.json.

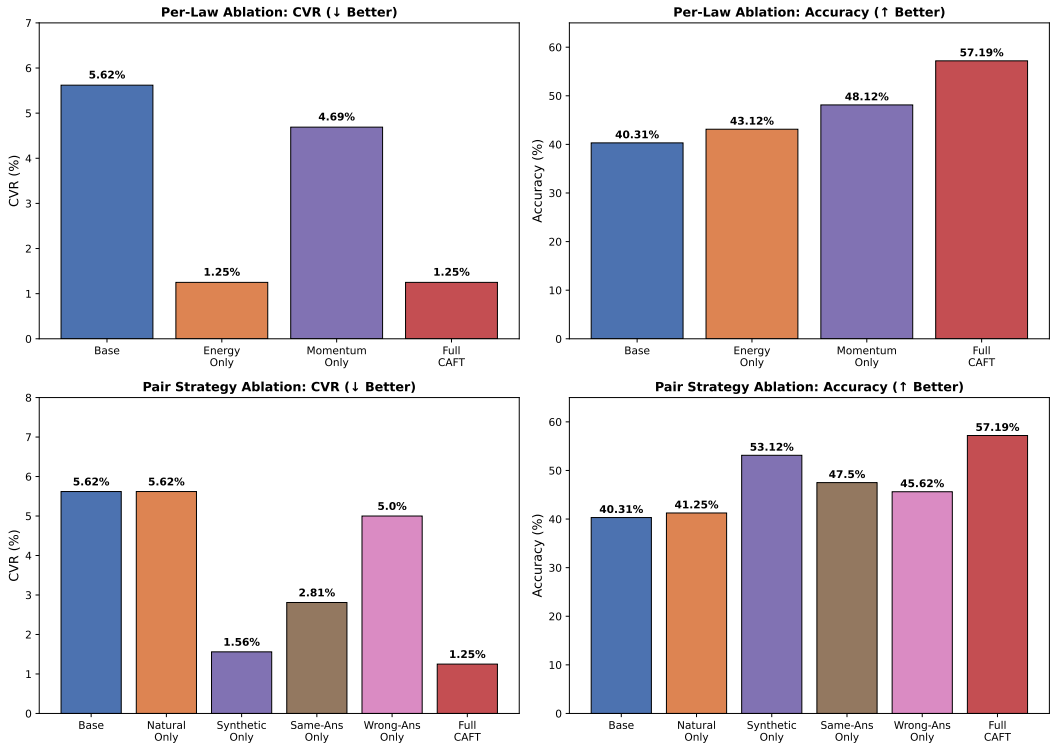

Figure 3: Ablation study results under sampled decoding. Top: per-law ablation (energy-only vs. momentum-only). Bottom: pair strategy ablation (four individual strategies vs. full CAFT). No single ablation matches full CAFT on both CVR and accuracy.

Table 14: OOD results on 80 unseen scenarios (1.5B). CAFT generalizes well, improving sampled accuracy by 16.25 pp while keeping CVR near zero; note that base greedy accuracy (61.25%) exceeds CAFT's (55.00%) under deterministic decoding.

Method	Decoding	CVR (%) ↓	Acc (%) ↑	CAA (%) ↑	PVS ↑
Base	Sampled	1.25	37.81	37.81	0.9989
	Greedy	0.00	61.25	61.25	1.0000
SFT	Sampled	0.00	39.38	39.38	1.0000
	Greedy	0.00	53.75	53.75	1.0000
CAFT	Sampled	0.31	54.06	54.06	0.9999
	Greedy	0.00	55.00	55.00	1.0000

L Robustness Details

This appendix reports the per-seed cells of the 1.5B and 8B replications and the matched-budget outcome-only comparison summarized in Section 4.6. Table 19 lists all seed runs; Table 20 lists the outcome-only arms.

M Full Scale-Up Table and 32B Detail

Table 21 reports the scale-up cells under the final evaluation protocol.

