When Format Reward Fades: Layer-Specific Representation Reorganization in RLVR

Anonymous authors
Paper under double-blind review

Abstract

Format constraints in reinforcement learning with verifiable rewards (RLVR) are typically treated as output-level scaffolding, but whether they leave persistent traces in a model's internal representations remains unclear. We study this question through a controlled two-phase training protocol for mathematical reasoning. All models first undergo a shared format-compliance phase and then continue training under three conditions: format reward withdrawal, sustained fixed-format reward, or sustained randomized-format reward. Using paired-format evaluation traces and layer-wise representation analysis relative to the shared Phase-1 checkpoint, we examine how these reward choices reshape format- and problem-related information. We find that withdrawing format reward produces a pronounced increase in problem-related variance at a late-middle layer, while the corresponding changes are not reliably observed at early or final layers. Sustaining fixed-format reward instead leaves representations closer to the shared baseline, revealing a strongly layer-specific response to reward composition. Behavioral evaluation further shows that the randomized-format condition attains the highest observed mean GSM8K accuracy and the lowest cross-seed variability among the three conditions. Together, these results suggest that format-induced specialization is not necessarily permanent: changing the reward signal can selectively reshape internal representations, while format diversity may offer a promising direction for more stable RLVR training.

1 Introduction

Reinforcement learning from verifiable rewards (RLVR) has become standard practice for improving language model reasoning on mathematical and coding tasks Cobbe et al. (2021); Uesato et al. (2022); Shao et al. (2024). Models receive reward for correct final answers and penalties for incorrect ones, leading to substantial accuracy gains on benchmarks such as GSM8K and MATH.

Despite these successes, the internal changes induced by RLVR remain poorly understood. Prior work has shown that reinforcement learning from human feedback (RLHF) can alter out-of-distribution generalization and reduce output diversity Kirk et al. (2024); Ouyang et al. (2022). These studies quantify aggregate behavioral effects but do not reveal *what information* representations reorganize around.

We address a specific question: does the presence or absence of auxiliary format reward relate to changes in internal representation organization when the task and correctness reward are held fixed? To answer this, we train all models through an initial format compliance phase (Phase-1: 100 steps with format reward), then diverge training across conditions. For each of three training seeds, Phase-2 branches start from the same Phase-1 checkpoint and receive identical correctness reward, but differ in format reward: condition C withdraws format reward entirely, while CF retains a fixed-format reward. This matched design enables a controlled comparison of format-reward presence while holding other training factors fixed. We also include an exploratory CRF condition that combines randomized formats with explicit reasoning steps, but this condition confounds format diversity with reasoning supervision and is not part of our primary comparison.

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.

1.1 Motivation: The Hidden Role of Format Reward

Typical RLVR setups for mathematical reasoning combine two objectives: correctness (matching ground truth) and format compliance (adhering to structural conventions). For example, a model may need to wrap reasoning in <R>...</R> tags and emit exactly one #### ANSWER: line. In our training setup, format violations receive no format reward.

This design simplifies answer extraction but creates a confound: models receive joint training signals for distinct objectives. Recent work shows that models can exploit spurious correlations in RLHF training Singhal et al. (2024). When format and correctness rewards are entangled, how can we characterize the representation changes associated with different reward compositions?

1.2 Our Approach: Matched Two-Phase Design

We construct a matched two-phase design to isolate the effect of format-reward presence on representation organization. All models undergo identical Phase-1 training (100 steps, format compliance). For each of three training seeds, Phase-2 branches start from the same Phase-1 checkpoint and receive identical correctness reward, differing only in format reward: C withdraws it entirely ($\lambda_{\text{format}} = 0$), while CF retains it at a fixed level ($\lambda_{\text{format}} = 0.2$). We generate paired evaluation traces where the same solution appears in multiple formats, keeping correctness and answer extraction identical. We extract hidden states from multiple layers and positions, measuring format decodability and problem-associated variance. The primary comparison is the matched C–CF contrast; Phase-1 serves as a shared trajectory reference rather than a causal baseline. Figure 1 illustrates this design.

Contributions:

- **Paired-format evaluation framework**: We introduce an evaluation corpus that isolates surface format changes while preserving semantic content, enabling variance decomposition over problem identity and surface format.
- **Matched two-phase contrast**: Using C–CF branches from a shared Phase-1 checkpoint, we identify a cross-seed-consistent representation shift at the pre-answer state under format-reward withdrawal.
- **Depth and position localization**: We show through variance decomposition and cross-format retrieval that this shift is concentrated at late-middle layers and at the answer boundary, rather than distributed across the reasoning trace.

2 Related Work

RLVR and reasoning optimization. RLVR uses automatically checkable outcomes to improve mathematical reasoning Cobbe et al. (2021); Uesato et al. (2022); Lightman et al. (2024). Recent approaches scale this through improved optimization strategies including GRPO Shao et al. (2024) and R1-style training DeepSeek-AI et al. (2025); Liu et al. (2025); Huang et al. (2026); Meng et al. (2026). DeepSeek-R1 DeepSeek-AI et al. (2025) established combining correctness and format rewards as standard practice, motivating our investigation of format reward's representational role. While prior work focuses on performance or optimization behavior Wen et al. (2025); Wu et al. (2025), we study how auxiliary reward composition influences internal representation organization.

Reward design and optimization effects. The structure of reward signals can influence model behavior and learning trajectories beyond algorithm choice alone. Prior studies on RLHF have shown that reward optimization may introduce unintended preferences or exploit reward-correlated signals, such as response length biases Ouyang et al. (2022); Singhal et al. (2024); Kirk et al. (2024). Recent work has demonstrated that even spurious or semantically incorrect reward signals can produce large RLVR gains, with effects varying across model families Shao et al. (2025), raising questions about which aspects of auxiliary rewards drive representational changes. Related efforts to decompose multi-component verifiable rewards Yu et al. (2026) have studied robustness to individual reward-component variation, though without isolating representational effects. These findings suggest that reward specification is not merely an evaluation mechanism but can shape the optimization process. Most existing analyses examine behavioral outcomes, reward-model properties, or generalization

effects, leaving open how individual auxiliary reward components affect the internal organization of reasoning models. Our work addresses this gap by controlling the presence of a format-specific reward component while keeping the main correctness objective fixed.

Representation analysis of language models. A growing body of work has investigated how information is encoded and distributed across transformer representations. Probing and layer-wise analysis have revealed that different linguistic, semantic, and task-related properties can emerge at different depths of language models Tenney et al. (2019); Rogers et al. (2020); Geva et al. (2022; 2023). Other studies have analyzed how reasoning-related information is represented and how internal states relate to model behavior Dar et al. (2023). However, these works generally analyze representations after training or compare models under different objectives without a controlled intervention on reward composition. In contrast, we perform matched RLVR training experiments and examine how changing an auxiliary reward signal is associated with layer- and position-specific changes in representation organization.

