THE VANISHING MIDDLE: REINFORCEMENT LEARNING ERASES THE UNCERTAIN CATEGORY IN ORDINAL CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In high-stakes decisions, such as clinical diagnosis or loan approval, the most useful answer is often "we cannot decide": an uncertain middle category that flags the case for a human and breaks the ambiguity. We show that reinforcement learning with verifiable rewards (RLVR), a standard way to improve large language model classifiers, systematically erases exactly this category. The setting is *ordinal* classification, whose labels lie on an ordered scale with a distinct middle. In detail, RLVR pushes answers toward the most common categories, and the rare middle loses recall. However, overall accuracy actually rises, making it hard to notice that the "we cannot decide" output is gone. Rebalancing the classes is not a cure we can deploy, and a second cause is at work: the middle is a two-sided range (two cutoffs), and at a matched frequency it collapses far more than a one-sided end (for example, a confidently positive or negative label). The fix therefore has to be structural. We propose **CURB** (Cumulative-ray Reparameterization of the Bounded middle), which rewrites the middle as two one-sided cutoff decisions at the model's output. CURB restores the middle across every setting and all three model families, without making it dominant.

1 INTRODUCTION

In high-stakes decisions, the most valuable answer a model can give is sometimes "I cannot decide." A clinician interpreting a genetic variant, a loan officer weighing an application, or a fact-checker adjudicating a claim would all rather receive an honest "the evidence is insufficient" than a confident answer that is wrong. These are *ordinal* classification problems: the labels lie on an ordered scale from one confident extreme to the other, with a distinct middle category that expresses this uncertainty. In clinical genetics, for example, the ACMG/AMP guideline sorts a variant into Pathogenic, Likely Pathogenic, Variant of Uncertain Significance, Likely Benign, or Benign, and the middle "Variant of Uncertain Significance" is the honest "we cannot call this" verdict. It is often the most consequential output, because a confident but wrong decisive call can directly harm a patient.

To improve models on such tasks, the standard tool is reinforcement learning with verifiable rewards (Lambert et al., 2024; DeepSeek-AI et al., 2025), typically Group Relative Policy Optimization (GRPO; Shao et al., 2024). It is preferred over supervised fine-tuning on the same labels because it optimizes the checkable objective directly and learns from the model's own wrong answers rather than imitating a reference. We find that this very mechanism has an unintended and dangerous effect: it **erases the uncertain middle category**.

When we apply GRPO with a right/wrong reward and low-rank adapter fine-tuning of Qwen2.5-7B-Instruct to the genetics task, recall on the uncertain category (the fraction of truly uncertain variants it still labels uncertain) falls from 0.214 to 0.057 (averaged over three runs). Meanwhile the two confident categories grow, and because they dominate the label mix, the *overall accuracy* actually *rises* from 0.41 to 0.49 (Figure 1). A practitioner who tracks only accuracy would conclude that the training helped, when in fact it has converted honest uncertainty into confident, and sometimes wrong, decisive answers. The failure is dangerous precisely because it is hidden: standard evaluation reports an improvement.

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.

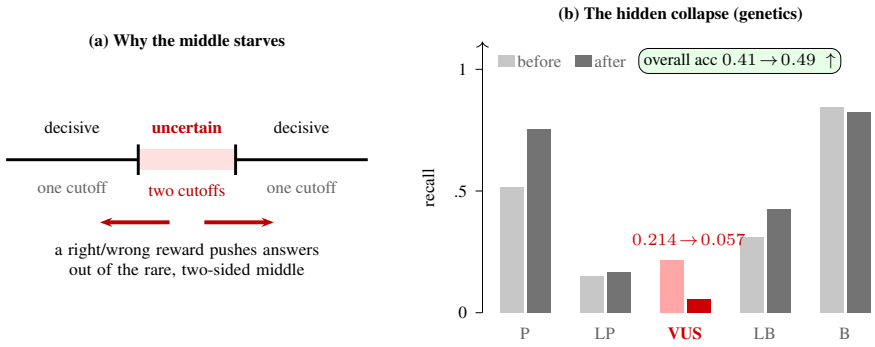

Figure 1: **The vanishing middle.** *(a)* On a three-category axis the uncertain middle is the only two-sided region; a right/wrong reward points only toward the ends. *(b)* Genetics per-category recall before (light) and after (dark) RLVR (GRPO, Qwen2.5-7B-Instruct).

Intuitively, a right/wrong reward acts as a *decisiveness ratchet*: because saying "uncertain" so rarely earns the reward, the model learns to stop hedging and commit to the most likely decisive answer. We trace the collapse to two causes. The first is prevalence: the training pushes answers toward the most common categories, so a rare middle collapses while a common one is strengthened. The second appears once frequency is controlled: the middle is a two-sided range, bounded by two cutoffs, and at a matched frequency it collapses far more than a one-sided end. This structural cause points to the fix. Rewriting the middle as two one-sided cutoff decisions *at the model's output*, which we call CURB, restores it, whereas a reformulation that still names the middle directly does not, so the cure is the shape of the label, not its name.

This paper makes three contributions.

- **A hidden failure across six settings** (Section 2.3): the rare uncertain category collapses while overall accuracy rises.
- **A prevalence law** (Sections 2.3 and 3): the middle collapses only when rare, shown causally by varying only its frequency, and is strengthened when common.
- **A structural second cause and a deduced fix** (Sections 3.2, 4): the middle is a two-sided range; CURB rewrites it as one-sided cutoff decisions at the output and restores it.

2 The Vanishing Middle

2.1 Measuring the middle

To measure whether training preserves the uncertain middle, we track per-category recall before and after reinforcement learning, holding the model, prompt, and evaluation set fixed and changing only the training stage. The training we study is standard RLVR. For each prompt the model samples several candidate answers, a verifier scores each one against the gold label (here a simple right/wrong reward), and the policy is updated to make the higher-scoring answers more likely. We use Group Relative Policy Optimization (GRPO; Shao et al., 2024), the now-standard RLVR optimizer: rather than learning a separate value network as in PPO (Schulman et al., 2017), it scores each sampled answer *relative to the other samples for the same prompt*, where the group average is the baseline. Concretely, starting from Qwen2.5-7B-Instruct, we run a brief supervised warm-up (supervised fine-tuning on a small set of examples) and then the GRPO stage, fine-tuning only low-rank adapters on the attention and feed-forward layers; the warm-up and reinforcement stages use disjoint prompts, so the model is never reinforced on the examples it was warmed up on. Throughout the paper, "before" is this warm-up model and "after" is the model after the reinforcement stage, so every change we report is attributable to the reinforcement stage. The collapse does not depend on this optimizer choice: it reappears under a variant without reward normalization (Dr. GRPO; Liu et al., 2025) and under the older PPO (Schulman et al., 2017).

