JUDGING RLVR BOUNDARY COMPARISONS AGAINST THEIR OWN MEASURED NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Whether reinforcement learning with verifiable rewards (RLVR) expands a base model's reasoning boundary is disputed in print. We show that both published answers can be read off one fixed (base, RLVR) pair at every sampling budget we test by moving only sampling temperature and scoring rule; the raw grid counts behind this are not floor-judged, and the base-ahead ones concentrate in chat-template cells at low temperature. Every movement is then judged against measured noise. Of the 30 registered comparisons between the two temperatures this debate uses, 11 are unjudgeable, moving less than the noise floor of the cells they compare; the rest sit on two Minerva cell pairs, and the 11 of them reseeded all clear again. In the sharpest instance the crossover point $k^\star$ moves from 10 to 48 on the same problems, scoring rule and seed. A post hoc observation at $k{=}1$ gives the movement a mechanism: temperature lowers the base arm's pass@1 by a larger factor than the RLVR arm's, and the sign of that asymmetry matches the direction $k^\star$ moves in every cell where it is defined, though a constant rule matches all but one. We turn the measurement into a decision procedure that returns AHEAD, CAUGHT or UNJUDGEABLE for a pair under a protocol, with its false-AHEAD rate set by a null control of two draws from one arm; neither the measured noise floor nor this null-calibrated gate has a counterpart we found in this literature. On seven released pairs the verdict changes with temperature for two of them at the null-calibrated gate: the answer belongs to the protocol as much as to the pair, and it names a finite-budget advantage rather than a support relation. The harness, the noise-floor and lineage instruments, and `judge_boundary.py` are in the supplementary material.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) trains a language model against an automatic checker rather than a learned reward model, and it reliably improves single-sample accuracy on mathematical reasoning (Shao et al., 2024; DeepSeek-AI, 2025). What it does to the model's *reasoning boundary*, the set of problems the model can solve at all given enough attempts, is contested. The standard probe is pass@k (Chen et al., 2021): draw k samples per problem and score a problem solved if any sample is correct. At $k{=}1$ the RLVR model wins; the disagreement is about what happens as k grows. Three answers are in print. Yue et al. (2025) report that the base model catches up and overtakes the RLVR model at large k, and conclude that RLVR reweights a distribution the base model already had. Wu et al. (2025a) report a related finding in more qualified terms: RLVR improves pass@1, but at larger budgets the empirical support it loses outweighs what it gains. A third line reports that with long enough training the boundary does expand, and that earlier negative results measured under-trained checkpoints (Liu et al., 2025a). These studies use different base models, sampling temperatures, k ranges, answer-extraction and answer-checking code, and RLVR checkpoints; any one of those could account for the disagreement. Either the literature disagrees about RLVR, or it disagrees about how to measure RLVR, and distinguishing the two requires holding the model pair fixed and moving only the protocol.

The problem, and the unit that avoids it. Showing that a protocol change moves the crossover point proves nothing on its own, because the crossover point moves anyway: it is a threshold on two noisy curves that approach each other slowly. A claim that "changing X moves $k^\star$ by a

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.

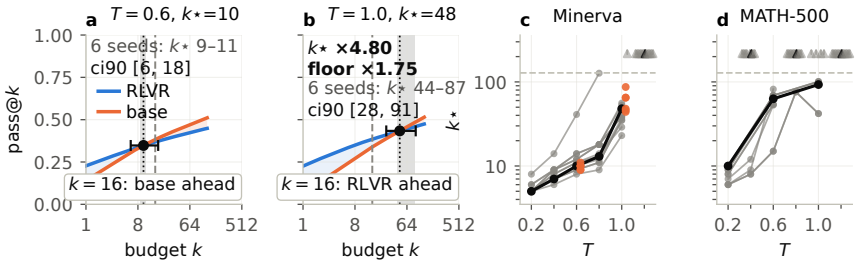

Figure 1: Same models, `minerva` problems, scoring rule and seed; only the sampling temperature differs, and both settings appear in the literature this paper is about. $k^\star$ moves from 10 to 48 against a measured local floor of $1.75\times$. In (a,b) the grey band is the spread of $k^\star$ over the 6 sampling seeds of that cell under this rule; whiskers, 90% bootstrap intervals over problems, [6, 18] and [28, 91]. A reader at budget k=16 (dashed) reports the base model ahead at T=0.6 and behind at T=1.0. (c,d) $k^\star$ against T on two ladders: black, this rule; grey, the other rules; orange, this rule at the 6 seeds of each reseeded cell where its floor is valid (Minerva only); triangles, censored at n=128 (dashed).

factor of r" is empty until r is compared against the movement produced by changing nothing, and to our knowledge no published boundary result reports that quantity. We measure it first, with two instruments (§3), repeated sampling seeds and disjoint half-splits of one run, and make the comparison rather than the number the unit: a comparison is *judgeable* when its effect clears the floor measured on the cells it compares, and *unjudgeable* when those cells supply no valid floor or the effect is smaller than the floor they do. An unjudgeable comparison is not a null result but one its own instrument cannot resolve, and reporting it as either outcome claims more than was measured.

What we find. On a general 7B base and a released RLVR checkpoint trained from it, with the prompt format held fixed at the model's own chat template, 11 of the 30 registered comparisons between the two standard temperatures are unjudgeable, each moving less than the floor of the cells compared. The 19 that clear it all sit on two Minerva cell pairs (one per top-p setting), none of the 8 MATH-500 comparisons clears its floor, and reseeding the Minerva pair changes no verdict (Appendix B). Figure 1 shows the sharpest instance, $k^\star$=10 at T=0.6 and $k^\star$=48 at T=1.0 against a local floor of $1.75\times$, monotone across a 0.2–1.2 ladder under 12/12 scoring rules (the 1.2 rung truncates 49% of samples and enters no verdict). Across the 684 grid points of our protocol sweep, both published conclusions are obtainable at every sampling budget we report from k=8 to k=512. The movement has a mechanism (§5): temperature lowers the base arm's pass@1 and not the RLVR arm's, while neither arm's set of ever-solved problems shrinks.

What the answer is once the protocol is fixed. Unjudgeability is a property of the comparison, not of the question. Under one registered protocol the procedure returns a verdict for every one of the seven released pairs, AHEAD for three; at T=1.0 it returns AHEAD for five, and two pairs change verdict (§4). Two limits scope this: the clearing comparisons all sit under the chat template on a base model, which §4 treats as an evaluation error; and the crossover is a property of the pair.

Contributions. (i) A decision procedure for one boundary comparison, with the criterion stated (§3.1) and shipped as `judge_boundary.py`, which reads two scored records and returns AHEAD, CAUGHT or UNJUDGEABLE on CPU, its gate set on a null control and its power measured on synthetic movement. (ii) A protocol sweep on a fixed model pair showing that both published RLVR-boundary conclusions can be produced from one harness by changing the protocol (§4); whether each study's own prompt, budget and scorer land on its conclusion is not settled here, with a mechanism stated as a prediction and measured on every pair (§5). (iii) Instruments we did not find in this literature: a measured noise floor for $k^\star$, with two silent ways of computing one that invert the conclusion (§3.1); a gate separating "could truncation explain this gap" from "could it explain the gap *moving*" (§3.2); and a weight-space check, with controls, that a checkpoint descends from its claimed base (§3.3). (iv) A reporting standard for boundary claims (§7) and a harness that emits it.

Table 1: The protocol axes. A3–A5 are post hoc over stored completions.