Anatomy of the 32B difference, in full. Because the 0.31-point CVR gap (1/320 vs. 2/320) is the only apparent 32B regression, we examined both flagged CAFT chains individually. One is a boundary case of the pattern verifier: the chain states "Energy is conserved, but

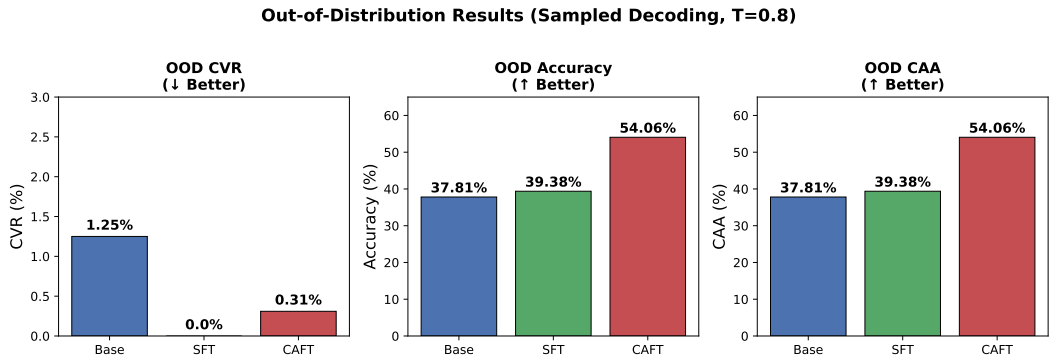

Figure 4: OOD results under sampled decoding. CAFT generalizes well to unseen scenarios, improving accuracy by 16.25 pp over the base model.

Table 15: OOD generalization at 8B and 32B scale. Sampled: $T = 0.8$, 320 chains. Greedy: 80 chains. At 32B, all methods achieve 0% OOD CVR, consistent with the ceiling effect observed in-distribution. Under the strip-first protocol, 8B OOD shows essentially no violations and no accuracy difference: base 0/320, CAFT 1/320 (0.31%; the single flag is a disclosed verifier false positive—the chain correctly states "the only external force acting on the system is internal," which the momentum pattern fires on without registering the internal-only scoping), replicas 0/320; base 75.62% sampled / 78.75% greedy vs. CAFT 78.44% (seed 43) and 79.38% (seed 44) sampled, 77.50%/81.25% greedy (paired sampled McNemar vs. base: p=0.32 and p=0.14). The 1.5B OOD greedy caveat (base 61.25% vs. CAFT 55.00%, Table 14) is unaffected and stands as the paper's genuine decoding-regime cost.

Method	Decoding	CVR (%) ↓	Acc (%) ↑	CAA (%) ↑	PVS ↑
32B OOD					
Base	Sampled	0.00	89.06	89.06	1.0000
	Greedy	0.00	88.75	88.75	1.0000
SFT	Sampled	0.00	90.31	90.31	1.0000
	Greedy	0.00	90.00	90.00	1.0000
CAFT	Sampled	0.00	90.62	90.62	1.0000
	Greedy	0.00	90.00	90.00	1.0000

mechanical energy is transformed into thermal energy due to friction"—physically correct (total energy is conserved; mechanical energy is dissipated)—but the friction-scenario pattern flags any unqualified "energy is conserved" claim. The other flagged chain, and the single base-model violation, occur on the same elastic-collision scenario where the model treats an elastic collision as inelastic; CAFT's flag is a different sampled chain of the same scenario. Formally, the base-vs-CAFT CVR difference rests on one discordant pair (McNemar exact p=1.0), the Wilson 95% CIs overlap almost completely ([0.06, 1.75] vs. [0.17, 2.25]), and the flagged-event rate is too small to support any directional claim. A clause-level semantic fix to the friction pattern would change CAFT-32B sampled CVR from 0.62% to 0.31% and no other cell (Appendix A), which is why we report the 32B result as no measurable CVR change rather than as a regression.

N Training Dynamics

Figure 6 shows the training loss curves for DPO and SFT at all three model scales. DPO training converges from the initial ∼0.69 (random preference) to 0.15 (1.5B), 0.17 (8B), and 0.22 (32B), indicating that the model learns to discriminate conservation-respecting from violating chains; SFT loss converges to 0.38 (1.5B), 0.23 (8B), and 0.15 (32B). The slightly higher DPO terminal loss at 32B is consistent with the larger model having less room for conservation improvement (ceiling effect)—the preference signal is less informative

Table 16: Full 32B results under both decoding strategies. All methods achieve near-zero CVR, confirming the ceiling effect. CAFT-32B improves accuracy but does not reduce CVR.

Setting	Method	CVR (%) ↓	Acc (%) ↑	CAA (%) ↑	PVS ↑
In-Distribution					
Sampled	Base	0.31	86.56	86.56	0.9990
	SFT	0.31	88.44	88.44	0.9999
	CAFT	0.62	88.12	87.81	0.9989
Greedy	Base	0.00	87.50	87.50	1.0000
	SFT	0.00	87.50	87.50	1.0000
	CAFT	0.00	88.75	88.75	1.0000
OOD					
Sampled	Base	0.00	89.06	89.06	1.0000
	SFT	0.00	90.31	90.31	1.0000
	CAFT	0.00	90.62	90.62	1.0000
Greedy	Base	0.00	88.75	88.75	1.0000
	SFT	0.00	90.00	90.00	1.0000
	CAFT	0.00	90.00	90.00	1.0000

Table 17: Per-category results at 32B scale under sampled decoding (CVR% / Accuracy%). Violations are extremely rare across all categories.