Table 1: Middle-category recall and overall accuracy, before → after training, in the six settings (controls in the lower block). Spreads are over three runs; values without a spread are single runs; change is relative to before.

task	middle category	how common	middle recall before → after	change	overall accuracy before → after
Genetics	Uncertain	14%	$0.214 \to 0.057 \pm .011$	-73%	$0.41 \to 0.49$
Biomedical QA	maybe	11%	$0.242 \to 0.061$	-75%	$0.71 \to 0.75$
Financial fact-checking	not-enough-info	19%	$0.325 \to 0.149 \pm .007$	-54%	$0.67 \to 0.70$
Political fact-checking	uncertain	15%	$0.444 \to 0.366 \pm .015$	-18%	$0.77 \to 0.78$
Security severity	medium	41%	no collapse[†]	—	$0.58 \to 0.58$
Central-bank language	neutral	49%	$0.824 \to 0.833 \pm .000$	$+1\%$	$0.74 \to 0.74$

[†]In the security task the reward does not rise during training (it stays near 0.61), so this task serves only as a frequency control, not as a probe of the training dynamics. The naive 3× middle-weighted reward is a provable no-op (Table 10 in the appendix) and is omitted here.

2.2 Datasets and settings

We measure per-category recall on six ordinal tasks, each pairing a distinct uncertain middle with two confident extremes. Four are tasks whose middle is rare, where we expect the failure: clinical genetics (our motivating, high-stakes task), biomedical question answering, and financial and political fact-checking. Two more serve as controls because their middle is instead common: computer-security severity scoring (a common "medium") and central-bank language (dovish / *neutral* / hawkish, with neutral common). Table 1 lists all six with the frequency of their middle category.

Each example presents the evidence a decision must be read off, and the label is the ordered verdict. In clinical genetics an example is a genetic variant with a set of evidence items; a published scoring rule assigns each item a positive or negative number of points, and the verdict is read off the summed total. For example, a variant with strong functional-damage evidence ($+4$) and a rare-in-population signal ($+2$) totals $+6$, reaching the Likely Pathogenic band, while one whose only signals are a $+2$ and a conflicting -1 totals $+1$, inside the uncertain band. Fixed cutoffs then map the total to the five ordered categories, from Benign at low totals through Likely Benign, Variant of Uncertain Significance in the middle band, and Likely Pathogenic, to Pathogenic at high totals. The other three collapsing tasks are text classification with a three-way ordered label read off the input text: biomedical question answering (yes / *maybe* / no), financial fact-checking (false / *not enough information* / true), and political fact-checking (refuted / *uncertain* / supported).

2.3 The middle collapses while accuracy rises

Table 1 reports what training does to each setting: wherever the middle is rare, its recall collapses while overall accuracy rises. On the genetics task the per-category recall moves (before → after; decisive tiers from a representative run, the uncertain value being the three-run mean): Pathogenic $0.515 \to 0.755$, Likely Pathogenic $0.150 \to 0.167$, **Uncertain** $0.214 \to 0.057$, Likely Benign $0.309 \to 0.426$, Benign $0.846 \to 0.824$: every confident category holds or grows while the middle collapses, and the collapse is consistent across three runs ($0.071, 0.054, 0.045$; mean 0.057, s.d. 0.011). The same pattern holds in all four rare-middle settings, and the two common-middle controls do not collapse. Crucially, because the confident categories dominate the label mix, the overall accuracy *rises* in every collapsing setting (on genetics, $0.41 \to 0.49$): a reader who tracks only accuracy would conclude that training helped, when it has in fact converted honest uncertainty into confident, and sometimes wrong, decisive answers (Figure 1). The collapse is not an artifact of category difficulty, answer length, formatting, or the number of distinct evidence combinations; none of these tracks which categories survive (Appendix E). *Why* the rare middle collapses, and why the cure must change the model's output rather than the data, is the subject of Sections 3 and 4.

3 Why the middle vanishes

The collapse is governed primarily by two properties of the category: how rare it is, and whether it is bounded on both sides. We establish each the same way, by holding everything else fixed and moving only that property, and we address the most plausible alternatives, the category's *meaning*, its

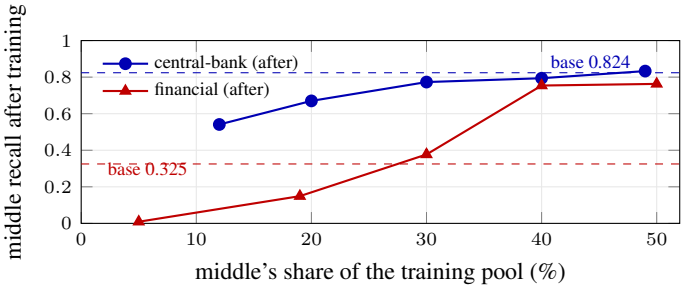

Figure 2: Middle recall after training as only the middle's pool share varies (warm-up fixed); dashed lines mark each task's pre-training recall.

Table 2: Rebalancing on biomedical QA: middle (`maybe`) recall under the natural pool versus a class-balanced pool ($1/3$ each), from the same warm-up model.

task	natural pool	balanced pool ($1/3$ each)	base (pre-training)
Biomedical QA	0.061	0.273	0.242

difficulty, and model *confidence*, by showing each is largely a confound of these two rather than a cause in its own right. We close by showing the failure is specific to reward-based training.

3.1 Cause 1: how common the category is

The dominant cause is prevalence. Reinforcement learning pushes answers toward the most common categories, so a middle that is rare collapses while a middle that is common is strengthened. The middle is rare in exactly the settings that collapse (about 14% of the genetics examples and 11% of the biomedical ones) and common in the two controls, which do not.

We test this causally by holding the task and the warm-up model fixed and changing *only* how often the middle appears in the reinforcement-learning pool. Lowering the middle's frequency lowers its post-training recall, in a graded, monotone way (Figure 2). On the central-bank control the neutral category's recall crosses its starting value (0.824) at about a 45% frequency, collapsing below that and strengthening above it. The same graded response appears *inside* a genuine collapse domain, not only in the control: sweeping the financial not-enough-information fraction from 5% to 50% moves its post-training recall from 0.009 up to 0.763, crossing its base (0.325) near 28%. We call a middle "rare" or "common" only for readability. The boundary is not a threshold we impose but a continuous crossover the sweeps reveal, and its location is task-dependent (near 28% here and 45% for the central-bank control).