A1 sampling	$T \in \{0.6, 1.0\}$, plus a $\{0.2, 0.4, 0.8, 1.2\}$ ladder on the two benchmarks with the most headroom; top-p 1.0, with a 0.95 probe
A2 prompt	the model's own chat template, or a 4-shot prompt with no template
A3–A5 scoring	every $k \le n$, reported at $k \in \{4, 8, \ldots, 512\}$; `boxed_strict`, `boxed_then_marker`, `boxed_then_last_num` or `last_number` extraction × string, numeric (10^{-6}) or numeric-else-string checking
A6 generation	2048 new tokens in a 4096-token context: a threat to validity, not a free axis (§3.2)

2 SETUP

The quantity. For a problem i with c_i correct completions out of n drawn samples, the unbiased pass@k estimator (Chen et al., 2021) is

$$\widehat{\text{pass@}k} \;=\; \frac{1}{P}\sum_{i=1}^{P}\left[1 - \binom{n-c_i}{k} \Big/ \binom{n}{k}\right], \qquad k \le n. \tag{1}$$

We compute it at every $k \le n$ rather than on a power-of-two ladder, which quantises an argmin into agreement. The *crossover point* is

$$k^\star \;=\; \min\{\, k \le n \;:\; \widehat{\text{pass@}k}^{\,\text{base}} \;\ge\; \widehat{\text{pass@}k}^{\,\text{rlvr}} \,\}, \tag{2}$$

undefined when no such k exists within the budget. It is the quantity the debate turns on, censored at n, and a threshold on two noisy curves, inheriting noise that neither curve's own interval displays.

Models. Our main pair is a general base model and a released RLVR checkpoint trained from it: `Qwen2.5-7B` and `Qwen-2.5-7B-SimpleRL-Zoo` (Zeng et al., 2025), the literature's own population, since the contested results are reported on general bases. six further pairs under the same recipe (Table 3) receive the standard two-temperature slice for §4; `Qwen2.5-Math-1.5B` (Yang et al., 2024) with `Oat-Zero` (Liu et al., 2025b), a different recipe, appears only in Appendix J; Open-Reasoner-Zero-7B (Hu et al., 2025), a second recipe on the `7b` base, only in Appendix O. We train nothing; every checkpoint is a public release, hence §3.3.

The protocol grid. We sweep the six axes of Table 1. A *cell* is one generation coordinate (benchmark, format, T, top-p, n, seed) with both arms; a *grid point* is a cell under one scoring rule, and $k^\star$ is a property of a grid point; a *comparison* is a set of grid points differing in one axis only (Appendix D), so one generation run expands into 12 scoring rules at no extra GPU cost. Benchmarks are AIME 2025 (30 problems, n=512), AMC 2023 (40, n=256), MATH-500 (Lightman et al., 2024) (200, n=128) and Minerva Math (Lewkowycz et al., 2022) (272, n=128), n larger where P is smaller so the samples per cell stay comparable; the main grid is 57 cells × 12 rules = 684 grid points. Sampling uses vLLM (Kwon et al., 2023), and every run records the model path and revision, the tokeniser's chat-template hash, the realised KV-cache size, and the fraction of completions that hit the token limit. A benchmark both models eventually solve everywhere cannot exhibit a crossover, so we report the headroom at the maximum budget (Appendix I): AMC 2023 is near-saturated (5.0%) and read accordingly, and the case study is drawn from Minerva (67.6%).

3 INSTRUMENTING THE QUESTION

Three instruments. The noise floor was registered before the main grid ran and its two corrections followed a pilot verdict; the verdict gate was recalibrated on the null control afterwards (§3.1). The truncation gate's cross-cell form (a between-arm asymmetry moving by more than 3 pp between the compared cells) was written after the 7B grid and its first verdict were seen; the registered rule, truncation within 5 pp between arms on the cells a verdict rests on, holds there (largest gap 3.8 pp), so the verdict does not depend on which is applied. They are a noise floor for $k^\star$, a truncation gate, and a lineage check that every boundary result assumes and none we know of performs.

3.1 A NOISE FLOOR FOR THE CROSSOVER POINT

Instrument 1: seed repeats. We rerun six protocol cells with every other coordinate held fixed, five of them at six sampling seeds and one at three, and record how far $k^\star$ moves (Appendix B).

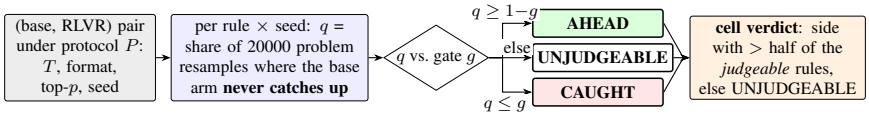

Figure 2: The decision procedure: q is the share of 20000 problem resamples in which the base arm never catches up; a side is read only when q clears $g = 15\%$; the cell takes the majority of its judgeable rules. Figure 7 draws the whole instrument.

The floors differ by a factor of 3.3 across cells: the largest movement anywhere is $5.86\times$ on `math500/chat/0.6/128` (a maximum over its 6 surviving rules), the median across cells and rules $1.52\times$. A rule with $k^\star$ undefined on more than 25% of seeds yields no floor, since a ratio over surviving seeds conditions on definedness. This is the stronger instrument, but it exists only where we ran the extra seeds: only 5 of the 6 yield a floor at all.

Instrument 2: half-splits. For every cell we split each arm's n samples into two disjoint halves (exactly, by hypergeometric sampling of each problem's correct count), recompute both curves on each half and take $k^\star$ on each, using the 95th percentile over 20000 splits. It needs no extra generation and is available in every cell. It is weaker in one convenient direction: the noise is measured at $n/2$ while the ratios are measured at n, so we treat clearing it as conservative; for a censored, non-smooth crossover this is the direction of the sample-size argument rather than a proof.

Two ways to compute a floor that invert the answer. Both cost us a verdict before we caught them (Appendix F). A cell whose $k^\star$ is undefined on more than 25% of splits has no valid floor, and a comparison with no valid floor is *unjudgeable*, not passing; and each comparison is judged against the largest *valid* floor among its group's cells, not the global maximum, which would veto every axis. A set of comparisons passes when more than 50% of those with a valid floor clear it; both corrections were committed before the grid they judge ran.

3.2 Bounding what the generation budget can explain

A completion that hits the token limit is scored wrong, so if the two arms are truncated at different rates, part of the gap between their curves is the budget rather than the models. On the main grid the base arm truncates on 0.4–10.2% of completions outside the ladder's top rung (median 2.8%). Whether the budget could explain the gap *within* a cell has a bound, and the bound is vacuous (Appendix J). Whether it could explain the gap *moving* is what gates a protocol claim, and that has a sharp answer: if the base-minus-RLVR truncation asymmetry is the same in every cell of a comparison, the budget acts equally on both sides in rate. We take that as evidence the budget did not move $k^\star$ rather than a proof: equal rates can truncate different problems. On the sharpest instance the two arms' truncated sets overlap with Jaccard 0.16 at both temperatures, and dropping truncated-and-wrong completions from n instead of scoring them wrong moves $k^\star$ only from 10 to 9 and from 48 to 44: the movement is not which problems truncate. We drop a comparison as truncation-suspect when the spread of that asymmetry across its cells exceeds 3 percentage points; dropped comparisons leave the count rather than failing. The test removes every whole-ladder temperature comparison, since the top rung clips roughly half of the base arm's completions (Figure 1(c)), and leaves every standard-pair comparison standing.