Category	Base 32B	SFT 32B	CAFT 32B
Elastic collision	2.5 / 20.0	2.5 / 20.0	2.5 / 25.0
Inelastic collision	0.0 / 95.0	0.0 / 92.5	0.0 / 87.5
Projectile energy	0.0 / 97.5	0.0 / 100.0	0.0 / 100.0
Pendulum energy	0.0 / 100.0	0.0 / 100.0	0.0 / 100.0
Spring energy	0.0 / 97.5	0.0 / 100.0	0.0 / 100.0
Friction dissipation	0.0 / 95.0	0.0 / 100.0	2.5 / 95.0
Explosion/separation	0.0 / 87.5	0.0 / 95.0	0.0 / 97.5
Roller coaster energy	0.0 / 100.0	0.0 / 100.0	0.0 / 100.0

when the base model already produces mostly conservation-respecting chains. Converged-loss magnitudes are not comparable across methods or scales: SFT minimizes a different objective (next-token likelihood on chosen chains), so its lower terminal value at 32B reflects that easier objective plus the ceiling, not superior preference learning.

O Scale-Up, Protocol, and Robustness Details

This appendix collects the detail behind Sections 4.5, 4.6, and 4.6: the scale summary table, the 32B pair-scarcity and flag analysis, the derivation and caveats of the deployment rule, the full seed-variance inference, the matched-budget and step-matched controls, and the two zero-training alternatives.

O.1 Scale summary table

O.2 32B: pair scarcity versus ceiling

The 32B run had the fewest natural violating pairs (37 of 1,282, vs. 97/1,409 at 1.5B and 154/1,379 at 8B, since natural pairs can only be harvested where the base model violates), but two observations argue that scarcity is a consequence of the ceiling rather than an independent cause: the synthetic-injection channel is deliberately scale-independent (761 synthetic pairs at 32B, the largest of any scale), yet produced no CVR reduction; and the single base violation anchoring the 0.31% rate leaves almost no violation mass for any signal to remove. The full anatomy of the flagged chains—including one hedged-but-correct statement that the friction pattern flags, and an elastic-collision scenario where the flagged

Table 18: Violation type breakdown (sampled decoding, 320 chains). CAFT reduces all violation types, with the largest reduction in energy violations.

Violation Type	Base (%)	CAFT (%)
Energy violation	4.06	0.31
Wrong law application	1.25	0.94
Momentum violation	0.31	0.00
Total (CVR)	5.62	1.25

Base Model (Violation: energy_violation)

Problem: Perfectly inelastic collision on a frictionless surface.

Reasoning: "Since there's no friction involved and we're dealing with perfectly inelastic collisions, momentum and kinetic energy are conserved during the collision."

Error: Incorrectly claims kinetic energy is conserved in an inelastic collision (only momentum is conserved).

Answer: Incorrect

CAFT (No Violation)

Reasoning: "For a perfectly inelastic collision where objects stick together, the initial momentum equals the final momentum of the combined object."

Correct: Applies only momentum conservation, recognizing that kinetic energy is not conserved in inelastic collisions.

Answer: Correct

Figure 5: Qualitative comparison on an inelastic collision problem. The base model incorrectly applies energy conservation to an inelastic collision, while CAFT correctly applies only momentum conservation.

CAFT chain is a different sample of the same scenario as the single base violation—is in Appendix M.

O.3 Deployment guideline: derivation and caveats

The rule screens for potential headroom, not realized benefit. The 1.5B OOD cell shows its granularity: the base OOD CVR interval [0.49, 3.17]% technically excludes zero, yet the OOD improvement is not significant (4 vs. 1 discordant events, p=0.375). Two practical caveats follow. Scenario difficulty mediates the ceiling (a 32B model at 86.6% accuracy leaves little violation mass for any intervention), so harder distributions shift the threshold scale at which the rule fires; and the rule is calibrated on sampled decoding, so practitioners should evaluate both decoding modes at their operating point (at 1.5B, greedy CAFT CVR is nominally higher than base; Section 4.2). The deepest caveat is that the guideline's first clause is only as trustworthy as the evaluation protocol that measures CVR: at 8B the base model's CVR interval under our strip-first protocol is $[0.00, 1.19]$%, so the rule recommends skipping CAFT at that scale.