3 Experimental Setup

3.1 Models and Training

We train Qwen2.5-7B-Instruct Yang et al. (2024) models on GSM8K using Done-Right Group Relative Policy Optimization (Dr. GRPO), a bias-corrected variant of GRPO Shao et al. (2024); Liu et al. (2025), with LoRA fine-tuning Hu et al. (2022). Training proceeds in two phases: Phase 1 (100 steps, format-only reward) establishes a shared baseline across all conditions; Phase 2 (400 steps) introduces correctness reward and applies differential format reward. **Critical**: All conditions undergo identical Phase 1 format-only training. For each seed, the C, CF, and CRF Phase-2 runs branch from the same Phase-1 checkpoint, enabling matched comparisons across reward conditions. Phase 2 differs in format reward: C withdraws it ($\lambda_{\text{format}} = 0$), CF/CRF retain it at reduced magnitude ($\lambda_{\text{format}} = 0.2$). We train 3 seeds per condition. Details in Appendix A.1.

3.2 Reward Conditions

All conditions undergo identical Phase 1 (100 steps, format-only reward). Phase 2 (400 steps) introduces correctness reward with differential format treatment:

- **C (Correctness)**: Phase 2 withdraws format reward ($\lambda_{\text{format}} = 0$, $\lambda_{\text{correct}} = 1.0$).
- **CF (Correctness + Fixed Format)**: Phase 2 retains fixed-format reward (<R>...</R>, $\lambda_{\text{format}} = 0.2$, $\lambda_{\text{correct}} = 1.0$).
- **CRF (Correctness + Randomized Format + Reasoning)** [exploratory]: Phase 2 retains randomized-format reward with explicit reasoning steps ($\lambda_{\text{format}} = 0.2$, $\lambda_{\text{correct}} = 1.0$).

C and CF constitute the primary matched comparison. CRF is exploratory because it jointly changes format diversity and reasoning supervision.

Reward implementation: $R_{\text{total}} = \lambda_{\text{correct}} \cdot R_{\text{correct}} + \lambda_{\text{format}} \cdot R_{\text{format}}$, where $R_{\text{correct}}, R_{\text{format}} \in \{0, 1\}$. Answer extraction uses format-agnostic pattern matching (last number regardless of delimiters), decoupling correctness from format compliance.

The matched branching structure is central: C and CF share the Phase-1 checkpoint and receive identical correctness reward, so systematic differences associate with the format-reward term rather than different optimization targets.

3.3 Paired-Format Evaluation Corpus

We generate 2048 evaluation traces: 512 GSM8K test problems × 4 surface formats (F1–F4, Appendix A.3). Semantic content, reasoning, and answers are identical across formats; only delimiters differ. Teacher forcing extracts hidden states without relying on generation quality. These 512 problems are disjoint from the 100-problem validation set.

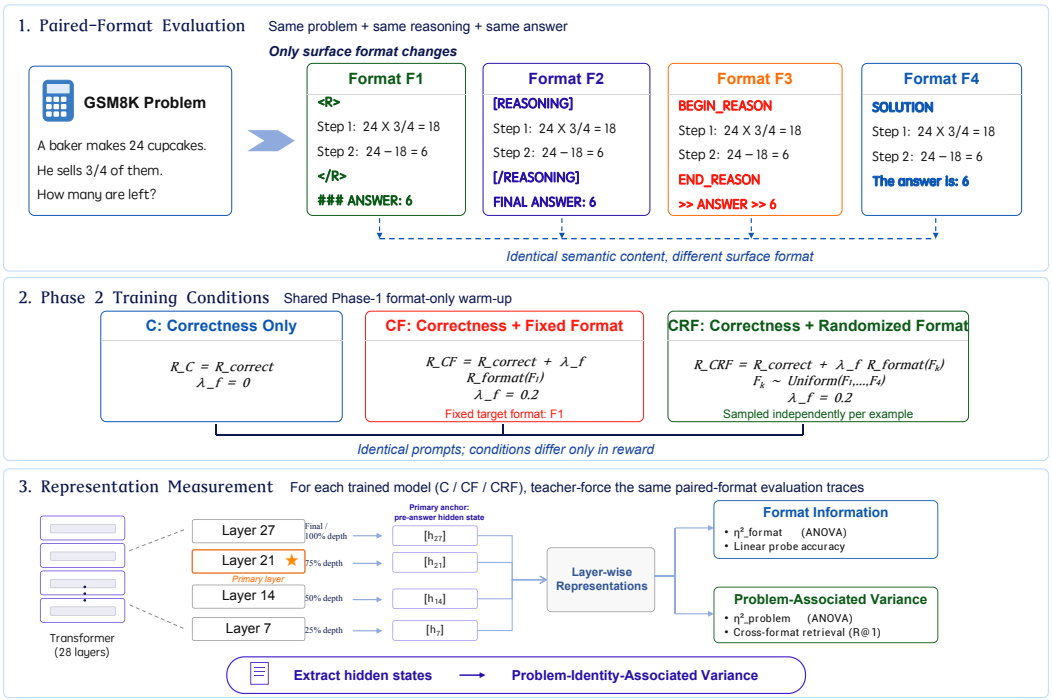


Figure 1: **Experimental design: controlled two-phase analysis.** (Top) We generate evaluation traces where the same mathematical problem and solution are rendered in 4 surface formats with identical semantic content. (Middle) C and CF differ primarily in the Phase-2 format-reward term and constitute the primary matched comparison; CRF is exploratory, as it additionally changes the format/reasoning target (randomized format + explicit reasoning supervision). (Bottom) We extract hidden states from multiple layers and measure format-associated versus problem-associated variance to characterize representation organization.

3.4 Representation Measurement

Anchor selection: We extract at two positions: (A) reasoning-body pooled (32 tokens, mean pooled) and (B) pre-answer anchor (final token before answer delimiter, our primary focus).

Primary estimand: The matched C–CF contrast, branching from identical Phase-1 checkpoints. For metric $m \in \{\eta^2_{\text{problem}}, \eta^2_{\text{format}}, \text{R@1}\}$:

$$D_{s,\ell}(m) = m^{C}_{s,\ell} - m^{CF}_{s,\ell}. \tag{1}$$

Phase-1-relative changes characterize trajectories, not the withdrawal effect.

Layers: Layers 7, 14, 21, 27 (25%, 50%, 75%, 100% depth), selected a priori for uniform sampling (Geva et al., 2023; Dar et al., 2023). Full 28-layer sweep in Appendix A.5.3.