Rebalancing confirms the same cause from the other direction: if rarity drives the collapse, equalizing the classes should undo it. Holding the warm-up model fixed and rebuilding the reinforcement-learning pool so the classes are equally common, the biomedical middle recovers from its collapsed 0.061 back above its pre-training level (Table 2). And where the middle is common to begin with, in the two control tasks, it never collapses in the first place.

3.2 Cause 2: the two-sided middle

Prevalence is not the whole story, because rarity alone does not say *which* rare category collapses. The answer is geometric. On the score axis that the label is read from, the two end categories are one-sided, each defined by a single cutoff (the lowest scores, the highest scores), while the middle is a two-sided range pinned between two cutoffs. In a ranked task this is not a coincidence but a necessity: the extremes are open on one side and only interior categories are bounded on both. The uncertain middle always sits in the interior and is therefore two-sided, and it is this two-sidedness, not merely its rarity, that makes it fragile.

We show the effect is structural with a matched-frequency ablation (Table 3; details in Appendix B): we make each category in turn the minority at the same 12% frequency, so rarity is held equal and only the geometry differs. The two-sided middle then collapses several times more than a one-sided end

Table 3: Matched-frequency ablation: recall of the minority category (each in turn at a 12% pool share) before → after training, on the control and the clinical task; the genetics arm uses a *graded* criteria reward; relative drop in parentheses.

minority category (12%)	side	recall before → after	drop
Central-bank language (3-tier; all-or-nothing reward)			
neutral (middle)	two-sided	0.824 → 0.485	−0.339 (−41%)
dovish (extreme)	one-sided	0.697 → 0.597	−0.100 (−14%)
hawkish (extreme)	one-sided	0.641 → 0.555	−0.086 (−13%)
none (balanced)	—	0.824 → 0.798	−0.026 (−3%)
Clinical genetics (5-tier; VUS = two-sided middle; *graded* criteria reward)			
VUS (middle)	two-sided	0.214 → 0.027	−0.187 (−87%)
benign (extreme)	one-sided	0.846 → 0.736	−0.110 (−13%)
pathogenic (extreme)	one-sided	0.515 → 0.362	−0.153 (−30%)
VUS, balanced	—	0.214 → 0.214	0.000 (0%)

Table 4: Recall before → after training for the two collapsing middles and four survivors under the same recipe; the degree-type middles are separate control tasks, outside the six settings of Table 1; change is relative to before; genetics rows are from the representative run of §2.3.

category (task)	kind	recall before → after	change
Variant of Uncertain Significance (genetics)	uncertain middle	0.214 → 0.071	−67%
maybe (biomedical QA)	uncertain middle	0.242 → 0.061	−75%
Likely Pathogenic (genetics)	confident tier	0.150 → 0.167	+11%
neutral (sentiment)	degree middle	0.543 → 0.504	−7%
three stars (rating)	degree middle	0.550 → 0.600	+9%
three stars, hard (rating)	degree middle	0.325 → 0.400	+23%

made equally rare, on both the control and the clinical task, and in this ablation's balanced arm, where every category is equally common, the middle does not collapse at all. Rarity and two-sidedness together erase the middle; neither alone does.

The two causes point to where a fix must act: the middle collapses because a rare, two-sided region is starved under a reward that only ever points one way, so the remedy is to remove the two-sided region from what the model must output. We develop this fix in Section 4.

3.3 Not a confound, and not a quality trade-off

Having attributed the collapse to prevalence and two-sidedness, we now rule out the alternatives a reader might reach for first: that the middle is a *special kind* of category, either an uncertain hedge or simply the hardest one; that the effect is really about model *confidence*; and that it is a generic property of any classifier rather than of reward-based training. None of these holds.

The middle is not lost because it is a special category. Two versions of this worry arise together. One is that the collapsing categories are all epistemic hedges, answers that decline to commit ("uncertain", "maybe", "not enough information"), so training might simply suppress hedging. The other is that the middle is just the hardest category to get right. Table 4 addresses both by comparing the collapsing middles against four categories that survive the same training. Difficulty is ruled out because the survivors enclose the collapsers in difficulty: their base recalls run from 0.150 (Likely Pathogenic, harder than either collapsing middle) through 0.325 (comparable) up to 0.550 (far easier), so the collapsers sit strictly inside the survivors' difficulty range and no difficulty threshold can separate the two groups. The hedging worry is subtler, because the survivors happen to be degree-type middles that merely name a midpoint while the collapsers are hedges, which looks like a meaning effect. But hedging is confounded with rarity, since the hedging middles are also the rare ones. Once frequency is controlled (Table 5), the outcome follows frequency and not meaning: a hedging middle made common is strengthened, while a substantive, non-hedging middle made rare collapses. Answer length, formatting failures, the number of distinct evidence combinations, and the spread of the confusion pattern are ruled out the same way (Appendix E).

Table 5: Middle recall before → after training when a hedging versus a substantive middle is made rare or common (pool share in parentheses; two points on the frequency axis of Figure 2).

middle category	made rare	made common
hedging (financial, not-enough-info)	$0.325 \rightarrow 0.149$ (19%)	$0.325 \rightarrow 0.754$ (40%)
substantive (central-bank, neutral)	$0.824 \rightarrow 0.541$ (12%)	$0.824 \rightarrow 0.833$ (49%)

The collapse is also orthogonal to model confidence, the most reflexive reading of our result being that reinforcement learning merely makes the model overconfident. Three observations rule that out. First, in the frequency sweep of Section 3 the warm-up model is identical across every column, so the model's starting confidence on the middle is held fixed while only frequency changes; the collapse still tracks frequency, which a confidence account cannot produce. Second, on the genetics task the model's token entropy is essentially unchanged by training (mean 0.192 versus 0.188), so there is no entropy collapse to blame. Third, it is not a failed run: on the central-bank control the reward rises ($0.65 \rightarrow 0.77$) while the common neutral is *strengthened*. A tier that stops being emitted cannot be recovered by recalibrating confidence, which is why the fix in Section 4 changes the output rather than the confidence.

Finally, the collapse is a property of reward-based training rather than of the label distribution itself. On the financial task, an ordinary supervised model (cross-entropy loss) trained on the *same* rare-middle pool that the reinforcement run collapses keeps the middle alive (recall $0.523 \pm .033$ across three seeds) and does so at a *higher* overall accuracy (0.72–0.74) than the reinforcement run (0.69), which nonetheless drives the same middle to 0.149. The reinforcement stage does not trade accuracy for the collapse; it is the reward-based objective, not the rare label, that erases the middle.