O.4 Seed variance: full inference

Pooling all three seeds against base (960 aligned chain pairs) confirms both effects jointly (154 vs. 231 accuracy pairs, p=1.0×10^{-4}; 51 vs. 20 CVR pairs, p=3.0×10^{-4}), and a scenario-level sign-flip randomization (240 blocks, 20,000 permutations) gives clustered p=0.0012 (per-seed p=0.012/0.067/0.38). Seed-level inference that respects the run structure agrees: the three per-seed CVR tests combine under Fisher's method to p=0.003 and under Stouffer to Z=3.43 (one-sided p=3.0 × 10^{-4}), with 3/3 concordant directional reductions. On accuracy, per-seed gains are +16.88/+3.13/+4.07 pp (p<10^{-4}/0.42/0.30), so the seed-42 magnitude is one draw from a wide run distribution, although every seed is non-negative; the OOD accuracy co-benefit replicates more stably (+16.25/+10.00/+7.81 pp across seeds 42/43/44, each individually significant).

Table 19: Seed variance of CAFT at 1.5B and 8B (in-distribution, sampled decoding, 320 chains per run). Discordant pairs are base-violating/CAFT-clean vs. the reverse; accuracy p-values compare each seed against the base model. [a]8B protocol. The 8B rows use the strip-first protocol (Section 4.5): the verifier is applied to thinking-stripped chains, under which CVR is 0/320 in every arm (the corresponding OOD cells are base 0/320, CAFT 1/320—a disclosed verifier false positive). Seeds 43/44 were evaluated under this protocol; their accuracy deltas are computed against the strip-first base row. Strip-first accuracy differences are not significant (base 77.81% vs. CAFT 81.25%, paired exact McNemar 33 vs. 22 discordant pairs, p=0.177; Section 4.5); the SFT row is the original SFT-8B checkpoint re-evaluated under the strip protocol (a step-matched SFT control is discussed in Section 4.6).

Scale	Run	CVR (%)	95% CI	Acc (%)	ΔAcc vs. base	McNemar p (CVR)
1.5B	Base (no FT)	5.62	[3.59, 8.72]	40.31	—	—
	CAFT seed 42	1.25	[0.49, 3.17]	57.19	+16.88 ($p<10^{-4}$)	0.004
	CAFT seed 43	2.19	[1.06, 4.45]	43.44	+3.13 (p=0.42)	0.035
	CAFT seed 44	3.75	[2.16, 6.44]	44.38	+4.07 (p=0.30)	0.327
8B (verifier applied to thinking-stripped chains; Section 4.5)						
8B	Base (no FT)	0.00	[0.00, 1.19]	77.81	—	—
	CAFT seed 42	0.00	[0.00, 1.19]	81.25	+3.44 (p=0.177)	—
	SFT	0.00	[0.00, 1.19]	39.69	−38.12 ($p<10^{-23}$)	—
8B (seeds 43/44)						
8B	CAFT seed 43	0.00	[0.00, 1.19]	76.25	−1.56	—
	CAFT seed 44	0.00	[0.00, 1.19]	76.25	−1.56	—

Table 20: Outcome-only DPO at matched budget vs. CAFT (1.5B, in-distribution, sampled decoding, 320 chains). All arms share the base model, DPO hyperparameters, and 1,409-pair optimization budget. p: paired exact McNemar vs. base.

Run	Pairs	CVR (%)	95% CI	Acc (%)	McNemar p (CVR)
Base (no FT)	—	5.62	[3.59, 8.72]	40.31	—
Outcome-only	1,409 (416 uniq.)	4.06	[2.39, 6.83]	44.06	0.458
CAFT (seed 42)	1,409	1.25	[0.49, 3.17]	57.19	0.004

O.5 Outcome-only DPO: matched-budget controls

"Matched budget" means matched optimizer steps, not matched gradient diversity (each unique pair is exposed ∼3.4× per epoch); a matched-uniqueness control behaves consistently, so the conclusion does not hinge on the upsampling. Head-to-head at matched budget, CAFT outperforms the outcome-only baseline on both axes (accuracy 83 vs. 41 discordant pairs, p=0.0002; violations 4 vs. 13, p=0.049 unclustered, p=0.063 clustered).