3.3 Checking that the checkpoint is the one being claimed

Every pass@k boundary result attributes the difference between a released RLVR checkpoint and a base model to RLVR, which assumes the checkpoint was trained from that base, usually on no evidence beyond a model card. RLVR is a few policy-gradient updates on a large model, so a correctly paired checkpoint stays close to its base in weight space. We stream both tensor by tensor and compare the median relative Frobenius distance against a *positive control* with declared lineage and a *negative control* pairing tensor i against tensor $i{+}1$ of the same models. On 60 tensors the test pair's median is 0.000390, against 0.000985 and 1.423775 for the controls, $3653\times$ below the mispaired one. The control separates the declared base from an unrelated tensor rather than from a sibling checkpoint of the same base; so the test pair is not an unrelated model; descent from this

Table 2: Movement in $k^\star$ per protocol axis, general 7B pair. *Cells*: distinct cell pairs behind the row (ladder: cell groups; rule row: cells re-scored under every rule), the number to weigh *comparisons* against. *Has floor*: passes the truncation gate and has a valid local floor (§3.1); the criterion runs on that column, and everything outside *clears it* is unjudgeable rather than a pass. The whole-ladder row loses 24 of 35 to the truncation gate alone and its survivors sit on one cell pair (minerva/chat/p0.95/n128); the abstract counts the two-setting row, 19 of 30 on 2 cell pairs. *The 1.2 rung clips 49% of Minerva base-arm completions (RLVR 1.2%). *Class*: pre-registered, whether both settings are defensible.

axis	comparisons	cells	has floor	clears it	median	max	class
temperature*	35	3	11	8	2.3×	2.6×	both-defensible
0.6 vs 1.0 only	30	3	30	19	2.5×	4.8×	both-defensible
top-p	23	2	22	3	1.4×	2.2×	both-defensible
extraction/check	19	19	10	5	1.9×	14.1×	both-defensible
prompt format	26	3	6	4	1.8×	12.1×	one-side-contested

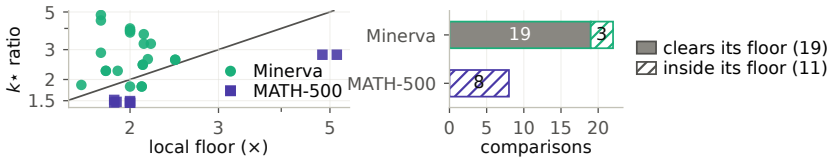

Figure 3: The temperature axis against its own floor, on the 30 standard-pair comparisons. Above the diagonal the movement exceeds the floor of the cells compared. All four axes on one scale: Figure 8.

base rather than from a sibling checkpoint is not tested. Appendix N confirms all six further pairs the same way.

4 Results

RQ1: how far each axis moves the crossover point. Table 2 summarises the four contrast axes: every comparison that moves one axis with all else fixed, its max-over-min $k^\star$ ratio judged against the local floor of the participating cells, with median and max taken over comparisons surviving the truncation gate (Figure 3 draws the temperature rows against their floors; Figure 8, all four axes).

Temperature is the axis that carries the result, and it is a dose-response curve. Figure 1(c,d) shows the axis at six temperatures instead of two: on the benchmark with the most stable floor, $k^\star$ rises monotonically with T under every one of the $12/12$ scoring rules (median span $7.2\times$), censored at the ladder's top rung, where the budget is no longer neutral: it clips 49% of base-arm completions against 1.2% for RLVR. On MATH-500 the same test passes only $2/12$ rules, mostly because $k^\star$ is undefined at an *interior* rung (§5): the clean dose-response is a property of the most stable benchmark rather than of the ladder. Between the two standard settings the movement is a median of $2.5\times$ across 30 comparisons, clearing the local floor in 19 of the 30 that have one, with the truncation asymmetry of every such pair inside the 3 pp tolerance (0 dropped). The other 11 are unjudgeable, each for want of an effect larger than its floor. All 19 that clear are Minerva rows from two cell pairs, one per top-p setting, and none of the 8 MATH-500 rows with a floor clears it (Appendix G); reseeding the default-top-p pair 5 times leaves 11 of its 11 replicated comparisons clearing under every seed (Appendix B). **Between the two temperatures every paper in this debate uses,** 11 **of** 30 **comparisons are unjudgeable, and the** 19 **that clear their floor are all on two Minerva cell pairs.**

The scoring rule and top-p move $k^\star$ but do not clear their own noise. The extraction-and-checking axis has a median ratio of $1.9\times$ and a maximum of $14.1\times$, but only 5 of the 10 comparisons with a floor exceed it; the top-p probe (0.95 against default) behaves the same way, 3 of 22. A single global floor ($5.86\times$, §3.1) is stricter, not looser: it would pass no axis at all, the temperature axis included. The prompt-format axis has the largest ratio of all ($12.1\times$), but 12 of 26 of its comparisons are truncation-suspect and it was classified *one-side-contested* in advance, since a chat template on a base model that never saw one is an evaluation error rather than a protocol variant, so we report it and exclude it from the criterion. **Extraction, top-p and format move $k^\star$ but not past the floors of their own groups: unjudgeable, not null.**

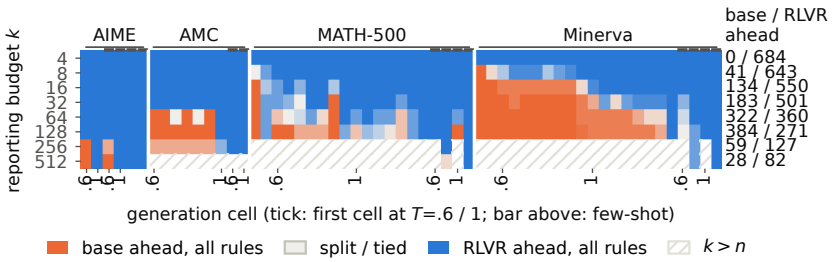

Figure 4: Which conclusion each cell supports at each reporting budget: one column per cell, one row per budget; colour is the share of the cell's 12 rules reporting the base model ahead minus the share reporting RLVR ahead; hatched: $k > n$. Row totals are the counts of Appendix H.

The sharpest instance. Figure 1(a,b) needs no aggregation: same benchmark, format, models, problems and seed, and $k^\star$ moves from 10 to 48 under the same rule, a factor of 4.80 against a local floor of $1.75\times$, in the same direction under every scoring rule (Appendix C). The claim does not rest on the selection: the *smallest* per-rule ratio in the same cell pair, $2.64\times$, also clears the floor, and a bootstrap over the same problems under both temperatures puts $\Delta k^\star$ in $[18, 83]$ and the ratio in $[2.31, 9.80]$ at 95% (censored on one side in 3.4% of draws), an interval that does not use the floor. The direction is the opposite of the naive one (§5).

RQ2: both published conclusions are obtainable at every budget. Figure 4 shows, per sampling budget and per cell, which conclusion the 684 grid points support; Appendix H tabulates the counts. At $k{=}4$ the RLVR model is ahead everywhere; from $k{=}8$ onwards both conclusions are available at every budget we report, the base-model share rising monotonically through $k{=}128$ without ever becoming unanimous (the last two budgets are a different slate, $n{\geq}256$ cells only). The base-ahead points sit where the protocol favours the base model: at $k{=}64$, 221 of 322 are chat cells at $T{\leq}0.6$.

From $k{=}8$ upward the harness alone supplies both published headlines: at $k{=}64$, 322 of 684 grid points say the base model is ahead, 360 say the RLVR model is, and $k^\star$ is undefined at 262.