4 The fix and why it is specific

4.1 Why the obvious fixes fail

The obvious remedy is the standard one for a rare class: up-weight it, multiplying the reward for a correct middle answer by a factor $c > 1$ while any other correct answer keeps reward 1 and a wrong answer 0. Under GRPO no choice of c can work. As Section 2.1 described, GRPO scores each sampled answer only relative to the other samples for the same prompt, and every sample is judged against that prompt's single gold label. On a prompt whose gold is the middle, the only correct answers are middle answers, so the up-weight turns the group's rewards from $\{1, 0\}$ into $\{c, 0\}$; that is, it multiplies every reward in the group by the same c. The advantage is invariant to that scaling, because the group mean and standard deviation scale by the same constant:

$$A_i \;=\; \frac{c\, r_i - c\, \bar{r}}{c\, \sigma_r} \;=\; \frac{r_i - \bar{r}}{\sigma_r}.$$

Up-weighting the middle is therefore an *exact no-op* on the gradient, for every $c > 1$.[1] We confirm this at $c = 3$: from identical warm-up models, the $3\times$ reward matches plain RLVR within run-to-run noise on every task and backbone tested (0.149 vs. 0.149 on the financial task; Table 10 in the appendix).

The second obvious remedy is to rebalance the training pool, which does remove the collapse (§3) but repairs the data rather than the model: in exactly the settings where the middle matters it is rare in the world, upsampling recycles a handful of examples, downsampling discards the plentiful decisive ones, and the training distribution drifts from the one the model will face. A usable fix has to work at the natural class frequencies. Grading the reward is not enough either: on the hard clinical task a graded, cumulative-threshold reward still leaves the middle collapsed (Appendix D). With the reward side and the data side exhausted, the remaining lever is the output.

4.2 CURB

The middle can be missed from *both* sides: an answer that is too low and one that is too high both count as wrong. It is a narrow ledge between two open cliffs, and under an all-or-nothing reward

[1] The one exception is an unparseable answer, which carries a fixed penalty rather than a scaled reward; with a parse rate of ≈ 1 throughout our runs this term is negligible.

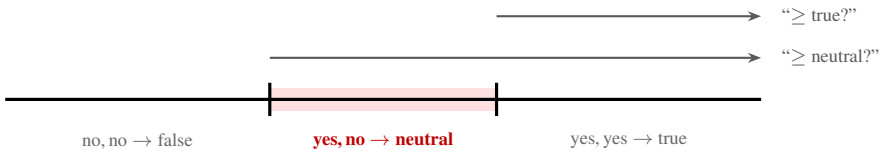

Figure 3: CURB removes the two-sided target of Figure 1(a): the model answers one one-sided cutoff question per boundary, and the middle is the leftover gap, never a region that can be missed from both sides.

Table 6: Middle recall, before → after training, under the single-label format (repeating Table 1) and under CURB, on the same natural pools; each arm starts from its own format-matched warm-up, so the two "before" values differ. Spreads are over three runs; values without a spread are single runs; change is relative to before.

	single-label RLVR		CURB	
task	before → after	change	before → after	change
Genetics	$0.214 \rightarrow 0.057 \pm .011$	-73%	$0.277 \rightarrow 0.229 \pm .034$	-17%
Biomedical QA	$0.242 \rightarrow 0.061$	-75%	$0.485 \rightarrow 0.273$	-44%
Financial fact-checking	$0.325 \rightarrow 0.149 \pm .007$	-54%	$0.246 \rightarrow 0.321 \pm .023$	$+30\%$
Political fact-checking	$0.444 \rightarrow 0.366 \pm .015$	-18%	$0.531 \rightarrow 0.502 \pm .006$	-5%

the training backs away from the ledge toward the safe, one-sided cliffs; this is the outward push of Figure 1(a). The fix is to remove the ledge. Instead of asking "which category is it?", a question the middle can fail from either side, we ask a set of one-sided questions, "is the label at least category t?", one for each of the $K-1$ boundaries, and read the category off the answers (Figure 3). We name this output-side fix **CURB** (*Cumulative-ray Reparameterization of the Bounded middle*). In the three-category case (`false<neutral<true`) the model answers "at least neutral?" and "at least true?", and `neutral` is simply *yes then no*: the gap between the two boundaries, never a target in its own right. In the five-category genetics label the four questions leave the uncertain category as the pattern (yes, yes, no, no); for the monotone answers the model almost always gives, the category is just how many boundaries it clears. Because every question is one-sided, the bounded middle is *never itself a target*: the all-or-nothing reward has no two-sided region to starve. CURB is the classical Frank–Hall decomposition; our contribution is not the technique but the diagnosis of *why* it specifically undoes the reward-based collapse.

4.3 CURB RESTORES THE MIDDLE AT NATURAL RARITY

Table 6 shows the result that the cheap fixes could not deliver: with the training pool left at its natural frequencies, so the middle is exactly as rare as when it collapsed, CURB restores the middle in all four collapsing settings. Its role is to *undo the collapse*, not to push the middle above its pre-training level: on the financial task it roughly doubles the collapsed recall ($0.149 \rightarrow 0.321$) and returns it to about the single-label base (0.325), and on the clinical task the restoration is likewise to approximately the pre-training level ($0.057 \rightarrow 0.229 \pm .034$, around the 0.214 start). We therefore claim restoration *of the collapse*, not improvement beyond baseline.

The restoration is not specific to one backbone. On financial fact-checking, with only the base model varied, the middle collapses under single-label training and is restored by CURB for Qwen2.5-7B ($0.325 \rightarrow 0.149$ vs. CURB $0.246 \rightarrow 0.321$), Mistral-7B ($0.439 \rightarrow 0.263$ vs. $0.351 \rightarrow 0.596$), and OLMo-7B ($0.430 \rightarrow 0.237$ vs. $0.342 \rightarrow 0.430$). The one dependence is task strength: on a weak-collapse task there is little to restore.