O.6 Step-matched SFT control

Because the SFT-eligible set is smaller than the DPO pair set, the original SFT runs took fewer optimizer steps at equal epochs (195/219/225 vs. 477/468/219 at 1.5B/8B/32B; Appendix I), so part of SFT's underperformance could in principle reflect under-training. Retraining SFT-1.5B on an upsampled chain set (1,272 samples; 1,144 after the same held-out split as every other trainer) yields 429 optimizer steps—more than double the original 195 and within 10% of DPO's 477—yet the step-matched model behaves like the original SFT run rather than like CAFT: CVR 5.31% (17/320) vs. base 5.62% (14 vs. 13 discordant pairs, p=1.0), accuracy 35.62% vs. 40.31% (p=0.23), and far from CAFT on both axes (violations 4 vs. 17, p=0.007; accuracy 115 vs. 46 CAFT-only-correct, $p=5.3\times10^{-8}$). The dose–response prediction of the under-training hypothesis therefore fails: what separates SFT from CAFT is the contrastive exposure to violating chains, not the number of gradient steps (Appendix L).

Table 21: Scale-up comparison: 1.5B vs. 8B vs. 32B under sampled decoding ($T = 0.8$, 320 chains) on in-distribution scenarios. CAFT reduces CVR and improves accuracy at 1.5B. At 32B, the base model already has near-zero CVR (0.31%); CAFT does not further reduce sampled CVR (0.62%) but improves accuracy (+1.56 pp). Under greedy decoding, all methods achieve 0% CVR at 32B (Table 16). Under the strip-first protocol, base, SFT, and CAFT at 8B all score 0.00% CVR (accuracies 77.8%/39.7%/81.2%); CAA and PVS are reported at the 1.5B and 32B scales.

Scale	Method	CVR (%) ↓	Acc (%) ↑	CAA (%) ↑	PVS ↑
1.5B	Base	5.62	40.31	38.12	0.9925
	SFT	5.62	34.06	31.56	0.9895
	CAFT	1.25	57.19	56.56	0.9992
32B	Base	0.31	86.56	86.56	0.9990
	SFT	0.31	88.44	88.44	0.9999
	CAFT	0.62	88.12	87.81	0.9989

8B is summarized separately: under the strip-first protocol all three methods sit at the violation floor (CVR 0.00% every arm), with accuracies 77.81%/81.25%/39.69% (base/CAFT/SFT) and no significant CAFT-vs.-base difference (Appendix 19); CAA and PVS are reported at the 1.5B and 32B scales.

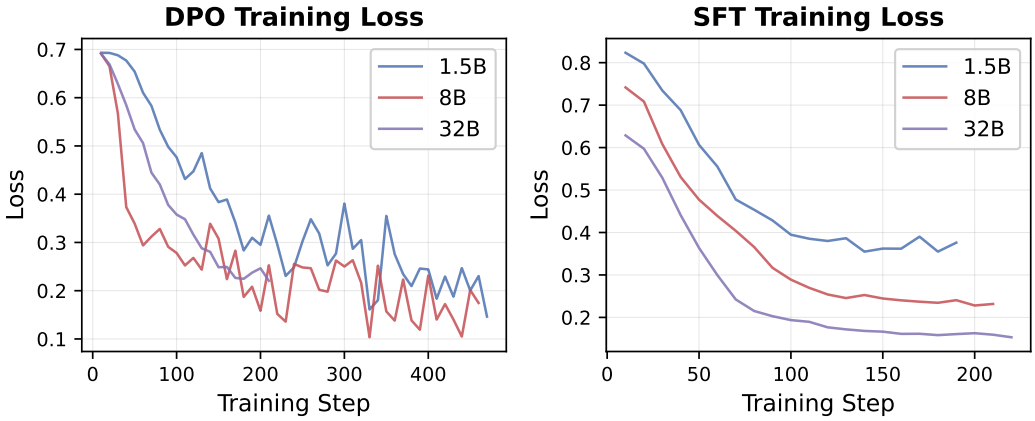

Figure 6: Training loss curves for DPO (left) and SFT (right) at 1.5B, 8B, and 32B scales.