Where the effect is not: prompt format. Every temperature comparison that clears its floor is inside the chat format, because something more extreme happens under few-shot prompting: no few-shot cell pair has $k^\star$ defined at *both* temperatures, so no ratio exists to judge (AIME 2025 $12/12$ rules at $T{=}0.6$ and $0/12$ at 1.0, MATH-500 the *opposite* way, Minerva $6/12$ to $0/12$; Table 4). The effect is expressed there entirely as benchmark-dependent censoring, so we scope the quantitative claim to the chat protocol, which three of the released pairs cannot run at all. The reading that n was merely too small is refuted by quadrupling the budget: at $n{=}512$ Minerva few-shot falls from $6/12$ defined to $2/12$ at $T{=}0.6$ and stays at $0/12$ at $T{=}1.0$, and since the estimator is unbiased at *fixed* k, the 4 rules that disappear never cross up to $k{=}512$ (Appendix A). The budget-matched control our own Limitations named (the same Minerva chat cell rerun at $n{=}512$, three seeds, registered before it ran) does not thin: $k^\star$ stays defined under $12/12$ rules at *every* seed. **Under few-shot the temperature effect is total censoring, not a smaller ratio: quadrupling n removes crossovers while the budget-matched chat cell keeps $12/12$ at three seeds.**

RQ3: the answer depends on the pair. Everything above holds the model pair fixed. Table 3 (drawn by scale in Figure 10) varies it: seven released pairs under one recipe (SimpleRL-Zoo), each ancestry-verified (§3.3), spanning scale (1.5B–14B), specialisation and family, on the standard two-temperature slice; the three pairs with no chat template are evaluated few-shot (§4).

Existence and location move with scale, and specialisation does not subtract uniformly. At 1.5B the crossover essentially does not exist: $k^\star$ is defined at $6/72$ grid points, all on Minerva. At 14B it is defined at $57/72$ points and arrives as early as $k{=}6$, under the same recipe and protocol, and a six-rung ladder at 14B keeps RLVR ahead at pass@1 by 7.8–22.7 points at every rung (Appendix L). The two math-specialised pairs agree the crossover mostly disappears and disagree where it survives: Qwen2.5-Math-7B defines $k^\star$ only on MATH-500 ($18/24$), DeepSeek-Math-7B almost only on AIME 2025 and Minerva ($1/24$ on MATH-500).

Table 3: The crossover across seven released pairs, one training recipe. *Defined* and min-$k^\star$ use one standard slice for every pair (three shared benchmarks, $T \in \{0.6, 1.0\}$, seed 0, default top-p, the pair's own format), so levels are comparable. *Later*: measurable contrasts where every rule moves *strictly* later at T=1.0, on any benchmark, so its slate is wider. *Verdict at* P_0 is what the released procedure returns for the pair, † marking one whose benchmarks do not all agree. The verdict does not consult the token budget: at P_0 Mistral-7B's RLVR arm truncates 35.6–72.6% of its samples against 2.1–14.0% for its base, and the most truncated base arm (76.5%, qwen1p5b/math500) faces an RLVR arm nearly as truncated. Per-benchmark rows are in Appendix L.

pair	$k^\star$ defined	later at T=1.0	past floor	min $k^\star$	verdict at P_0 (g=0.15)
Qwen2.5-7B (general, main pair)	55/72	9/12	6/8	8	CAUGHT
DeepSeek-Math-7B (specialised)	35/72	0/1	0/1	1	AHEAD†
Llama-3.1-8B (general, few-shot)	36/72	–	–	5	CAUGHT
Qwen2.5-Math-7B (specialised)	18/72	–	–	61	AHEAD†
Mistral-7B-v0.1 (general, few-shot)	48/72	1/2	0/1	1	CAUGHT
Qwen2.5-14B (general)	57/72	2/2	2/2	6	CAUGHT
Qwen2.5-1.5B (general)	6/72	–	–	102	AHEAD

The temperature effect survives the population where it can be measured. Each population contrast gets the same judgement, each rule's ratio against its own two cells' half-split floor. Of the 17 measurable standard-pair contrasts, 8 clear their floors in a majority of the rules with a floor, among them *both* benchmarks of the 14B pair; 5 have no valid floor, and 4 are judged and do not clear. 12 of 17 move later at T=1.0 under *every* rule (the MATH-500 direction is not stable across its 6 seeds), and five further cells of the standard slice, defined under every rule at T=0.6, are censored entirely at T=1.0; the one contrast moving the other way, DeepSeek-Math on AIME 2025, fails its floor with both arms at $\leq$0.2% pass@1 (composition and seed replicates: Appendix L). An eighth pair from a second recipe on the same base (Open-Reasoner-Zero-7B, Appendix O) moves up at both temperatures, as the 7b pair does, but no local floor survives the validity gate, so the registered recipe-transfer test is unevaluable either way.

RQ4: under one fixed protocol the question has an answer, and the answer moves with the protocol. At P_0 (T=0.6, the pair's own format, default top-p, seed 0, the three shared benchmarks, all 12 rules) we ask whether the base arm reaches the RLVR arm within the budget. A rule-seed judgement is CAUGHT when the crossover is present, AHEAD when the RLVR arm stays ahead through the budget, each in all but a fraction g of 20000 problem resamples, and UNJUDGEABLE otherwise; a pair's verdict is the majority of its 36 rule-benchmark judgements (Figure 2). AHEAD names a finite-budget advantage and CAUGHT its absence within n, not a support relation. The registered gate was g=0.25. A null control run after P_0 was registered (two draws of one base arm differing only in the sampling seed, 45 comparisons on 3 cells) fixes what a verdict is worth: two curves from one distribution meet at some k, so CAUGHT is the procedure's reading of zero signal, and at the registered gate the null still returned AHEAD 5 times (11.1%). We therefore tightened the gate to g=0.15, where the null returns AHEAD 2 times (4.4%), CAUGHT 35 and UNJUDGEABLE 8; its power, on synthetic movement, is a property of the cell (Appendix M). At that gate all seven pairs get a verdict, the same under three bootstrap seeds: four CAUGHT, three AHEAD (DeepSeek-Math-7B, Qwen2.5-Math-7B and Qwen2.5-1.5B), none unjudgeable (Table 3). The registered gate's AHEAD pairs keep it; the third, DeepSeek-Math-7B, was that gate's narrowest call (15 CAUGHT to 12 AHEAD) and turns AHEAD only because tightening leaves 24 of its judgements unjudgeable: a gate-sensitive verdict. Moving one axis, to T=1.0 with everything else held, returns two CAUGHT and five AHEAD; two pairs (Llama-3.1-8B and Mistral-7B-v0.1) change verdict, in the direction §5 predicts.

AHEAD carries the evidence: three pairs at T=0.6, five at T=1.0, two change between the two.

5 What Moves the Crossover: Mechanism and Failure Cases

§4 measured how far $k^\star$ moves; we now state what moves it as a prediction on a quantity every cell has, and test it on all seven pairs. Unlike the gate of §3, it is post hoc, designed after the grid ran.

On the main pair, temperature moves the k=1 end of the base curve and leaves its support. Figure 5 follows the ladders of Figure 1(c,d) one quantity at a time. Between T=0.2 and T=1.0 the

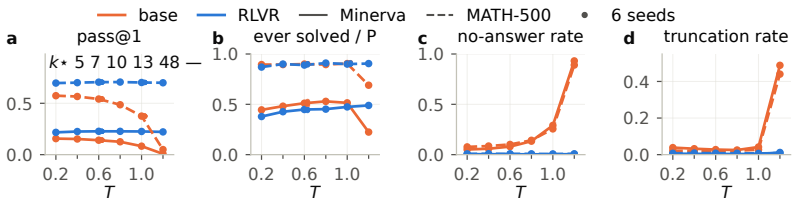

Figure 5: The main pair's ladder under the case-study rule, Minerva (solid) and MATH-500 (dashed): each arm's pass@1, the fraction of problems it ever solves within n, its no-answer rate and its truncation rate, against T. Dots are 6-seed values; the Minerva $k^\star$ runs along the top of (a).