4.4 ONLY THIS DECOMPOSITION, AND ONLY AT THE OUTPUT

The fix works at the output, not at the reward (Table 7). On the clinical task every reward- or advantage-side intervention leaves the middle collapsed: a graded cumulative-threshold reward (ODRPO) and a correctness-clipped baseline (CoRPO) both land near the plain $0/1$ level (0.045–0.085 VUS recall), while moving the *same* cumulative decomposition to the policy *output* (CURB) holds it; the optimizer-side focal reweighting F-GRPO fails likewise on the financial task (Appendix C). Strikingly, once the output is the threshold vector the reward shape barely matters: a

Table 7: Reward-side versus output-side interventions on the clinical task (middle = VUS); same natural pool, recipe, and evaluation, each method from its format-matched warm-up; change is relative to base. †every $\tau \in \{0.5, 0.75, 0.9\}$ ($0.045/0.08/0.054$). ‡a $0/1$-exact reward on the CURB output (seeds $0.205, 0.268$).

method	where it acts	VUS recall (base → after)	change
plain RLVR ($0/1$)	reward	$0.214 \to 0.057$	-73%
ordinal-distance reward	reward	$0.214 \to 0.071$	-67%
ODRPO (cumulative-graded)	reward	$0.214 \to 0.085$	-60%
CoRPO (baseline clip)	advantage	$0.214 \to 0.045$–$0.08^{\dagger}$	-79 to -63%
CURB (threshold output)	**output**	$0.277 \to 0.229 \pm .034$	$\mathbf{-17\%}$
CURB, $0/1$-exact reward	**output**	$0.277 \to 0.237^{\ddagger}$	$\mathbf{-14\%}$

Table 8: Middle recall under three answer formats on two domains; same data per domain, only the answer format changes, each with its own format-matched warm-up; change is relative to before.

answer format	how the middle is reported	before → after	change
Financial fact-checking (middle = not-enough-info)			
single label	named directly	$0.325 \to 0.149$	-54%
three independent yes/no	one question names the middle	$0.114 \to 0.035 \pm .009$	-69%
two one-sided cutoffs (CURB)	never a direct target	$0.246 \to 0.321$	$+30\%$
Biomedical QA (middle = maybe), a second strong-collapse domain			
single label	named directly	$0.242 \to 0.061$	-75%
three independent yes/no	one question names the middle	$0.364 \to 0.061$	-83%
two one-sided cutoffs (CURB)	never a direct target	$0.485 \to 0.273$	-44%

harsh $0/1$-exact reward on the CURB output holds the middle as well as the graded reward (0.237 vs. 0.229), so it is the output reparameterization, not the reward, that fixes the collapse.

And the rewriting must be the one-sided one (Table 8). On the financial fact-checking task we keep the data fixed and change only how the model reports its answer. An alternative reformulation that asks three independent yes/no questions, one of which is still "is this the middle category?", *trains well* (its reward reaches 0.74) yet still destroys the middle, because it asks the model to name the two-sided middle directly. Only removing that direct two-sided target, by asking the two one-sided cutoff questions of CURB instead, restores it. The pattern replicates on a second strong-collapse domain (biomedical QA, Table 8): naming the middle directly collapses it in either format, while the cumulative cutoffs preserve it, so the effect is not specific to one dataset.

5 Related work

Reinforcement learning for language model post-training. Reinforcement learning is now a standard part of language model post-training, beginning with learning from human preferences (Christiano et al., 2017; Ouyang et al., 2022). Reinforcement learning with verifiable rewards (RLVR) replaces the learned preference model with an automatic checker (Lambert et al., 2024), is now a dominant recipe for reasoning-oriented training (DeepSeek-AI et al., 2025), and sharpens performance beyond a supervised start on checkable tasks (Yue et al., 2025); the property behind this edge, learning from the model's own wrong samples, is what drives the failure we study. Most RLVR systems are critic-free policy-gradient methods, chiefly Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which replaces the value model of PPO (Schulman et al., 2017) with a group-relative advantage, with later variants adjusting normalization and clipping (Liu et al., 2025; Yu et al., 2025).

RLVR for evidence-based reasoning. Our tasks are not flat label assignment. In clinical variant interpretation, biomedical question answering, and claim adjudication, a model must gather, weigh, and aggregate evidence before committing to a verdict, which the fact-verification literature frames as a multi-stage decision process rather than single-step classification (Yang et al., 2026), and which clinical guidelines instantiate as synthesizing graded evidence into graded recommendations (Tziakouri & Menolascina, 2025). Because the final verdict is verifiable against an expert label while the intermediate weighing is not, these tasks are natural targets for RLVR (Lambert et al., 2024; Shao

et al., 2024), extended from math and code to domains whose answers are only partially checkable (Gunjal et al., 2025); and it delivers real gains on decision-making classification (Yaldiz et al., 2026), clinical assessment (Lee et al., 2026), and evidence-grounded verdict prediction (He et al., 2025). What sets these label spaces apart is a tier that asserts no direction: the verdict that the evidence is insufficient or conflicting, appearing as *not enough evidence* in claim verification (Schlichtkrull et al., 2023), *maybe* in biomedical question answering (Jin et al., 2019), and *uncertain significance* under the ACMG criteria (Richards et al., 2015). Prior work has noticed this tier is underused, but attributes it to scale: larger, more confident models avoid the *not enough info* verdict (He et al., 2025). We show it is not merely underused but actively eroded by training, within a fixed model and orthogonally to confidence.

Reward-side versus output-side interventions. The methods above act on the *reward* or the optimizer. CoRPO clips GRPO's group baseline (Garg et al., 2025), F-GRPO reweights the advantage toward rarely sampled correct trajectories (Plyusov et al., 2026), and ODRPO decomposes the ordinal reward into cumulative one-sided threshold indicators (Patel et al., 2026). CURB applies that same cumulative decomposition at the *policy output* instead, and *where* it is applied is what matters: under an identical reward, the decomposition prevents the collapse only when it sits on the output, not the reward (Table 7; Appendices C, D). Concurrent work reaches the same conclusion for a binary abstain gate: making abstention an action saturates a gate that kills the reward gradient and the KL anchor together, and the repair moves the decision to an always-emitted output (Che et al., 2026). In our own head-to-head the collapse tracks the tier's two-sided structure rather than lost reward magnitude (Appendix D), and all of these interventions are complementary to CURB.

Ordinal decomposition and the evaluation setting. Rewriting an ordered label as a set of one-sided cutoff decisions is a classical idea (Frank & Hall, 2001; Li & Lin, 2006; Niu et al., 2016; Cao et al., 2020; Shi et al., 2023); our contribution is not the decomposition itself but the explanation of why this specific move cures a reward-based failure. We study it in clinical variant classification, where an uncertain tier (a variant of uncertain significance) is defined by a two-sided rule between the pathogenic and benign ends (Richards et al., 2015); the collapse and its cure replicate on yes/no/maybe biomedical question answering (Jin et al., 2019) and on financial and political fact-checking, while degree-type middles from sentiment and star-rating tasks (Socher et al., 2013) serve as surviving controls. Further adjacent failure modes and uncertainty-handling methods are surveyed in Appendix F.