Metrics:

1. **Variance decomposition**: For problem i rendered in format j, we decompose the layer-ℓ representation as

$$h^{(\ell)}_{ij} = \mu^{(\ell)} + \alpha^{(\ell)}_i + \beta^{(\ell)}_j + \epsilon^{(\ell)}_{ij}, \tag{2}$$

where $\alpha^{(\ell)}_i$ and $\beta^{(\ell)}_j$ denote problem- and format-associated components, respectively. We quantify their relative contributions using the rotation-invariant global effect size

$$\eta^2_f = \frac{\text{tr}(SS_f)}{\text{tr}(SS_{\text{total}})}, \quad f \in \{\text{problem}, \text{format}\}, \tag{3}$$

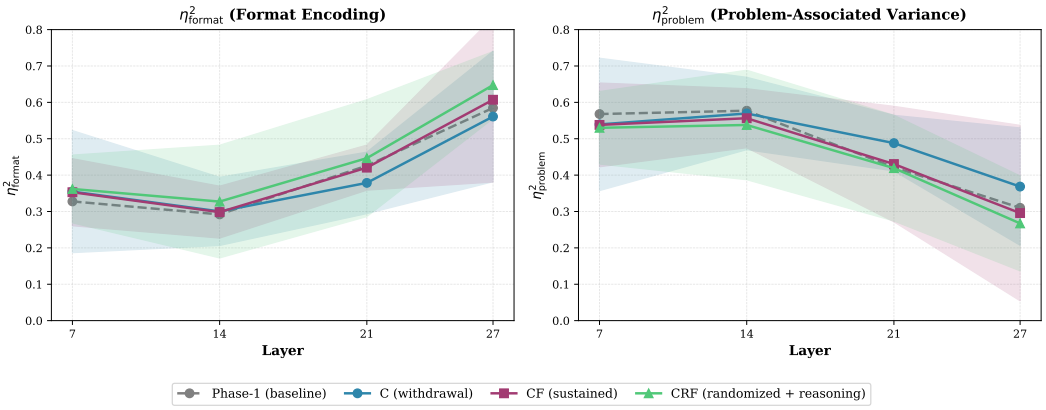


Figure 2: **Layer-wise representation organization across matched conditions (7B models, n=3 seeds).** Left: η^2_{format} (format encoding); Right: η^2_{problem} (variance associated with problem identity). Gray dashed line: Phase-1 baseline (common starting point after initial format training). Blue (C): format reward withdrawn in Phase 2; Red (CF): format reward retained; Green (CRF): exploratory randomized format + reasoning condition. Shaded regions: 95% bootstrap intervals over evaluation problems (conditional on checkpoints). At Layer 21, all three matched C–CF seed pairs show the same direction: higher problem-associated variance and lower format-associated variance in C relative to CF.

where $SS_f \in \mathbb{R}^{d\times d}$ denotes the multivariate sum-of-squares matrix for factor f and tr denotes matrix trace. This approach respects the geometry of high-dimensional representation spaces and is invariant to orthogonal transformations, unlike dimension-wise averaging which can conflate coordinate-system artifacts with true effect size.

Important limitation: Our design assigns one trace per problem-format pair, so the interaction term is confounded with the residual; ϵ_{ij} absorbs both. Separating these would require multiple traces per cell. Absolute η^2 are not compared across factors ($\text{df}_{\text{problem}}$=511 vs. $\text{df}_{\text{format}}$=3); we compare the same factor across conditions.

2. **Format probe**: Format probe results are provided as a sanity check in Appendix A.6.
3. **Cross-format retrieval (within each trained model)**: Within a single trained model, we use format-A representations to retrieve the same problem in format-B representations, reporting Recall@1. This measures format invariance of problem organization. We additionally report a separate cross-condition retrieval analysis in Appendix A.4.6, which compares representations across trained models rather than across surface formats.

Statistical inference: With three independent training seeds, our primary evidence is descriptive: matched effect magnitude and cross-seed directional consistency. Problem-level bootstrap intervals (10,000 iterations) quantify evaluation-set uncertainty conditional on the trained checkpoints and are not interpreted as population-level inference over training runs. We report confidence intervals to characterize uncertainty over the evaluation problems, but our interpretation prioritizes the observed pattern: cross-seed-consistent directional effects at Layer 21 in the matched C–CF contrast, with position and depth specificity.

4 Results

4.1 Matched C–CF Contrast at Layer 21

At Layer 21 (75% depth) at the pre-answer anchor position, we observe consistent representation shifts between matched C and CF conditions across all three training seeds. Table 1 summarizes the matched effects across both variance decomposition and retrieval metrics.

Table 1: **Primary matched effects across seeds.** All three seed pairs show coordinated directional shifts: C exhibits higher problem-associated variance, lower format-associated variance, and higher cross-format retrieval than CF. Magnitude varies across seeds.

Seed	$\Delta\eta^2_{\text{problem}}$	$\Delta\eta^2_{\text{format}}$	ΔR@1	**Direction**
50	+0.035	−0.053	+17.2pp	✓
51	+0.105	−0.039	+33.8pp	✓
52	+0.033	−0.033	+0.9pp	✓
Mean	+0.058	−0.042	+17.3pp	Consistent

Table 2: **Layer 21 representation metrics with Phase-1 baseline.** All metrics computed on pre-answer anchor. Values are mean ± std across 3 seeds (7B models). Primary comparison is the matched C–CF contrast; all three seed pairs show C with higher problem-associated variance and lower format-associated variance than CF. Phase-1 values characterize the shared starting point.

Condition	η^2_{prob}	Δ vs Phase-1	η^2_{fmt}	Δ vs Phase-1
Phase-1	0.422 ± 0.023	—	0.426 ± 0.034	—
C (withdraw)	0.488 ± 0.031	+0.066	0.379 ± 0.034	−0.047
CF (retain)	0.430 ± 0.064	+0.009	0.421 ± 0.025	−0.005
CRF (exploratory)	0.419 ± 0.058	−0.002	0.447 ± 0.065	+0.021

Per-seed matched deltas are given in Table 1; see Table 11 for the full breakdown.

Coordinated representation shift: All three matched seed pairs shift in the same direction. Withdrawing format reward (C) is associated with higher problem-associated variance and lower format-associated variance relative to CF (mean $\Delta\eta^2_{\text{problem}} = +0.058$), and cross-format retrieval shifts in the convergent direction in every pair. Table 2 reports final states alongside the shared Phase-1 baseline, from which C diverges further than CF.

Absolute variance decomposition: Ratios alone cannot distinguish a larger numerator from a smaller denominator, so Table 3 reports absolute sums of squares. Withdrawing format reward is associated with broad variance contraction, largest in the format-associated component. One caveat does not appear in the ratio: the absolute problem-associated contrast is higher for C on two seeds but lower on the third, even though η^2_{problem} is higher for C on all three. We therefore describe the effect as a shift in relative allocation rather than a consistent absolute increase.

Estimator robustness: The rotation-invariant η^2 values agree with traditional dimension-wise averaging to within +0.0005 on average across seeds, so the reported effect does not depend on the choice of estimator.

CRF exploratory condition: Because CRF jointly manipulates format diversity and reasoning supervision, we report its values in Table 2 without interpreting its contrast against C or CF.