O.7 Two zero-training alternatives

Test-time selection (best-of-4 with verifier ranking, BoV-4) applied to the base model drives CVR to 0% in every cell but lowers accuracy relative to single-sample base at 1.5B (37.5% vs. 40.3%), because selection optimizes conservation compliance rather than answer correctness—partly mechanically, since the verifier's tie-breaking prefers shorter chains. CAFT improves both simultaneously (57.2% vs. 37.5% at 1.5B) and bakes the improvement into the weights, so the two are complementary: BoV-4 suits settings where the verifier is available at inference and accuracy is secondary, CAFT suits single-pass, verifier-free deployment. A prompt-only control—appending an explicit conservation-check directive to the system prompt—does not reduce violations (CVR 6.88%, 22/320, vs. base 5.62%; 15 vs. 19 discordant pairs, p=0.61; nominal direction worse) and degrades accuracy (−4.1 pp, p=0.30); against CAFT it is dominated (22/320 vs. 4/320; p=0.0003). Instruction alone therefore neither suppresses the misconception patterns nor preserves accuracy, indicating that the learned contrast, not task framing, drives the reduction; both controls use one scale (1.5B) and one phrasing, and stronger prompting regimes are untested (Appendix L).

Table 22: Scale summary, sampled decoding (T=0.8, 320 chains, in-distribution). The 8B rows report the strip-first protocol (Section 4.5). Greedy decoding: CVR=0% for all methods at 32B and for SFT at 1.5B.

Scale	Run	CVR (%) ↓	Acc (%) ↑	CAA (%) ↑
1.5B	Base	5.62	40.31	38.12
	CAFT	1.25	57.19	56.56
8B	Base	0.00	77.81	77.81
	CAFT	0.00	81.25	81.25
32B	Base	0.31	86.56	86.56
	CAFT	0.62	88.12	87.81

O.8 The 8B seed runs as a diagnostic

The 8B replicas (seeds 43/44) score 0/320 violations with 76.25% accuracy because they were evaluated under the strip-first prompt, under which no checkpoint emits thinking content.

P Full Limitations

This appendix holds the full, unabridged list of the eleven limitations summarized in Section 5; numbering is identical to the body list.

(1) Run-to-run variance. DPO at β=0.1 with ∼1.4k pairs has nontrivial run-to-run variance. At 1.5B the CVR reduction replicates directionally across seeds (1.25%/2.19%/3.75% vs. base 5.62%) with pooled significance, but per-seed McNemar significance is unstable (seed 44: p=0.33) and the accuracy co-benefit ranges from +3 to +17 pp, so any single-run magnitude is one draw from a wide run distribution; ablation conclusions resting on 1–4 chain differences could similarly reorder. At 8B the seed runs are diagnostic rather than confirmatory: the replicas score 0/320 under the strip-first prompt, under which no checkpoint emits visible thinking content, so their agreement reflects the protocol; the three-seed Fisher combination (p=4.5×10^{-15}) compares across a format divide and is not a replication.

(2) Verifier coverage and self-referential validation. The verifier covers only mechanics, and regex detection can miss subtle numerical violations that a symbolic-computation verifier would catch; the friction pattern also flags hedged-but-correct statements of the form "energy is conserved, but mechanical energy is dissipated" (the single 32B instance in Section 4.5), which a clause-level semantic check would remove. Validation is partly self-referential (Suites 1/3/4 are labeled by the verifier itself; Suite 5's labels partially lean on verifier outputs), and the most independent evidence is a blind multi-referee audit (Appendix H; inter-referee κ = 0.62–0.76 across four raters) in which precision against auditor consensus is 67.9% (69.6% under four-rater consensus), bounded below by the disclosed hedge-pattern false positives, with the in-distribution CAFT effect surviving on verifier-clean control chains (p=0.026).

(3) Scale and model-family coverage. Experiments stop at 32B, and the scale axis co-varies with model family (Qwen2.5 at 1.5B vs. Qwen3 at 8B/32B); our Qwen3-1.7B control (zero CVR in both splits, Section 4.5) shows the cross-scale trend is not a pure scale effect, and a same-family sweep with CAFT training at each scale is a natural next step to separate the two.

(4) Scenario complexity and OOD scope. Scenarios are closed-form, limiting complexity relative to open-ended physics, and OOD scenarios vary numerical parameters within the same templates—re-instantiation rather than structural generalization.

(5) Sample size, clustering, and multiplicity. Evaluation uses 80 scenarios (320 chains) per split; at 32B this yields single-digit violation counts, so CVR differences at that scale are unresolvable by design, and violations cluster within scenarios, making the effective sample size smaller than the chain count suggests. We therefore supplement every headline McNemar test with a scenario-clustered sign-flip randomization test (Appendix C); the 1.5B

CVR and accuracy effects survive clustering, the CAFT-vs.-outcome-only CVR contrast weakens to marginal (p=0.063), and the 8B comparisons are moot under the strip-first protocol. We report ∼30 paired tests without a global multiplicity correction; within the second-constraint family no test survives Holm correction (smallest family-wise adjusted p=0.083, from a raw p=0.0138 at 32B in-distribution over 6 tests), and we frame those results as exploratory.