6 Limitations

The usual boundaries of a study of this size apply. The collapse reproduces under GRPO, Dr. GRPO, and PPO (§2.1), and further reinforcement-learning variants such as RLOO (Ahmadian et al., 2024) remain to be tested; backbone generality is shown for three model families on one task (§4), and the size of the restoration tracks the strength of the collapse, so a weak-collapse task leaves little to restore. Statistical power varies with how rare a category is: per-category test counts and Wilson intervals are given in Table 9, the two best-powered collapse tasks are individually significant there, and arms quoted without a spread are single runs. The location of the rare-to-common crossover differs across tasks (near 28% and 45% in our two sweeps), and what determines it is left to future work. Finally, downstream clinical harm, how often an erased uncertain case becomes a confident wrong answer, is reported descriptively (Appendix E) rather than tied to patient outcomes.

7 Conclusion

A standard verifiable-reward recipe erases the ordinal middle exactly when it is rare, and overall accuracy hides the loss. The failure has two causes, how common the category is (a smooth, two-directional response to frequency) and the fact that the middle is a two-sided range rather than a one-sided cutoff, and the same object that causes it prescribes the cure. Rewriting the two-sided middle as one-sided cutoff decisions at the model's output, CURB, restores the middle in every collapsing setting at the natural class frequencies, without rebalancing any data and without making the middle dominant. In practice, any reward-based training on an ordered task whose middle category carries the risk should be judged by the recall of each category, not by overall accuracy.

Reproducibility statement

All tasks use public data and the public model Qwen2.5-7B-Instruct. The per-category recalls, the number of runs, the training settings, and how the warm-up and reinforcement pools are built are given in the appendix; the alternative answer formats are specified alongside Table 8. Code will be released on acceptance.

Ethics statement

The genetics results are a study of a training method on public curated data, not a deployed diagnostic tool. The safety point is a warning: overall accuracy can hide the erasure of the "uncertain" category.

Use of AI Assistance

Following the ICLR 2027 policy on the use of AI, we disclose our use of large language model (LLM) assistants in preparing this submission. LLMs were used to aid and polish writing and to draft sections of the paper; to retrieve and discover related work; to support research ideation and execution, including designing and implementing experiment code and analyzing and interpreting results; and to derive and check the mathematical argument for the reward-rescaling no-op (Section 4.1). LLMs were not used to generate synthetic datasets. The authors reviewed all AI-assisted output, including all citations, claims, code, and results, verified it, and take full responsibility for the content of this paper.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024. RLOO.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. CORAL; arXiv:1901.07884.

Xujun Che, Yuchen Yuan, Weida Zhao, and Chenyang Yu. Abstention as an action can kill both the reward gradient and the KL anchor: Collapse law and repair for error-penalized reinforcement learning. *arXiv preprint arXiv:2608.00301*, 2026.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.03741.

Ganqu Cui et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *ECML*, volume 2167 of *LNCS*, pp. 145–156. Springer, 2001. doi: 10.1007/3-540-44795-4_13.

Fardin Ganjkhanloo, Emmett Springer, Erik H. Hoyer, Daniel L. Young, and Kimia Ghobadi. Joint score-threshold optimization for interpretable risk assessment. *arXiv preprint arXiv:2510.21934*, 2025.

Anisha Garg, Claire Zhang, Nishit Neema, David Bick, Ganesh Venkatesh, and Joel Hestness. Corpo: Adding a correctness bias to grpo improves generalization. *arXiv preprint arXiv:2511.04439*, 2025. v1–v2 titled "The Peril of Preference: Why GRPO fails on Ordinal Rewards"; retitled CoRPO in v3.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

Qi He, Cheng Qian, Xiusi Chen, Bingxiang He, Yi R. Fung, and Heng Ji. Veri-r1: Toward precise and faithful claim verification via online reinforcement learning. *arXiv preprint arXiv:2510.01932*, 2025.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019. arXiv:1909.06146.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*, 2025.

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 *ICLR*, 2024. arXiv:2310.06452.

Nathan Lambert et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. Introduces the term RLVR.

Chaeeun Lee, T. Michael Yates, Pasquale Minervini, and T. Ian Simpson. Process-supervised multi-agent reinforcement learning for reliable clinical reasoning. *arXiv preprint arXiv:2602.14160*, 2026.

Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. In *NIPS*, volume 19, pp. 865–872, 2006.

Enlu Lin, Qiong Chen, and Xiaoming Qi. Deep reinforcement learning for imbalanced classification. *arXiv preprint arXiv:1901.01379*, 2019.

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. Dr.GRPO.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, pp. 4920–4928, 2016.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. InstructGPT; arXiv:2203.02155.

Nirmal Patel, Fei Wang, and Inderjit S. Dhillon. Odrpo: Ordinal decompositions of discrete rewards for robust policy optimization. *arXiv preprint arXiv:2605.12667*, 2026.

Daniil Plyusov, Alexey Gorbatovski, Boris Shaposhnikov, Viacheslav Sinii, Alexey Malakhov, Daria Korotyshova, and Daniil Gavrilov. F-grpo: Don't let your policy learn the obvious and forget the rare. *arXiv preprint arXiv:2602.06717*, 2026.

Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L. Rehm. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–424, 2015.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. arXiv:2305.13117.

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, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. Introduces GRPO.

Xintong Shi, Wenzhi Cao, and Sebastian Raschka. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 26:941–955, 2023. CORN; arXiv:2111.08851.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, 2013.

Anni Tziakouri and Filippo Menolascina. Reinforcement learning for clinical reasoning: Aligning LLMs with ACR imaging appropriateness criteria. *arXiv preprint arXiv:2510.05194*, 2025.

Shaohang Wei, Zikun Su, Feifan Song, Wen Luo, Wei Li, Guangyue Peng, and Houfeng Wang. Verifier-induced support reshaping in on-policy optimization. *arXiv preprint arXiv:2608.00220*, 2026.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Jingxiang Chen, Mohammad Kachuee, Teja Gollapudi, Yiwei Liao, Nicolas Scheffer, Rakesh Wang, Anuj Kumar, Yu Meng, Wen-tau Yih, and Xin Luna Dong. Truthrl: Incentivizing truthful llms via reinforcement learning. *arXiv preprint arXiv:2509.25760*, 2025.