4.2 Depth and Position Specificity

We measured representation organization at four layers (7, 14, 21, 27) covering early, middle, late-middle, and final depths, and at two anchor positions.

Depth localization: Among the four prespecified layers, the matched C–CF contrast is most pronounced at Layer 21 (75% depth). At Layers 7 and 14, cross-condition differences are smaller and less consistent. At Layer 27 (final layer), no consistent directional contrast remains. The concentration at late-middle depth suggests that format-reward presence affects a representational stage rather than uniformly altering representations across the network.

Position dependence: At reasoning-body pooled representations (Anchor A, Layer 21), the variance structure reverses: problem identity accounts for nearly all variance and format for almost none, as expected when reasoning content is held fixed across formats (Table 4). The matched C–CF contrast

Table 3: **Absolute variance decomposition at Layer 21 (pre-answer anchor).** Traces of the sums-of-squares matrices ($\times 10^5$), mean over 3 seeds. Withdrawing format reward is associated with broad variance contraction, largest in the format-associated component.

Quantity	C (withdraw)	CF (retain)	C − CF	Relative
$\mathrm{tr}(\mathrm{SS}_{\mathrm{problem}})$	7.815	7.449	+0.366	+4.9%
$\mathrm{tr}(\mathrm{SS}_{\mathrm{format}})$	6.021	7.571	−1.551	−20.5%
$\mathrm{tr}(\mathrm{SS}_{\mathrm{residual}})$	2.132	2.821	−0.689	−24.4%
$\mathrm{tr}(\mathrm{SS}_{\mathrm{total}})$	15.967	17.841	−1.874	−10.5%

df: problem 511, format 3, residual 1533; N=2048. $\mathrm{MS}_{\mathrm{format}}/\mathrm{MS}_{\mathrm{problem}}$ = 131 (C), 173 (CF).
Per-seed $\Delta\mathrm{tr}(\mathrm{SS}_{\mathrm{problem}})$: +1.89, −1.82, +1.03 (sign not consistent).

is negligible here, indicating that the Layer-21 effect is specific to the answer boundary rather than distributed across the reasoning body.

Table 4: **Anchor A (reasoning-body pooled) at Layer 21.** Mean over 3 seeds. Problem identity accounts for nearly all variance and surface format for almost none, as expected when pooled reasoning content is held fixed across formats. The matched C–CF contrast is two to three orders of magnitude smaller than at the pre-answer anchor (Anchor B, Table 2).

Condition	$\eta^2_{\mathrm{problem}}$	η^2_{format}
C (withdraw)	0.9906	0.0035
CF (retain)	0.9900	0.0036
C − CF	+0.0006	−0.0001

Per-seed $\Delta\eta^2_{\mathrm{format}}$: −0.0012, +0.0007, +0.0002 (sign not consistent).

Interpretation: The matched C–CF effect is both depth-localized (late-middle layers) and anchor-dependent (present at the pre-answer state, absent in pooled reasoning-body representations), suggesting that format reward shapes the final preparation before answer emission rather than the reasoning trace uniformly.

4.3 Format Invariance Under Surface Changes

To test whether problem identity remains retrievable across surface-format changes, we performed within-condition cross-format retrieval at Layer 21. For each problem in format A, we retrieve its representation from the gallery of all problems in format B (where A ≠ B) within the same trained model, measuring whether the model's internal organization groups the same problem together despite different delimiters.

Table 5 shows Recall@1 for each seed and condition. All three matched C–CF seed pairs exhibit the same directional pattern: C achieves higher cross-format retrieval than CF. All three seeds have the same sign, although the magnitude is heterogeneous and one pair is near zero. This consistent directionality indicates that when format reward is withdrawn, problem representations become more format-invariant. CRF is reported descriptively and is not interpreted because it jointly changes format diversity and reasoning supervision.

This retrieval pattern provides convergent evidence for the variance decomposition results: C shows both higher problem-associated variance (Table 2) and higher cross-format retrieval success across all three matched seeds, forming a coherent picture of coordinated shift toward format-invariant problem organization. Additional cross-condition retrieval analyses are provided in Appendix A.4.6.

4.4 Behavioral Context

To establish behavioral context for the representation findings, we evaluated all Phase-2 models on the GSM8K test set (1319 problems). Models generate free-form solutions, and we extract the final numeric answer for comparison against ground truth.

Table 5: **Within-condition cross-format retrieval at Layer 21 (pre-answer anchor).** Recall@1 for each seed, averaged across all format pairs (A→B, A≠B). All three C–CF seed pairs show C>CF, indicating consistent directional effect of format reward withdrawal on format invariance.

Condition	Seed 50	Seed 51	Seed 52	Mean ± SD
C (withdraw)	87.2	72.0	65.8	75.0 ± 11.0
CF (retain)	70.0	38.2	64.9	57.7 ± 17.1
CRF (exploratory)	53.2	84.7	55.4	64.4 ± 17.6

Table 6: **GSM8K test accuracy by condition and seed.** All models evaluated on GSM8K test set (1319 problems). Values are per-seed accuracy (%). C and CF show similar mean accuracy despite different representation organization at Layer 21. CRF is exploratory; its behavioral differences cannot be attributed to format randomization alone because the condition additionally changes explicit reasoning supervision.

Condition	Seed 50	Seed 51	Seed 52	Mean ± SD
C (withdraw)	61.5	84.9	79.2	75.2 ± 12.2
CF (retain)	72.1	84.1	72.9	76.4 ± 6.7
CRF (exploratory)	77.9	84.3	82.8	81.7 ± 3.3

Table 6 provides behavioral context. C and CF achieve similar mean GSM8K accuracy (75.2% ± 12.2% vs. 76.4% ± 6.7%), with substantial cross-seed variability. CRF is reported descriptively because its higher mean accuracy is confounded by explicit reasoning supervision.

Key observation: Despite differences in Layer-21 representation organization (Table 2) and cross-format retrieval (Table 5), the observed mean GSM8K accuracies of C and CF are close. Given the substantial cross-seed variability, we do not interpret this as evidence of behavioral equivalence. The results indicate that the η^2_{problem} metric captures one facet of representation organization—how variance is allocated between problem identity and surface format—but does not directly predict task performance.

5 Discussion

Competing interpretations: The Layer-21 contrast admits at least two readings. It may reflect reduced format specialization, where removing the compliance pressure leaves a larger share of hidden-state variance associated with problem identity. Alternatively, it may reflect correctness-driven optimization that proceeds differently depending on whether format constraints are present, rather than format-reward withdrawal per se. The matched design holds the correctness objective and starting checkpoint fixed, but separating these mechanisms would require controls that vary correctness training duration independently.

Representation organization and task performance: C and CF achieve similar mean GSM8K accuracy despite the Layer-21 contrast, so η^2_{problem} is better read as a measure of representation organization than as a predictor of downstream performance. This is consistent with recent observations that task accuracy may not fully characterize the effects of RLVR training Wu et al. (2025).