(6) The 8B cell under the strip-first protocol. Under the strip-first protocol 8B shows no measurable CVR or accuracy difference (base 0/320, CAFT 0/320 in-distribution and 1/320 OOD, a disclosed false positive of the momentum external-force pattern; accuracy 77.8% vs. 81.2%, p=0.177), so we claim neither an 8B benefit nor an 8B cost; no 8B difference—in either decoding mode or either split—is significant under that protocol, and we claim no 8B cost on any split. Inelastic collision remains the hardest category at 1.5B.

(7) Ceiling at 32B. CAFT does not reduce CVR (0.31% vs. 0.62%, p=1.0); harder scenarios are needed to test CAFT near the violation floor.

(8) Checker coverage and the armed-but-silent mass check. The guarded arithmetic checker validates only explicitly stated single-operation equalities, so slips in unstated mental arithmetic or multi-step derivations are invisible to it and the near-zero measured rates are lower bounds; the dimensional-homogeneity checker likewise skips lexically ambiguous left-hand sides, so its coverage is conservative. Relatedly, the mass-conservation check is armed in all 320 in-distribution scenarios and fires correctly on synthetic mass-violation probes, but never fires on any evaluated chain—no model at any scale verbalizes mass creation or destruction—so its silence is a floor effect of closed-form mechanics rather than a disabled check.

(9) Quantization at 32B. 32B uses QLoRA 4-bit quantization, which may mask CAFT's effect.

(10) Baseline, seed, and ablation coverage. The outcome-only baseline is run at 1.5B, the scale with the clearest effect, so 8B/32B process-vs.-outcome attribution rests on the wrong-answer-only ablation, an indirect analogue with unmatched pair budgets, and extending the matched-budget baseline to the larger scales is a natural next step; seed replications cover 1.5B and 8B but not 32B, and the 8B OOD seed evaluations use a different scenario-generation seed than the headline OOD runs, so a paired comparison there is left to future work. The leave-one-strategy-out ablation is run at a single seed and 1.5B with small per-arm discordant counts (4–16 events), so its division-of-labor attribution is read as directional; the same-answer-process and wrong-answer removals differ from full CAFT by 4–6 discordant chains.

(11) Strength and scope of verifier validation. Our strongest checks on the verifier beyond the six suites are the numeric-recomputation cross-check (on the 21 flagged chains whose quantities admit full balance recomputation, the recomputed verdicts agree with the pattern verifier after manual adjudication; Appendix G) and the blind multi-referee audit (substantial inter-referee agreement, $\kappa = 0.62$–0.76, median 0.72; precision 67.9% against three-rater consensus, 69.6% against four-rater; recall 38.8–40.0%; the in-distribution effect preserved on verifier-clean control chains at Fisher p=0.026 three-rater, p=0.048 four-rater). The referees are LLM-based rather than human experts and the audit sample is small (110 chains), so this complements—but does not replace—a larger human expert audit.

Q Analysis Details

Process-vs.-outcome probes (full detail). Same-answer-process pairs (Strategy C) show that conservation-law satisfaction provides a process-level signal independent of answer correctness, and the matched-budget outcome-only baseline (Section 4.6, Table 20) closes the comparison empirically: at an identical 1,409-pair budget, removing the process signal eliminates the violation reduction (4.06% vs. CAFT's 1.25%) and shrinks the accuracy gain to the generic-preference-optimization level (+3.75 pp, p=0.32 vs. base). The accuracy-side separation is robust to scenario clustering, while the CVR-side separation is directionally consistent across three independent probes (no CVR reduction from outcome-only training

at matched budget; head-to-head discordance 13 vs. 4 in CAFT's favor, p=0.049 unclustered and p=0.063 clustered; and the natural-budget variant behaving consistently) but does not individually clear the clustered significance bar—so we present the process-signal conclusion as supported by converging evidence rather than by any single test.