Fang Wu, Aaron Tu, Weihao Xuan, et al. Position: The hidden costs and measurement gaps of reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2509.21882*, 2025.

Duygu Nur Yaldiz, Evangelia Spiliopoulou, Zheng Qi, Siddharth Varia, Srikanth Doss, and Nikolaos Pappas. Balancing classification and calibration performance in decision-making LLMs via calibration aware reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12537–12553, 2026. arXiv:2601.13284.

Rongxin Yang, Shenghong He, Siyuan Zhu, and Chao Yu. From verdict to process: Agentic reinforcement learning for multi-stage fact verification. *arXiv preprint arXiv:2606.13262*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Suqin Yuan, Jinkun Chen, Jiyang Zheng, Muyang Li, Lei Feng, Dadong Wang, Tao Xiang, Tongliang Liu, and Bo An. Understanding diversity collapse in rlvr via the lens of overtraining, 2026.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025.

Xianzhou Zeng, Jing Huang, Chunmei Xie, Gongrui Nan, Siye Chen, Mengyu Lu, Weiqi Xiong, Qixuan Zhou, Junhao Zhang, Qiang Zhu, Yadong Li, and Xingzhong Xu. Ucpo: Uncertainty-aware policy optimization. *arXiv preprint arXiv:2601.22648*, 2026.

A EXPERIMENTAL DETAILS

A.1 Per-category test counts, recall, and significance. Table 9 gives, for each setting's middle category, the test count n and the middle recall before and after training, with Wilson 95% intervals. The collapse is individually significant (non-overlapping intervals) on the two best-powered collapse tasks (genetics, n=112; financial, n=114); biomedical (n=33) and political (n=81) show the same direction with overlapping intervals and are corroborative. The controls do not collapse. The primary evidence for the cause is the within-task causal frequency sweep (§3, Figure 2), not this cross-section; full per-category recalls for all tiers and seeds are in the released logs. **A.2 Training settings.** 400

Table 9: Middle-category recall with Wilson 95% intervals (test count n). The change is significant where the before/after intervals do not overlap.

setting	middle	n	before recall [95% CI]	after recall [95% CI]
Genetics (clinical)	VUS	112	$0.214\,[.148, .299]$	$\mathbf{0.057}\,[.027, .116]$
Financial	not-enough-info	114	$0.325\,[.246, .415]$	$\mathbf{0.149}\,[.095, .226]$
Biomedical QA	maybe	33	$0.242\,[.128, .410]$	$0.061\,[.017, .197]$
Political	uncertain	81	$0.444\,[.341, .552]$	$0.366\,[.269, .475]$
Central-bank (control)	neutral	233	$0.824\,[.770, .868]$	$0.833\,[.780, .875]$
Security (control)	medium	812	$0.657\,[.624, .689]$	$0.680\,[.647, .711]$

update steps, eight sampled answers per prompt, low-rank adapter rank 32, learning rate 5×10^{-6}, entropy weight 0.04, sampling temperature 0.9. **A.3 The naive-reward no-op.** Table 10 reports the controlled comparison behind the no-op argument of Section 4.

B MATCHED-FREQUENCY ABLATION: DETAILS

This appendix gives the details behind the matched-frequency ablation of Section 3.2 (Table 3). Holding the task and warm-up model fixed (central-bank language; base recall dovish 0.697, neutral 0.824, hawkish 0.641; parse rate 1.00), we make each category in turn the 12% minority in the reinforcement-learning pool (the other two balanced at 44%), plus a balanced $33/33/33$ control. The two-sided middle collapses by 0.34 while a one-sided extreme made equally rare drops only ~ 0.09, and the gap survives normalizing by base recall (-41% vs. -13 to -14% relative). A generic rare-trajectory-forgetting account (e.g. F-GRPO, Plyusov et al., 2026), which has no term for whether a label is one-sided or two-sided, predicts no such difference. On the clinical genetics task the two-sided VUS at a matched 12% is nearly erased ($0.214 \to 0.027$, -87%) while one-sided extremes lose far less (-13 to -30%). These matched-structure pools (one category at 12%, the other two balanced) differ from the frequency sweep of Section 3, which downsamples only the middle; the neutral collapse at 12% reads 0.485 here versus 0.541 there, both large drops from 0.824, the small gap reflecting pool construction and run-to-run variation.

C HEAD-TO-HEAD WITH AN OPTIMIZER-SIDE REWEIGHTING (F-GRPO)

F-GRPO (Plyusov et al., 2026) reweights the group-relative advantage by a per-prompt focal factor $g(x)=(1-\hat{p}_{\text{correct}}(x))^{\gamma}$ that up-weights low-success (hard) prompts. Because the minority middle is exactly the low-success case, one might expect it to rescue the middle. On the financial task, from the identical warm-up model, it does not (Table 11): the middle stays collapsed at every γ, and γ=1 *deepens* the collapse (up-weighting the hard middle prompts only amplifies the gradient toward the decisive extremes on them). Removing the effective-learning-rate side effect (mean-preserving g) matches plain RLVR exactly. Only the output-side cumulative-cutoff fix restores the middle, confirming the two interventions act on different axes (advantage magnitude versus output structure) and are not substitutes. (These are single fresh runs from one batch, so plain RLVR and the fix read slightly higher here, 0.167 and 0.360, than the multi-run values in the main tables, 0.149 and 0.321; the within-batch comparison is the meaningful one.)

Table 10: Middle recall after plain RLVR ($0/1$) versus the naive $3\times$ middle-weighted reward, both arms trained from the identical warm-up model, across tasks and backbones.

setting	warm-up	RLVR ($0/1$)	naive ($3\times$)	$\lvert\Delta\rvert$
Across tasks (Qwen2.5-7B)				
Financial fact-checking	0.325	0.149	0.149	0.000
Political fact-checking	0.444	0.366	0.370	0.004
Central-bank (control)	0.824	0.833	0.854	0.021
Across backbones (financial fact-checking)				
Mistral-7B‡	0.368	0.263	0.254	0.009
OLMo-7B†	0.430	0.303	0.329	0.026

†Middle recall at matched training steps averaged over two seeds: for this backbone a single best-validation checkpoint is a noisy estimate of the small middle category (the two arms' best checkpoints happen to fall at different steps), so the matched-step average is the controlled statistic; the single best-validation run quoted in §4 reads $0.430 \to 0.237$. The per-step gap between the two rewards ($\overline{|\Delta|}\approx 0.02$) is within the same-reward seed-to-seed spread. ‡This pair shares its own warm-up, a separate supervised run from the backbone comparison of §4 (whose Mistral base is 0.439), hence the different base.