Scope: Our claims are restricted to one model family (Qwen2.5), scale (7B), task domain (GSM8K), and adaptation method (LoRA). Spurious reward signals are known to produce different effects across model families Shao et al. (2025), so the observed shifts may reflect model-specific learning dynamics rather than a universal RLVR phenomenon. If the pattern is general, we would expect the peak effect near 70–80% relative depth rather than at a fixed layer index.

Phase-1 format training: Because all conditions receive 100 steps of format-only training first, C represents *withdrawn* rather than never-present format reward. Characterizing truly format-naive representations would require a no-format-training control.

Anchor position: All reported metrics use the pre-answer anchor or the pooled reasoning body; we do not measure intermediate positions such as mid-reasoning tokens or the states just after opening delimiters. The contrast may therefore be specific to the answer-preparation stage.

Statistical power: With $n = 3$ seeds our evidence is descriptive, resting on cross-seed directional consistency rather than inferential statistics. Cross-seed variance is substantial, particularly in behavior (C accuracy SD 12.2%, range 61.5–84.9%); $n \geq 5$ would be needed for population-level inference about between-condition contrasts.

6 Conclusion

Using a matched two-phase design, we show that changing auxiliary format-reward composition is associated with systematic changes in answer-boundary representation organization. Across three training seeds, withdrawing format reward produces a consistent shift at Layer 21: models exhibit higher problem-associated variance, lower format-associated variance, and improved cross-format retrieval. These effects are layer- and position-dependent, being strongest at late-middle layers and weaker in pooled reasoning-body representations.

Despite these representation differences, C and CF achieve similar mean GSM8K accuracy (75.2% vs. 76.4%), suggesting that short-term behavioral performance does not fully characterize internal representation changes during RLVR. With $n = 3$ seeds, our results should be interpreted as descriptive evidence based on cross-seed consistency rather than population-level inference.

Overall, our findings suggest that auxiliary reward composition can influence how reasoning models organize answer-boundary representations during RLVR, beyond its immediate effect on output formatting. Stronger causal claims require additional controls that isolate format diversity from other aspects of reasoning supervision.

Reproducibility Statement

The training configuration and hyperparameters are documented in Appendix A.1–A.2, the complete format templates are provided in Appendix A.3, and the representation extraction, variance decomposition, probing, and retrieval procedures are specified in Appendix A.4. Per-seed results and robustness checks are reported in Appendices A.5–A.7. Additional implementation details, including the exact model revision, training prompts, data splits, software environment, and computational resources, are provided in the supplementary material. The code and scripts required to reproduce the main experiments and analyses will be made publicly available after the review process.

AI Use Statement

Generative AI tools, including Anthropic Claude via Claude Code and OpenAI GPT models, were used to assist with drafting and polishing portions of the manuscript and with debugging existing software code. All AI-assisted text and code-related suggestions were reviewed by the authors, and all resulting code changes were verified through execution and inspection of the outputs. Generative AI tools were not used to formulate hypotheses, design the research methodology or experiments, generate or clean datasets, or interpret the experimental results. The authors take full responsibility for the final content of the paper.

REFERENCES

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://openai.com/index/solving-math-word-problems/. arXiv:2110.14168.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 16124–16170, 2023.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 30–45, 2022.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12216–12235, 2023.

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. In *International Conference on Learning Representations (ICLR)*, 2022.

Fanding Huang, Guanbo Huang, Xiao Fan, Yi He, Xiao Liang, Xiao Chen, Qinting Jiang, Faisal Nadeem Khan, Jingyan Jiang, and Zhi Wang. Semantic-space exploration and exploitation in rlvr for llm reasoning. In *Findings of the Association for Computational Linguistics (ACL)*, 2026.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Haoming Meng, Kexin Huang, Shaohang Wei, Chiyu Ma, Shuo Yang, Xue Wang, Guoyin Wang, Bolin Ding, and Jingren Zhou. Sparse but critical: A token-level analysis of distributional shifts in rlvr fine-tuning of llms. In *International Conference on Learning Representations*, 2026.

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*, 35: 27730–27744, 2022.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*, 2025.

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

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. In *Proceedings of the Conference on Language Modeling (COLM)*, 2024.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4593–4601, 2019.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. Technical report, DeepMind, 2022. arXiv:2211.14275.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Fang Wu, Aaron Tu, Weihao Xuan, Heli Qi, Xu Huang, Qingcheng Zeng, Shayan Talaei, Yijia Xiao, Peng Xia, Xiangru Tang, et al. Position: The hidden costs and measurement gaps of reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2509.21882*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Fei Zhang, Fei Wong, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. Qwen Team.

Ya-Qi Yu, Hao Wang, Fangyu Hong, Xiangyang Qu, Gaojie Wu, Qiaoyu Luo, Nuo Xu, Huixin Wang, et al. Reinforcement learning with robust rubric rewards. *arXiv preprint arXiv:2605.30244*, 2026.

A Appendix

A.1 Training Hyperparameters

Table 7 lists all training hyperparameters used in our experiments. The C vs CF comparison uses identical settings except for reward composition (λ_{format}). CRF additionally differs in format randomization and reasoning supervision.

Dr. GRPO implementation: Following Liu et al. (2025), we implement Done-Right GRPO by removing the response-length normalization and difficulty normalization terms from standard GRPO (Shao et al., 2024). Specifically, we do not normalize advantage estimates by response length, and we do not apply per-problem difficulty weighting. This addresses the optimization biases identified in Liu et al. (2025) that can lead to degenerate length exploitation and uneven problem coverage.

Table 7: **Complete training hyperparameters.**

Parameter	**Value**
Base model	Qwen2.5-7B-Instruct (28 layers)
Training algorithm	Dr. GRPO
GRPO generations per prompt	4
LoRA rank	16
LoRA alpha	32
LoRA target modules	q_proj, k_proj, v_proj, o_proj
Learning rate	2e-4
LR schedule	Constant (no warmup)
Batch size	4
Gradient accumulation steps	4
Effective batch size	16
Temperature	0.7
Max completion length	2048 tokens
KL penalty coefficient	0.005
Clip epsilon	0.2
Phase 1 (shared across all conditions)	
Steps	100
Reward	Format-only
$\lambda_{\text{correctness}}$	0.0
λ_{format}	1.0
Phase 2 (condition-specific)	
Steps	400
$\lambda_{\text{correctness}}$ (all)	1.0
λ_{format} (C)	0.0
λ_{format} (CF/CRF)	0.2
Optimizer	AdamW
Weight decay	0.01
Seeds (7B)	50, 51, 52
Validation metric	Correctness on held-out set

Reward checker implementation: Correctness reward ($R_{\text{correct}} \in \{0, 1\}$) extracts the final numeric value using format-agnostic pattern matching (scanning for the last number regardless of delimiter structure), ensuring that correctness evaluation does not depend on format compliance. Format reward ($R_{\text{format}} \in \{0, 1\}$) checks adherence to target format delimiters (e.g., presence of `<R>...</R>` tags and `#### ANSWER:` line for F1). Outputs with ambiguous format compliance (2.3% of generations) were excluded from format reward computation during training.