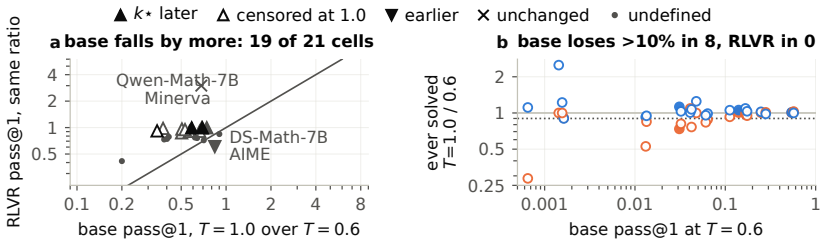

Figure 6: The prediction on every cell of the standard slice across seven pairs. (a) The base arm's pass@1 at T=1.0 over its value at T=0.6, against the same ratio for RLVR; below the diagonal the base arm fell by more, and the marker is what $k^\star$ did. (b) Each arm's ever-solved count at T=1.0 over its count at T=0.6, against the base arm's pass@1 at T=0.6; dotted line, the 10% support-loss threshold. Filled, the main pair; hollow, the other pairs; in (b) orange is base, blue RLVR.

base arm's pass@1 on Minerva falls from 15.6% to 8.3% (47%), the RLVR arm's from 21.7% to 22.6%, within 4% of itself. Ever-solved within n=128 does not fall: base 121 → 140 of 272, RLVR 103 → 129. The base arm's no-answer rate rises from 5% to 29%, the RLVR arm's stays under 0.5%, and truncation is flat on both arms until the top rung. Over the same range $k^\star$ moves from 5 to 48; MATH-500 gives the same picture (Appendix C). Extraction accounts for part of the drop but not all of it: under the lenient `last_number` rule the base arm's no-answer rate at T=1.0 is 3% and its pass@1 still falls by 44% against 2% for RLVR, $k^\star$ moving 6 to 39; seed noise on the same cell is smaller (Appendix C). The crossover moves later because the base curve starts lower at k=1 and reaches the same value at large k rather than because its reachable set shrinks.

Prediction, and its measurement across seven pairs. If that is the mechanism, the direction of the movement in $k^\star$ between T=0.6 and T=1.0 is fixed by one quantity available in every cell: whether the base arm's pass@1 falls by a larger factor than the RLVR arm's. Where it does, $k^\star$ moves later or leaves the budget; where it does not, it moves earlier. A boundary reading predicts instead that $k^\star$ follows the base arm's ever-solved count. The further pairs test it (Figure 6(a), the 21 (pair, benchmark) cells of the standard slice). The base arm falls by more in 19 of 21 cells under the case-study rule and 17 under a majority of the 12 rules; the exceptions are Qwen2.5-Math-7B on Minerva and DeepSeek-Math-7B on AIME 2025. $k^\star$ is defined at T=0.6 in 13 cells, 12 of them with a direction (4 later, 7 censored at T=1.0, 1 earlier) and 1 without (Mistral-7B on AIME 2025: $k^\star$=1 at both temperatures). The direction matches the sign of the pass@1 asymmetry in 12 of 12. A rule answering *later* unconditionally scores 11 of 12, so the rule's discriminating power depends on the one cell that moves earlier, DeepSeek-Math-7B on AIME 2025, where the base arm's pass@1 falls by less and both arms score at most 0.2% at pass@1. Fisher's exact one-sided p on that 2×2 is 0.083: one cell cannot separate the rules, so this is an observation rather than a confirmed prediction.

Where the boundary reading does apply. Figure 6(b) plots each arm's ever-solved count at T=1.0 over its count at T=0.6. The RLVR arm loses more than 10% of its ever-solved problems in 0 of 21 cells, the base arm in 8, every one a cell where its pass@1 at T=0.6 is at most 6.4%, and every cell above it keeps its support. Of those 8, 3 are censored at T=1.0 and the rest had no $k^\star$ at either temperature, while the 4 cells where $k^\star$ moves later all keep the base arm's support. Two regimes, then: k=1 compression where the base arm scores a few percent pass@1 or more, and support loss below that, where $k^\star$ is mostly undefined to begin with. The second is the boundary reading.

Where the instrument fails. Two cases in our own grid. The MATH-500 ladder: 8 of 12 rules lose $k^\star$ at an interior rung (Figure 1(d)) because the two curves run within a few points of each other at large k (Figure 5(b)); $k^\star$ is then the threshold of two curves that touch and separate, and its floor is the grid's widest ($5.86\times$ on `math500/chat/0.6/128`). DeepSeek-Math-7B on AIME 2025 (§4): a $188.50\times$ movement its own floor rejects. The instrument's own failures are in §3.1, §3.2 and Appendix B. Negative results: the scoring-rule and top-p axes (Table 2), few-shot prompting, which yields no ratio (§4), and two abandoned grids (Appendix K). **Temperature moves $k^\star$ by compressing the base arm's curve at $k{=}1$, not by shrinking its support: the compression's sign matches the direction in 12 of 12 cells, and support shrinks only where pass@1 is already at most 6.4%.**

6 Related Work

The RLVR boundary debate. Yue et al. (2025) report the base model overtaking at large k and conclude that RLVR redistributes mass within a boundary the base already had; Wu et al. (2025a) reach a compatible conclusion, RLVR anchored to the base distribution and shedding more support than it gains; Liu et al. (2025a) report the opposite under much longer training; later work reads the two as phases of one dynamic (Yao et al., 2025; Yuan et al., 2026) and localizes pass@k inversion to boundary prompts (Zhou, 2026; Yu, 2025).

pass@k as a boundary probe. The unbiased estimator is from Chen et al. (2021), where it prices sampling many candidate programs under a test oracle; the boundary literature reads it as a capability claim, adding the assumption that pass@k at large k approximates the policy's support.

Evaluation sensitivity. That results move under harness choices is established outside the RLVR setting: formatting choices span accuracy ranges comparable to gaps between models (Sclar et al., 2024), single-prompt evaluation is not identifiable (Mizrahi et al., 2024), and harness-level decisions silently change published numbers (Biderman et al., 2024; Wu et al., 2025b). We add what a threshold statistic needs and a mean does not: a measured floor. Aggregation over an unsettled configuration (Xu et al., 2026; Xiao et al., 2025) has no crossover to lose, since protocols that disagree about whether $k^\star$ exists have no mean; axes A4–A5 recur wherever an LLM's output is machine-scored (Xiao et al., 2026b;a), and here they carry the result. No published boundary result we know reports how far $k^\star$ moves under a change of seed (§3.1) or verifies in weight space that a checkpoint descends from its claimed base (§3.3).

7 What a Boundary Claim Should Report

On one fixed model pair, with no training, the sampling temperature alone moves the crossover point past the largest valid floor of its comparison group. It does so by compressing the base arm's curve at $k{=}1$ (§5), and both published conclusions about the RLVR boundary can be obtained at every budget from $k{=}8$ to $k{=}512$. Of the 30 registered comparisons that carry this disagreement, 11 are unjudgeable at the budgets they ran at. At one stated protocol all seven released pairs return a verdict, and the temperature alone changes two of them (§4). A boundary claim should carry **a measured noise floor for the reported statistic**, neither conditioned on definedness nor set by one unstable cell (§3.1); **the full protocol coordinate**: temperature, top-p, format and template hash, scoring rule, budget, n, revision; **the censoring**, since "no crossover within n" is a bound, not a value; **the truncation asymmetry between arms** (§3.2); **the population**, since existence and location are properties of a pair; and **lineage evidence** (§3.3).

Limitations. The dose-response curve is the main pair's, inside the chat format. The sharpest instance borrows its floor, its $T{=}1.0$ cell's own being invalid under the definedness gate at six seeds as at one (§3.1); the paired interval on $\Delta k^\star$ does not. We sweep the protocol, not the training; §5, the null calibration that replaced the registered gate and the synthetic power curve are post hoc. Two results would change our mind, a movement small against a measured floor or a chat cell dissolving at $n{=}512$; we ran the second (§4).