Table 11: F-GRPO focal reweighting vs. the cumulative-cutoff fix on the financial task (middle recall and overall accuracy, before $\to$ after; focal arms share one warm-up model, the fix uses its own). No focal setting restores the middle; only the output-side fix does. Change is the relative change from before.

method	middle (not-enough-info) before $\to$ after	change	overall
plain RLVR	$0.325 \to 0.167$	-49%	0.687
F-GRPO focal (γ=0.5)	$0.325 \to 0.219$	-33%	0.688
F-GRPO focal (γ=1.0)	$0.325 \to 0.044$	-86%	0.668
F-GRPO focal, mean-preserving	$0.325 \to 0.167$	-49%	0.692
cumulative cutoffs (ours)	$0.246 \to \mathbf{0.360}$	$+\mathbf{46}\%$	0.695

D HEAD-TO-HEAD WITH A REWARD-SIDE ORDINAL DECOMPOSITION (ODRPO)

The nearest rival to our *fix* decomposes the ordinal reward, not the output: ODRPO (Patel et al., 2026) rewards a monolithic single-label policy by how many cumulative thresholds $\mathbb{I}[\text{tier} \geq k]$ it matches, i.e. $R = 1 - |\Delta\text{rank}|/(K-1)$. This is exactly the per-threshold score CURB uses, but applied to the label rather than emitted as the output. We run it on the clinical task under conditions identical to CURB (same natural pool, warm-up recipe, and evaluation; each method uses the warm-up matching its output format), at two seeds. Table 12 shows the reward-side grading does *not* rescue the two-sided middle: VUS recall falls to 0.085 on average, statistically indistinguishable from the plain $0/1$ baseline (0.057) and the ordinal-distance reward (0.071), while the output-side decomposition holds it at $0.229 \pm .034$.

The failure is diagnostic of *why* it matters where the fix is applied. Grading the reward *is* informative, since under ODRPO the one-sided tiers improve (e.g. Likely Benign $0.31 \to 0.63$, Pathogenic $0.52 \to 0.70$), but the two-sided middle still competes in a 5-way argmax and bleeds to *both* neighbors: of the truly VUS cases, the fraction correctly kept as VUS falls from 0.21 to 0.07–0.10 while mass moves *both* down (to B/LB, ≈ 0.55) and up (to LP/P, $0.24 \to 0.37$), the same two-directional bleed as the untreated collapse. Reward grading is thus orthogonal to the two-sided structure; only removing the bounded interval from the output space (CURB) prevents the leak.

E RULING OUT ALTERNATIVE EXPLANATIONS FOR THE MIDDLE COLLAPSE

Beyond difficulty, we check four alternative explanations on the saved genetics predictions (natural-pool run, overall accuracy $0.41 \to 0.49$; consistent across three seeds). None tracks the collapse (Table 13). **Difficulty:** the *hardest* category, Likely Pathogenic (base recall 0.150), *survives* ($\to 0.167$) while the easier VUS (0.214) collapses ($\to 0.071$). **Answer length:** VUS answers are mid-length, the two longest categories (LP, P) have opposite outcomes, and the overall correlation of answer

Table 12: Reward-side ordinal decomposition (ODRPO) vs. the output-side decomposition (CURB) on the clinical task: VUS recall, before → after (each method from the warm-up matching its output format; same pool and evaluation). Reward grading collapses the two-sided middle like the $0/1$ baseline; only the output-side fix holds it. Change is the relative change from before.

method	output / reward	VUS recall before → after	change
plain RLVR	monolithic / $0/1$	$0.214 \to 0.057$	-73%
ordinal-distance reward	monolithic / graded ($\lvert\Delta\rvert$)	$0.214 \to 0.071$	-67%
ODRPO (cumulative-threshold)	monolithic / graded (cumulative)	$0.214 \to 0.085$	-60%
cumulative cutoffs (CURB, ours)	threshold vector / per-threshold	$0.277 \to \mathbf{0.229 \pm .034}$	$\mathbf{-17\%}$

length with correctness is 0.03. **Formatting:** every category parses at rate 1.00, so VUS collapses with perfect formatting. **Evidence-set diversity:** VUS has a middling count of distinct evidence combinations; LP and P have the most, with opposite outcomes. **Confusion pattern:** VUS errors are not diffuse: 57% go to the adjacent-lower category and 36% to the adjacent-higher one (only 7% stay), a systematic round-to-the-nearest-decisive shift, not the scattered errors a genuinely harder category would produce.

Table 13: Alternative explanations, on the genetics test set (before → after training). Recall (the collapse) does not track answer length, parse rate, or evidence-set diversity across categories, and the hardest category (LP) survives. Change is the relative change from before.

category	recall before → after	change	parse	distinct combos / n	answer length
Benign	$0.846 \to 0.824$	-3%	1.00	0.32	16
Likely Benign	$0.309 \to 0.426$	$+38\%$	1.00	0.40	18
VUS (middle)	$\mathbf{0.214 \to 0.071}$	$\mathbf{-67\%}$	1.00	0.56	24
Likely Pathogenic	$0.150 \to 0.167$	$+11\%$	1.00	0.82	32
Pathogenic	$0.515 \to 0.755$	$+47\%$	1.00	0.78	31

F Extended related work: adjacent failure modes

Beyond the directly comparable methods discussed in Section 5, several lines describe adjacent phenomena that our failure is distinct from. Reinforcement fine-tuning reduces diversity and narrows the output distribution (Kirk et al., 2024; Yue et al., 2025; Cui et al., 2025), favors common classes under imbalance (Lin et al., 2019), and can lift a headline metric while a rare behavior vanishes (Wei et al., 2026); these concern the distribution of solutions, not a valued output category. It also suppresses honest abstention and hides it in aggregate metrics (Wu et al., 2025; Wei et al., 2025; Kirichenko et al., 2025; Kalai et al., 2025; Yuan et al., 2026), with remedies that train the model to express uncertainty (Damani et al., 2025) or decouple its advantage (Zeng et al., 2026); all treat uncertainty as an *exit* from the label space, whereas ours is an interior, two-sided tier. A related calibration fix rescales decision-token probabilities to cure RLVR overconfidence (Yaldiz et al., 2026), but our failure is a tier that stops being emitted, which recalibration cannot recover. Finally, middle categories collapse under *supervised* cutoff training, repaired by a threshold-gap constraint (Ganjkhanloo et al., 2025); that stems from label scarcity and is separable from ours, which is specific to reward-based training (Section 3).