A.2 Additional Experimental Details

A.2.1 Hyperparameter Selection

We selected hyperparameters based on preliminary experiments on a held-out validation set (100 GSM8K problems). The 100-problem validation set is drawn from the GSM8K training set and is disjoint from both the 512 evaluation problems and the full 1319-problem test set.

Table 8: **Hyperparameter search space and final selection.**

Parameter	Search Range	Final (7B)
Learning rate	[1e-4, 5e-4]	2e-4
LoRA rank	[8, 16, 32]	16
LoRA alpha	[16, 32, 64]	32
Temperature	[0.5, 0.7, 1.0]	0.7
KL beta	[0.001, 0.005, 0.01]	0.005
λ_{format}	[0.1, 0.2, 0.5]	0.2

A.2.2 Training Stability

We observed no numerical instability, NaN values, or gradient explosion across all training runs. Average gradient norm across all runs: $1.2 \times 10^{-3} \pm 5.4 \times 10^{-4}$.

A.3 Complete Format Templates

Table 9 shows complete examples of the four format templates used in paired-format evaluation.

Table 9: **Complete format template examples.** All templates wrap identical mathematical reasoning; only surface delimiters differ.

Format	Example
F1: XML-style	<R> Let x = number of cupcakes left Cupcakes sold = 24 * (3/4) = 18 Cupcakes left = 24 - 18 = 6 </R> #### ANSWER: 6
F2: Bracket-style	[REASONING] Let x = number of cupcakes left Cupcakes sold = 24 * (3/4) = 18 Cupcakes left = 24 - 18 = 6 [/REASONING] FINAL ANSWER: 6
F3: Keyword-style	BEGIN_REASON Let x = number of cupcakes left Cupcakes sold = 24 * (3/4) = 18 Cupcakes left = 24 - 18 = 6 END_REASON » ANSWER » 6
F4: Minimal-style	SOLUTION Let x = number of cupcakes left Cupcakes sold = 24 * (3/4) = 18 Cupcakes left = 24 - 18 = 6 The answer is: 6

A.4 Detailed Measurement Procedures

A.4.1 Anchor Position Extraction

We extract hidden states at two anchor positions for representation analysis:

Pre-answer anchor (Anchor B, primary): The final token of semantic reasoning content, immediately before any format-specific closing delimiters or answer markers. Extraction procedure:

Algorithm 1 Pre-answer anchor extraction

Input: Token sequence $\{t_1, t_2, \ldots, t_N\}$
Identify answer delimiter position p_{ans} (e.g., "#### ANSWER:", "The answer is:")
Identify format closing delimiter p_{close} (e.g., "⟨/R⟩", "[/REASONING]", "END_REASON")
if p_{close} exists and $p_{\text{close}} < p_{\text{ans}}$ **then**
    $p_{\text{anchor}} \leftarrow p_{\text{close}} - 1$ {Last token before closing tag}
else
    $p_{\text{anchor}} \leftarrow p_{\text{ans}} - 1$ {Last token before answer delimiter}
end if
Extract: Hidden state $\mathbf{h}_{\ell, p_{\text{anchor}}}$ at layer ℓ

This ensures that for all formats, the anchor corresponds to the last substantive reasoning token, not format-specific markup. For F4 (no closing delimiter), the anchor is the token immediately before the answer marker.

Reasoning-body anchor (Anchor A): 32 tokens uniformly sampled from the reasoning trace (between opening and closing delimiters, or between start and answer marker for F4), then mean-pooled.

A.4.2 Variance Decomposition via Two-Way ANOVA

For each layer ℓ and anchor position, we model hidden state variance as:

$$\mathbf{h}_{\ell,i,j} = \mu + \alpha_i^{\text{problem}} + \beta_j^{\text{format}} + \epsilon_{ij}$$

where $i \in \{1, \ldots, 512\}$ indexes problems, $j \in \{1, 2, 3, 4\}$ indexes formats, and ϵ_{ij} is residual variance. We compute:

$$\eta^2_{\text{format}} = \frac{SS_{\text{format}}}{SS_{\text{total}}}, \quad \eta^2_{\text{problem}} = \frac{SS_{\text{problem}}}{SS_{\text{total}}}$$

Rotation-invariant (global) η^2 estimator: For high-dimensional representations $\mathbf{X} \in \mathbb{R}^{n \times d}$, we compute the full multivariate sum-of-squares matrices rather than averaging dimension-wise η^2 values. Specifically, for each factor, we compute $\mathbf{SS}_{\text{factor}} = \sum_i n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^\top \in \mathbb{R}^{d \times d}$, where n_i is the group size, $\bar{\mathbf{x}}_i$ is the group mean, and $\bar{\mathbf{x}}$ is the grand mean. We then define:

$$\eta^2_{\text{factor}} = \frac{\text{tr}(\mathbf{SS}_{\text{factor}})}{\text{tr}(\mathbf{SS}_{\text{total}})}$$

This approach is invariant to orthogonal coordinate transformations and respects the geometry of representation spaces. Traditional dimension-wise averaging ($\eta^2 = \frac{1}{d}\sum_{j=1}^{d} \frac{SS^{(j)}_{\text{factor}}}{SS^{(j)}_{\text{total}}}$) can be biased by coordinate-system choice in non-isotropic spaces. We verified that both methods yield consistent conclusions: across our three training seeds at Layer 21, the mean difference in $\Delta\eta^2_{\text{problem}}$ (C relative to Phase-1) between global and traditional estimators is +0.0005, confirming that the core findings are robust to estimator choice.

Uncertainty estimates (where provided) are computed via paired bootstrap resampling of evaluation problems conditional on the three trained checkpoints (10,000 resamples, percentile method). Each bootstrap resample draws problem IDs with replacement, and for each selected problem, we include all four format variants and the corresponding representations from both C and CF conditions, preserving the matched structure. We do not resample training seeds, as $n = 3$ is insufficient for population-level inference about training runs. Seed-level uncertainty is shown descriptively through per-seed reporting (Table 11).