Violation-type and convergence details. CAFT's CVR reductions concentrate where violations are most common (friction dissipation 22.5%→2.5%, inelastic collision 17.5%→2.5%); violation-free categories stay violation-free. By violation type, CAFT cuts energy violations from 4.06% to 0.31% and wrong-law-application from 1.25% to 0.94%, and eliminates momentum violations entirely (Appendix J); no chain at any scale verbalizes a mass non-conservation, so the armed mass check never fires (Limitations (8)). CAFT also attains the highest Process Validity Score (0.9992 vs. 0.9925 base, 0.9895 SFT), improving reasoning-step quality beyond explicit violations; PVS is computed by the same conservation verifier as CVR (Appendix A); it is a verifier-derived secondary metric, not an independent quality measure, and inherits the verifier's coverage limits. Conservation-based pairs (synthetic, same-answer-process) are also easier to learn (perfect training accuracy) than answer-based wrong-answer pairs (64.3%), supporting the hypothesis that conservation-law verification provides a cleaner training signal. DPO training converges smoothly from the initial ∼0.69 to 0.15 (1.5B), 0.17 (8B), and 0.22 (32B), indicating that the model learns to discriminate conservation-respecting from violating chains; terminal losses are not comparable across methods or scales, since SFT minimizes a different objective (Appendix J, Figure 6).

Downstream accuracy decomposition. To address whether CVR gains come at the expense of answer correctness, we decompose the 1.5B in-distribution chain-level transitions from base to CAFT. Of 320 aligned chains, 90 flip from incorrect to correct answers while 36 flip from correct to incorrect (net +16.9 pp, $p<10^{-4}$); conservation status improves or is retained in 316 of 320 chains (18 improve from violating to clean, 298 remain clean, 0 remain violating; the 4 deteriorations are the CVR discordant events of Section 4.2). The 36 accuracy regressions are spread across all eight categories (2–8 chains each) and are outweighed by gains in every category (e.g., projectile energy 4 lost vs. 15 gained; friction dissipation 4 lost vs. 14 gained). CAFT therefore does not trade conservation for accuracy at 1.5B, and the validated untargeted probes below show no measurable shift in other checkable facets of reasoning quality.

Second-constraint checker design and audit. Conservation laws are one instance of a broader family of programmatically checkable reasoning constraints. To probe whether CAFT's effect is specific to its training constraint or reshapes reasoning more broadly, we score all evaluation chains against a second, orthogonal family: numerical arithmetic consistency (each stated binary equality $a \oplus b = c$ inside math delimiters verified against exact arithmetic, relative tolerance 10^{-4}). The full checker description and audit are given in Appendix E. The checker applies digit boundaries, single-operator left-hand sides, backward scanning, and symbolic-right-hand-side and rounding exclusions, flagging at most 0.62% of chains in every cell, with no system pair differing significantly anywhere (all paired exact McNemar $p \geq 0.5$; Appendix E). Manual audit of all five residual flags across the six scale×split cells finds three genuine slips, one sign-shorthand transcription, and one deliberate reductio the model itself labels impossible; the residual flags are disjoint from conservation violations (0 of 5,760 evaluation chains are flagged by both checkers). A third, independently audited probe on dimensional homogeneity shows no systematic difference between systems either (in-distribution 8.1%/15.3%/0.3% base vs. 5.9%/16.6%/0.6% CAFT at 1.5B/8B/32B; all paired $p \geq 0.13$), and the converse training experiment (training on dimensional-consistency pairs) reduces conservation violations no better than chance (p=0.69), so neither method's behavioural footprint transfers to the other constraint family (Appendix D). The methodological lesson outlives the null result: an unvalidated probe can manufacture a plausible, statistically significant, mechanistically narratable "side effect"—ours did—so untargeted-constraint probes should be held to the same validation standard as the primary metric before their outputs are interpreted.

R Main-Results Figure

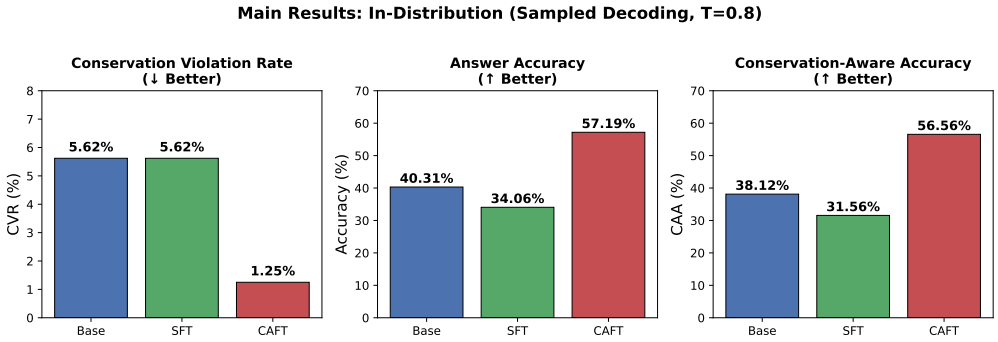

Figure 7: Main results under sampled decoding ($T = 0.8$). CAFT achieves the lowest CVR and highest accuracy and CAA among all methods under this decoding regime.