AI USE STATEMENT

Generative AI tools were used substantially in the preparation of this work, and we disclose the uses that fall under the required categories. LLMs helped develop the conceptual framework and refine the hypothesis this paper tests, gave feedback on research methodology and experiment design, implemented the evaluation harness (the sampling driver, the re-scoring pass, the pass@k and crossover code, the noise-floor instruments, the lineage check and the macro generator) and assisted in interpreting results. Under the recommended categories they were additionally used for code creation, literature search, brainstorming, and editing the text of this paper.

We have *not* used generative AI tools to produce, select or filter any experimental result. Every empirical figure in this paper is emitted by `eval/gen_macros.py` from the result artifacts and is inserted into the text as a macro annotated with the artifact it came from; protocol constants such as T, the token budget and the tolerances are written literally, and nothing else is, so a number that no longer follows from the data cannot survive a rebuild. We state this having found the guarantee broken: an earlier draft carried four hand-typed figures, three of them stale, and the artifact tests in `tests/` now fail if a quantity is formatted from a rounded copy rather than from its integer counts. The research question, the claims, the pre-registered decision rule and the conclusions are the authors'.

We have reviewed all AI-assisted work and take responsibility for the final content, including every claim, number and artifact. Two aspects of this particular paper deserve explicit statement because they concern exactly the failure mode a reader should worry about. First, the paper's gate, what counts as a real movement in $k^\star$, was written down before the grid it judges was generated, and two corrections to the noise-floor computation (§3.1) were made before that grid ran; the cross-cell form of the truncation gate was not, and §3 reports the registered rule beside it. The artifact ships the pre-registration and the dated decision records that amended it, each saying what changed and why; it does not ship the repository history that would date them independently, so this ordering is stated here rather than checkable from the package. Second, we report an axis that fails its own criterion (the scoring-rule axis, §4) and two settings in which the effect is absent (§4, §4), because a protocol-sensitivity paper that reported only the sensitive axes would be an instance of the problem it describes.

REPRODUCIBILITY STATEMENT

The artifact contains the sampling driver, the re-scoring pass, the analysis code producing every table, the per-problem scored records and provenance file for every one of the 57 cells in the grid, the decision procedure as a command (`judge_boundary.py --base X --rlvr Y --protocol P --g 0.15`, which needs only two scored records and carries its own self-test), and a driver (`reproduce.sh`) that rebuilds each analysis product the paper's numbers come from, at both gates, and compares it byte for byte with the shipped copy; we checked this claim by deleting the analysis outputs and rebuilding them from the records alone. The 3 cells of the budget-matched control (§4) ship the same records under their own directory and their own grid file, kept apart from the main grid so that adding them cannot move a number the grid already reported; that the main grid rebuilds byte-identically from its own records with the control present is part of the same check. Axes A3–A5 (budget, extraction, verification) are post hoc over stored completions, so every scoring result in the paper replays on CPU from the cache with no GPU and no network, under the pinned analysis environment shipped with the artifact. That pin is necessary rather than precautionary: replaying the same records under a newer `numpy` reproduces most grids exactly but moves a few resampled quantities by one unit, including one bootstrap interval endpoint, so we state the version the numbers were made under instead of implying the replay is version-free. Left out are the raw completion text (gigabytes, with no information the scored records lack), the lab queue scripts, the test suite and the paper sources; `MANIFEST.json` lists each with its reason. The generation runs behind the cache record their full protocol coordinate in a provenance file (model path, tokeniser chat-template hash, temperature, top-p, n, token budget, realised KV-cache size, free GPU memory at initialisation, and the fraction of completions that hit the token limit) and each was run with a fixed sampling seed. We note the limit of that seeding honestly: vLLM's batching is not bitwise reproducible across different batch compositions on a shared machine, which is one reason §3.1 measures the seed-to-seed spread instead of assuming it is zero. The lineage check of §3.3 runs on CPU from the two published checkpoints.

Ethics Statement

This work evaluates publicly released model checkpoints on public mathematics benchmarks. It involves no human subjects, no new data collection, and no personally identifying information. We see one misreading worth guarding against: this paper shows that a published disagreement is reproducible from evaluation choices, which is a statement about measurement and not a claim that any particular published result is wrong or was obtained in bad faith. The protocol settings we sweep are all defensible choices that appear in the literature, which is precisely what makes the sensitivity a problem for the field rather than for any individual paper.

References

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, et al. Lessons from the trenches on reproducible evaluation of language models. arXiv:2405.14782, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, et al. Evaluating large language models trained on code. arXiv:2107.03374, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, et al. Training verifiers to solve math word problems. arXiv:2110.14168, 2021.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638, 2025.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. arXiv:2503.24290, 2025. Source of the Open-Reasoner-Zero-7B checkpoint.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, et al. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, 2022.

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 *International Conference on Learning Representations*, 2024. arXiv:2305.20050.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. arXiv:2505.24864, 2025a.

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:2503.20783, 2025b. Source of the Oat-Zero checkpoint.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300, 2024.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why RLVR may or may not escape its origin. arXiv:2507.14843, 2025a.

Yuheng Wu, Azalia Mirhoseini, and Thierry Tambe. On the role of temperature sampling in test-time scaling. arXiv:2510.02611, 2025b.

Yifan Xiao, Zhongjun Wang, Tiantian Yan, Weiwei Li, Dazhong Wu, and Jie Ma. Time series anomaly detection: A hybrid deep learning framework with adaptive thresholding for electricity industry data. In *IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)*, 2025. DOI 10.1109/CSCLOUD66326.2025.00046.

Yifan Xiao, Shijie Li, and Yuhao Ge. RunbookFX: Type- and effect-safe LLM synthesis for executable incident diagnosis and mitigation. *Proceedings of the ACM on Programming Languages*, (ICFP), 2026a. DOI 10.1145/3828675.

Yifan Xiao, Shijie Li, and Yu Huang. CALM-TS: Risk-controlled LLM labeling for time-series via calibrated selective gating. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026b. DOI 10.1145/3770855.3817859.

Dexuan Xu, Yanyuan Chen, Zhongyan Chai, Yifan Xiao, Yandong Yan, Weiping Ding, Hanpin Wang, Zhi Jin, Wenpin Jiao, Weihua Yue, Hang Li, and Yu Huang. Knowledge fusion in deep learning-based medical vision-language models: A review. *Information Fusion*, 125:103455, 2026. doi: 10.1016/j.inffus.2025.103455.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. arXiv:2409.12122, 2024.

Xinhao Yao, Lu Yu, Xiaolin Hu, Fengwei Teng, Qing Cui, Jun Zhou, and Yong Liu. The debate on rlvr reasoning capability boundary: Shrinkage, expansion, or both? a two-stage dynamic view. arXiv:2510.04028, 2025.

Yang Yu. Pass@k metric for rlvr: A diagnostic tool of exploration, but not an objective. arXiv:2511.16231, 2025.

Suqin Yuan, Jinkun Chen, Jiyang Zheng, Muyang Li, Lei Feng, Dadong Wang, Tao Xiang, and Tongliang Liu. Understanding diversity collapse in rlvr via the lens of overtraining. arXiv:2606.15455, 2026.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, 2025. Oral; arXiv:2504.13837.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. SimpleRL-Zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2503.18892; source of the RLVR checkpoint evaluated here.

Todd Zhou. When rlvr shrinks the reasoning boundary: Diagnosing pass@k inversion. arXiv:2607.20543, 2026.