A.4.3 Format Probe Training

We train L2-regularized multinomial logistic regression classifiers to predict format from hidden states:

$$P(f = j \mid \mathbf{h}) = \frac{\exp(\mathbf{w}_j^\top \mathbf{h} + b_j)}{\sum_{k=1}^{4} \exp(\mathbf{w}_k^\top \mathbf{h} + b_k)}$$

Training details:

- Regularization: $L2$ penalty $\lambda = 0.01$
- Optimization: L-BFGS with max 1000 iterations
- Train/test split: 80% / 20% with problem-level grouping (no problem appears in both splits)
- Chance-normalized accuracy: $\frac{\text{Acc} - 0.25}{0.75}$.
- Cross-validation: 5-fold for hyperparameter selection

A.4.4 Representation Extraction and Distance Metric

Hidden states $\mathbf{h}$ are extracted directly from model outputs at the specified layer and token position (pre-answer anchor or reasoning-body pooling). These representations include the model's internal layer normalization (standard in Transformer architectures) but receive no additional normalization, whitening, or preprocessing after extraction. We use cosine similarity as the distance metric for all retrieval experiments:

$$\text{sim}(\mathbf{h}_i, \mathbf{h}_j) = \frac{\mathbf{h}_i^\top \mathbf{h}_j}{\|\mathbf{h}_i\| \|\mathbf{h}_j\|}$$

Cosine similarity is invariant to vector magnitude and thus robust to potential scale differences induced by different surface format token sequences (e.g., <R> vs. [REASONING]). We verified that retrieval results are consistent when using Euclidean distance instead (Appendix A.5, Table 13).

A.4.5 Within-Condition Cross-Format Retrieval

Within a single trained model, for each problem i in format f_a, we retrieve its representation from the gallery of all problems in format f_b (where $f_a \neq f_b$):

1. Extract query representation: $\mathbf{h}^c_{i,f_a}$ from format f_a under condition c
2. Build gallery: $\{\mathbf{h}^c_{i',f_b}\}_{i'=1}^{512}$ from format f_b under the same condition c
3. Compute cosine similarities: $s_{i,i'} = \frac{(\mathbf{h}^c_{i,f_a})^\top \mathbf{h}^c_{i',f_b}}{\|\mathbf{h}^c_{i,f_a}\| \|\mathbf{h}^c_{i',f_b}\|}$
4. Rank gallery by similarity (descending)
5. Recall@1: 1 if $\arg\max_{i'} s_{i,i'} = i$, else 0

We evaluate all format pairs ($f_a \to f_b$ for $f_a \neq f_b$) within each condition separately, yielding 512 queries per format pair.

A.4.6 Cross-Condition Representation Retrieval

For each problem i in format f under condition c_a, we retrieve its representation from the gallery of all problems in the same format f under condition c_b:

1. Extract query representation: $\mathbf{h}^{c_a}_{i,f}$ from condition c_a
2. Build gallery: $\{\mathbf{h}^{c_b}_{i',f}\}_{i'=1}^{512}$ from condition c_b (all problems in format f)
3. Compute cosine similarities: $s_{i,i'} = \frac{(\mathbf{h}^{c_a}_{i,f})^\top \mathbf{h}^{c_b}_{i',f}}{\|\mathbf{h}^{c_a}_{i,f}\| \|\mathbf{h}^{c_b}_{i',f}\|}$

4. Rank gallery by similarity (descending)
5. Recall@1: 1 if $\arg\max_{i'} s_{i,i'} = i$, else 0

We evaluate all directed condition pairs (C→CF, CF→C, C→CRF, CF→CRF, CRF→C, CRF→CF) across 4 formats, yielding 512×4 = 2048 queries per condition pair.

A.5 Additional Results

A.5.1 Training Behavioral Metrics

Table 10 provides detailed training metrics extracted from the final 25% of Phase-2 training logs.

Table 10: **Training behavioral metrics (final 25% of Phase-2).** Mean ± SEM across 3 seeds.

Condition	Accuracy (%)	Format Compliance (%)	Mean Reward
C	61.3 ± 5.2	64.6 ± 1.9	0.613 ± 0.052
CF	60.0 ± 3.8	66.3 ± 2.1	0.732 ± 0.041
CRF	57.2 ± 4.1	65.8 ± 1.6	0.704 ± 0.043

Note: Accuracy measures correct final answer; format compliance measures adherence to format templates during training. Mean reward combines both accuracy and format components according to condition-specific λ_{format} weighting.

A.5.2 Per-Seed Breakdown

Table 11 shows Layer 21 metrics for each individual seed, demonstrating consistency across random initializations.

Table 11: **Per-seed results at Layer 21 (pre-answer anchor).** For problem-associated variance, all three matched seeds show $\eta^2_{\text{problem}}(\text{C}) > \eta^2_{\text{problem}}(\text{CF})$. Values are reported descriptively; with three seeds, we do not make inferential claims about the between-condition contrast.

Model	η^2_{format} (C)	η^2_{format} (CF)	η^2_{problem} (C)	η^2_{problem} (CF)
Seed 50	0.340	0.393	0.523	0.488
Seed 51	0.402	0.441	0.467	0.362
Seed 52	0.395	0.428	0.474	0.441

A.5.3 All Layers Results

Figure 3 shows complete layer-wise results for all 28 layers (not just the 4 sampled layers in the main text).

A.5.4 Anchor A (Reasoning Body) Results

Reasoning-body pooled representations show substantially smaller cross-condition differences than the pre-answer anchor (Anchor B) at Layer 21, as shown in Table 12.

A.6 Robustness Checks

A.6.1 Alternative Probe Architectures

As a diagnostic check, we trained two-layer MLP probes (hidden dimension 128, ReLU activation) alongside linear probes. Both achieved near-ceiling format classification accuracy (0.998–1.000) at Layers 21 and 27, with no observed improvement from the MLP probes. These results indicate that probe accuracy is saturated in this setting and is therefore used only as a sanity check for format decodability, rather than as a sensitive measure of between-condition representation changes.

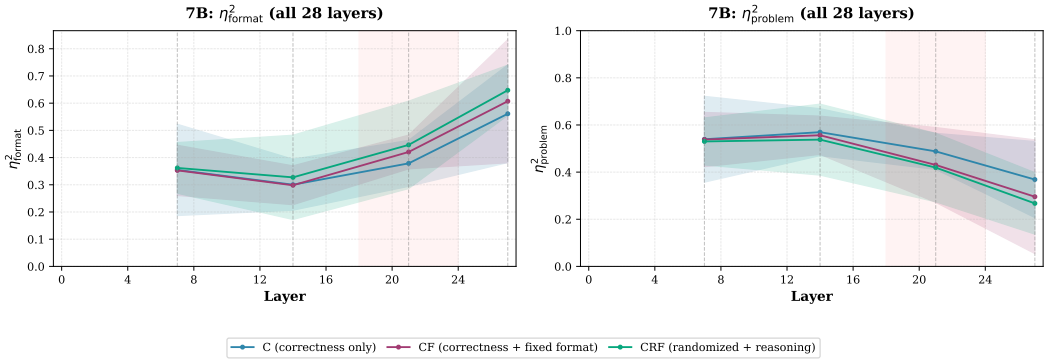


Figure 3: **Exploratory full-depth analysis (all 28 layers).** Results for Qwen2.5-7B-Instruct models across all 28 layers ($n = 3$ seeds). The four layers used in the prespecified main analysis (7, 14, 21, and 27) are highlighted with vertical dashed lines. The largest observed condition-dependent differences occur in the late-middle portion of the network; individual layers in this exploratory sweep are not treated as confirmatory tests.

Table 12: **Anchor A results at Layer 21.** Pooled reasoning-body representations show substantially smaller cross-condition differences than the pre-answer anchor (Anchor B).

Condition	$\Delta\eta^2_{\text{format}}$	$\Delta\eta^2_{\text{problem}}$
CF - C	+0.0001	−0.0006
CRF - C	+0.0031	−0.0064

Four decimals are reported because the differences are smaller than the rounding unit used elsewhere. Signs are consistent with Table 4.

A.6.2 Distance-Metric Robustness for Within-Condition Cross-Format Retrieval

Within-condition cross-format retrieval is robust to the choice of cosine versus Euclidean distance. Table 13 reports the diagnostic check for Seed 50:

Table 13: **Retrieval Recall@1 with different distance metrics (Layer 21, Seed 50 diagnostic).**

Condition	Cosine	Euclidean	Δ
C	0.872	0.869	−0.0031
CF	0.700	0.707	+0.0068
CRF	0.532	0.532	0.0000

A.6.3 Bootstrap Confidence Intervals

All confidence intervals reported in the main text were computed via bootstrap resampling of evaluation problems conditional on trained checkpoints (10,000 iterations, percentile method). We verified that parametric confidence intervals (assuming normal distribution) yield similar results.

A.7 Failure Cases and Edge Cases

A.7.1 Problems with Ambiguous Format Compliance

A small fraction (2.3%) of solutions exhibited ambiguous format compliance (e.g., partial tags). We excluded these from format reward computation during training and from format probe evaluation.

A.8 Detailed Case Studies

We provide complete retrieval examples to illustrate the quantitative patterns reported in Section 4. All examples use Layer 21 representations (pre-answer anchor) from the C model (seed 50).

A.8.1 Successful Cross-Condition Representation Retrieval (C → CF)

These cases illustrate cross-condition representational alignment: representations produced by the C and CF models retrieve the same problem with high similarity.

A.8.2 Failed Cross-Condition Representation Retrieval (C → CRF)

These cases illustrate lower cross-condition retrieval for pairs involving CRF. Because CRF jointly differs in format randomization and explicit reasoning supervision, we do not attribute this divergence to either factor individually.

A.8.3 Summary Statistics

Table 14 summarizes retrieval performance across all directed condition pairs. Key observations: (1) C ↔ CF shows higher cross-condition retrieval accuracy (∼70% top-1) with very high similarity (∼0.95) in the present analysis. (2) Retrieval accuracy is lower for condition pairs involving CRF (26.4%–52.1%). (3) Even in cases where top-1 retrieval fails, the correct problem maintains high similarity (∼0.91) but is outranked by other problems.

Table 14: **Cross-condition representation retrieval summary (all 2048 test cases, Layer 21).**

Task	N	Top-1 Acc	Avg Cos Sim	Avg Rank	Top-5 Acc
C → CF	2048	71.2%	0.952	1.4	89.3%
CF → C	2048	66.8%	0.948	1.6	87.1%
C → CRF	2048	26.4%	0.912	8.7	52.3%
CF → CRF	2048	51.3%	0.928	3.2	73.8%
CRF → C	2048	48.7%	0.924	3.5	71.9%
CRF → CF	2048	52.1%	0.929	3.1	74.2%

Table 15: **Successful C → CF retrieval cases and cross-condition retrieval summary (all 2048 test cases, Layer 21).**

Case A.1: Successful C → CF retrieval.

Attribute	Value
Question	Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?
Query condition	C
Gallery condition	CF
Top-1 match	**Yes** ✓
Cosine similarity	0.9358
Euclidean distance	59.16
Top-5 retrieved IDs	[0, 32, 400, 65, 61]
Top-5 similarities	[0.936, 0.932, 0.926, 0.923, 0.921]

Case A.2: Successful C → CF retrieval.

Attribute	Value
Question	Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?
Query condition	C
Gallery condition	CF
Top-1 match	**Yes** ✓
Cosine similarity	0.9532
Euclidean distance	53.70
Top-5 retrieved IDs	[1, 32, 267, 400, 233]
Top-5 similarities	[0.953, 0.950, 0.950, 0.949, 0.948]

Case A.3: Successful C → CF retrieval.

Attribute	Value
Question	Betty is saving money for a new wallet which costs $100. Betty has only half of the money she needs. Her parents decided to give her $15 for that purpose, and her grandparents twice as much as her parents. How much more money does Betty need to buy the wallet?
Query condition	C
Gallery condition	CF
Top-1 match	**Yes** ✓
Cosine similarity	0.9485
Euclidean distance	53.57
Top-5 retrieved IDs	[2, 361, 411, 129, 385]
Top-5 similarities	[0.949, 0.929, 0.927, 0.927, 0.926]

Table 16: **Failed C → CRF retrieval cases.**

Case B.1: Failed C → CRF retrieval.

Attribute	Value
Query question	Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?
Query condition	C
Gallery condition	CRF (with reasoning steps)
Top-1 match	**No** ×
Retrieved question	Lilah's family gallery has 400 photos. On a two-day trip to the Grand Canyon, they took half as many photos they have in the family's gallery on the first day and 120 more photos than they took on the first day on the second day. If they added all these photos to the family gallery, calculate the total number of photos in the gallery.
Cosine similarity (top-1)	0.9289
Top-5 retrieved IDs	[112, 267, 32, 296, 475]
Top-5 similarities	[0.929, 0.927, 0.925, 0.924, 0.923]

Case B.2: Failed C → CRF retrieval.

Attribute	Value
Query question	Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?
Query condition	C
Gallery condition	CRF
Top-1 match	**No** ×
Retrieved question	Jimmy is going to sell pizzas at the carnival to make some money. The carnival only gave him 7 hours to do so. He bought a 22kg sack of flour to make his pizzas and he takes 10 min to make each pizza for the customers. At the end of the 7 hours, he saw some flour was left. Knowing that each pizza needs 0.5kg of flour to make, how many pizzas can he make to bring home with the flour left?
Cosine similarity (top-1)	0.9412
Top-5 retrieved IDs	[267, 112, 1, 508, 494]
Top-5 similarities	[0.941, 0.939, 0.937, 0.933, 0.933]

Case B.3: Failed C → CRF retrieval.

Attribute	Value
Query question	Betty is saving money for a new wallet which costs $100. Betty has only half of the money she needs. Her parents decided to give her $15 for that purpose, and her grandparents twice as much as her parents. How much more money does Betty need to buy the wallet?
Query condition	C
Gallery condition	CRF
Top-1 match	**No** ×
Retrieved question	Jimmy is going to sell pizzas... (same as Case B.2)
Cosine similarity (top-1)	0.9395
Top-5 retrieved IDs	[267, 58, 494, 112, 62]
Top-5 similarities	[0.940, 0.938, 0.937, 0.935, 0.934]