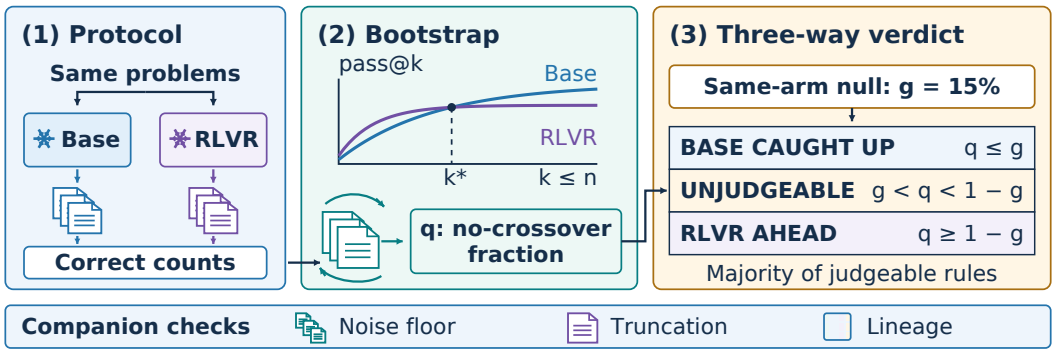

Figure 7: The instrument for one model pair under a matched protocol. Both arms answer the same problems; resampling problems gives q, the share of resamples in which the base arm never catches up within the budget n. The same-arm null sets $g = 15\%$, and each scoring rule returns CAUGHT, AHEAD or UNJUDGEABLE; the cell takes the strict majority of its judgeable rules and is UNJUDGEABLE without one. These are finite-budget verdicts, not support claims. Noise floors, truncation and lineage are companion measurements, not further gates. Curves are schematic.

A THE FULL GRID

Table 4 lists every generation cell: its protocol coordinate, the fraction of completions in each arm that ended by hitting the token limit, how many of the 12 scoring rules yield a defined $k^\star$, and the range $k^\star$ takes across those rules.

The grid is deliberately unbalanced in seeds, and Table 5 shows it: five coordinates carry six seeds each (the floor cells of §3.1, chosen before any of those runs), one carries three, and the remaining twenty-four carry one. The rows above are all at seed 0; an extremum over such a grid is not an extremum over a balanced design, and a reader comparing coordinates should check their seed depth before reading a difference as protocol sensitivity rather than sampling noise. The deep coordinates are deep because measuring the seed spread was the point; nothing else was resampled to match.

Three patterns in this table are worth naming. First, on AIME 2025 the temperature change does not relocate the crossover so much as abolish it: $k^\star$ is defined under all 12 rules at T=0.6 and under none at T=1.0, in both prompt formats. A ratio cannot be formed from that, so these cells contribute no comparison to Table 2: the axis summary is, if anything, understating the temperature axis by excluding the cells where its effect is total. Second, truncation rates are strongly format-dependent and only weakly arm-dependent under few-shot prompting, which is what the asymmetry test of §3.2 is built to exploit. Third, the two Minerva few-shot budgets are the experiment of §4: at T=0.6 definedness falls from $6/12$ at n=128 to $2/12$ at n=512 rather than rising, which is what refutes the reading that the few-shot crossover was merely out of budget: since the estimator is unbiased at *fixed* k, more samples cannot remove a real crossover below k=128 and can only add them above it, so a fall has one reading: the 4 rules that disappear never cross up to k=512, and the 2 survivors move from $k^\star$=22 to 71. That is half the comparison, because the chat cells it is set against ran at n=128, so a format reading was still confounded with a budget one. The starred rows close that comparison at a matched budget: Minerva chat at n=512 keeps $12/12$ at each of 3 seeds ($36/36$ points, of which 0 are budget-edge ties), so the format contrast at n=512 can be read down the column rather than across two budgets. Nor is that definedness a near miss: the latest of those $k^\star$ is 25, 4.9% of the budget, and the seed-repeat floor of the same cell *tightens* from $2.41\times$ at n=128 (Table 5) to $1.14\times$ here, over $12/12$ rules, so the larger budget makes $k^\star$ more stable in this cell rather than less, which is the opposite of what a budget-limited crossover would do. It remains true that a column-wise reading *within* the grid crosses budgets wherever a benchmark appears at two, and that the format comparisons entering Table 2 are all within one n: the control cells enter no axis, no floor, no flip count and no seed floor, only this table and §4.

B THE SEED-REPEAT FLOORS IN FULL

Table 5 gives the raw evidence behind the seed floors: each of 5 protocol cells generated at 6 sampling seeds with every other coordinate held fixed, and $k^\star$ recomputed under each rule. The

Table 4: The 30 registered generation cells at sampling seed 0 plus the control rows; the 27 seed replicates of the 6 reseeded cells are in Table 5. The rows marked * are the 3 cells of the budget-matched chat control (§4), not part of the grid and counted in no figure quoting 684 grid points. "$k^\star$ defined" counts the scoring rules under which the base curve reaches the RLVR curve within n samples, strictly: a dagger marks cells where $k^\star$ landed at $k{=}n$ with the two curves numerically equal, which is a tie and is counted as censored (§4). Since $k^\star$ is censored at n, the definedness column compares only within a budget; Minerva appears at two.

benchmark	format	T	top-p	seed	n	trunc. b/R (%)	$k^\star$ def.	$k^\star$ range
aime25	chat	0.6	1.0	0	512	10.2 / 2.0	12/12	164–214
aime25	chat	1.0	1.0	0	512	4.6 / 1.4	0/12	–
aime25	fewshot4	0.6	1.0	0	512	8.2 / 8.7	12/12	202–258
aime25	fewshot4	1.0	1.0	0	512	1.8 / 2.6	0/12	–
amc23	chat	0.6	1.0	0	256	3.4 / 1.8	12/12	56–60
amc23	chat	1.0	1.0	0	256	3.4 / 1.6	0/12$^{\dagger 6}$	–
amc23	fewshot4	0.6	1.0	0	256	3.7 / 4.2	0/12	–
amc23	fewshot4	1.0	1.0	0	256	0.7 / 1.1	0/12	–
math500	chat	0.2	1.0	0	128	2.2 / 0.7	12/12	6–10
math500	chat	0.4	1.0	0	128	2.1 / 0.9	4/12	8–12
math500	chat	0.6	1.0	0	128	1.9 / 0.8	12/12	15–82
math500	chat	0.8	1.0	0	128	2.2 / 0.8	2/12$^{\dagger 1}$	76–76
math500	chat	1.0	1.0	0	128	2.7 / 0.7	8/12	42–101
math500	chat	1.2	1.0	0	128	44.0 / 0.8	0/12	–
math500	fewshot4	0.6	1.0	0	128	2.2 / 2.5	0/12	–
math500	fewshot4	0.6	1.0	0	512	2.3 / 2.6	2/12$^{\dagger 10}$	291–291
math500	fewshot4	1.0	1.0	0	128	0.4 / 0.7	12/12	60–112
math500	fewshot4	1.0	1.0	0	512	0.4 / 0.8	0/12	–
minerva	chat	0.2	1.0	0	128	3.8 / 0.5	12/12	5–8
minerva	chat	0.4	1.0	0	128	3.2 / 0.5	12/12	6–14
minerva	chat	0.6	1.0	0	128	2.8 / 0.4	12/12	8–41
minerva	chat	0.6	0.95	0	128	2.7 / 0.4	12/12	7–30
minerva	chat	0.8	1.0	0	128	2.5 / 0.4	12/12	9–127
minerva	chat	1.0	1.0	0	128	4.2 / 0.3	11/12	23–56
minerva	chat	1.0	0.95	0	128	3.7 / 0.3	11/12	13–34
minerva	chat	1.2	1.0	0	128	48.8 / 1.2	0/12	–
minerva	fewshot4	0.6	1.0	0	128	8.5 / 8.6	6/12	22–109
minerva	fewshot4	0.6	1.0	0	512	8.3 / 8.5	2/12	71–71
minerva	fewshot4	1.0	1.0	0	128	1.5 / 2.3	0/12	–
minerva	fewshot4	1.0	1.0	0	512	1.6 / 2.4	0/12	–
minerva*	chat	0.6	1.0	0	512	2.8 / 0.4	12/12	7–25
minerva*	chat	0.6	1.0	1	512	2.8 / 0.4	12/12	7–25
minerva*	chat	0.6	1.0	2	512	2.8 / 0.4	12/12	7–22

high-temperature end of the Minerva chat ladder ($T{=}1.0$, $n{=}128$), which the first version of this grid carried at one seed, was extended to six under a reading rule fixed before the runs: the global floor was to be updated if it rose and reported unchanged otherwise. It did not rise; the floor stays at $5.86\times$ on `math500/chat/0.6/128`. The seeds enter no comparison: Table 2 is judged over the seed-0 grid (`DECISION_20260906`), and the pair replicates instead. Of the 19 seed-0 comparisons that clear the floor, 11 have replicates (the rest are top-p probe rows, never reseeded), and 11 of those clear under every one of the 5 added seeds, 53 of 53 replicate comparisons in all.

The spread across cells ($1.80\times$ on `amc23/chat/0.6/256` against $5.86\times$ on `math500/chat/0.6/128`) is why the verdict judges each comparison against the floor of its own cells: one global threshold would either be set by the noisiest cell and veto clean findings, or be set by the cleanest and wave through noise. Within every cell the maximum is set by the lenient `last_number` extractions, whose $k^\star$ is genuinely less stable; we keep the maximum rather than trimming it.

Table 5: Seed-repeat floors, the 5 reseeded cells that yield a floor. Per cell the floor is the maximum ratio over rules; the global maximum is $5.86\times$ (on `math500/chat/0.6/128`) and the median across all cells and rules is $1.52\times$. A dash marks a seed at which $k^\star$ is undefined; rules undefined on more than 25% of seeds are excluded by the validity gate of §3.1 and not listed.

cell	extraction	check	$k^\star$ per seed	ratio
amc23/chat/0.6/256	boxed_strict	exact_str	60, 55, 63, 47, 60, 63	1.34
	boxed_strict	numeric	60, 55, 63, 47, 60, 63	1.34
	boxed_strict	numeric_or_str	60, 55, 63, 47, 60, 63	1.34
	boxed_then_last_num	exact_str	56, 51, 66, 39, 66, 63	1.69
	boxed_then_last_num	numeric	56, 51, 66, 39, 66, 63	1.69
	boxed_then_last_num	numeric_or_str	56, 51, 66, 39, 66, 63	1.69
	boxed_then_marker	exact_str	60, 55, 63, 47, 60, 63	1.34
	boxed_then_marker	numeric	60, 55, 63, 47, 60, 63	1.34
	boxed_then_marker	numeric_or_str	60, 55, 63, 47, 60, 63	1.34
	last_number	exact_str	60, 51, 67, 40, 72, 63	1.80
	last_number	numeric	60, 51, 67, 40, 72, 63	1.80
	last_number	numeric_or_str	60, 51, 67, 40, 72, 63	1.80
math500/chat/0.6/128	boxed_strict	numeric	54, 43, 19, 75, 56, 18	4.17
	boxed_then_last_num	numeric	53, 45, 17, 77, 57, 16	4.81
	boxed_then_marker	numeric	54, 43, 19, 75, 56, 18	4.17
	last_number	exact_str	82, 21, 16, 78, 50, 14	5.86
	last_number	numeric	15, 21, 14, 60, 34, 13	4.62
	last_number	numeric_or_str	15, 21, 14, 60, 34, 13	4.62
math500/chat/1.0/128	boxed_then_last_num	exact_str	94, 78, 91, 89, 51, –	1.84
	last_number	numeric	42, 71, 47, 39, 27, 41	2.63
	last_number	numeric_or_str	42, 71, 47, 39, 27, 41	2.63
minerva/chat/0.6/128	boxed_strict	exact_str	11, 11, 10, 9, 10, 10	1.22
	boxed_strict	numeric	14, 15, 13, 13, 14, 14	1.15
	boxed_strict	numeric_or_str	9, 9, 9, 8, 9, 9	1.13
	boxed_then_last_num	exact_str	10, 9, 9, 8, 8, 9	1.25
	boxed_then_last_num	numeric	11, 11, 10, 9, 10, 10	1.22
	boxed_then_last_num	numeric_or_str	8, 8, 7, 7, 7, 7	1.14
	boxed_then_marker	exact_str	10, 11, 10, 9, 9, 10	1.22
	boxed_then_marker	numeric	14, 15, 12, 12, 14, 14	1.25
	boxed_then_marker	numeric_or_str	9, 9, 9, 8, 9, 9	1.13
	last_number	exact_str	41, 32, 20, 17, 21, 22	2.41
	last_number	numeric	12, 12, 12, 11, 12, 12	1.09
	last_number	numeric_or_str	12, 12, 12, 11, 12, 12	1.09
minerva/chat/1.0/128	boxed_strict	exact_str	49, 46, 44, 65, 65, 87	1.98
	boxed_strict	numeric	56, 50, 55, 59, 74, –	1.48
	boxed_strict	numeric_or_str	35, 31, 33, 37, 40, 42	1.35
	boxed_then_last_num	exact_str	37, 31, 35, 47, 41, 64	2.06
	boxed_then_last_num	numeric	29, 23, 31, 33, 33, 36	1.57
	boxed_then_last_num	numeric_or_str	23, 19, 23, 25, 24, 26	1.37
	boxed_then_marker	exact_str	48, 46, 44, 65, 65, 87	1.98
	boxed_then_marker	numeric	53, 50, 55, 59, 74, –	1.48
	boxed_then_marker	numeric_or_str	35, 31, 33, 37, 40, 42	1.35
	last_number	numeric	39, 28, 45, 44, 40, 54	1.93
	last_number	numeric_or_str	39, 28, 45, 44, 40, 54	1.93

C THE CASE STUDY, RULE BY RULE

Numbers the body summarises. On MATH-500 between $T{=}0.2$ and $T{=}1.0$ the base arm's pass@1 falls $57.4\% \to 37.5\%$, the RLVR arm's $69.7\% \to 70.3\%$, ever-solved flat. Over the 6 seeds of the Minerva $T{=}0.6$ cell, base pass@1 spans 14.1–14.2% and ever-solved 134–144; the support-loss threshold of §5 exceeds this spread. The ladder's top rung clips 49% of base-arm completions on Minerva. Llama-3.1-8B defines $k^\star$ under 12/12 at $T{=}0.6$, 0/12 at $T{=}1.0$ rules on Minerva and 12/12 at $T{=}0.6$, 0/12 at $T{=}1.0$ on MATH-500.

Table 6 expands Figure 1: the same cell under every scoring rule, at the two standard temperatures.

Table 6: One cell (minerva, chat format), every scoring rule, T=0.6 versus T=1.0. *ci*90 is the 90% bootstrap interval for $k^\star$ over problems, arms paired by problem index; it is disjoint between the two temperatures in 7 of the 11 rules with both endpoints defined. *floor* is the local noise floor for the group's cells; a dash marks a temperature at which $k^\star$ is undefined within the budget, itself an instance of the sensitivity being measured.

extraction	check	$k^\star$@T=0.6	ci90	$k^\star$@T=1.0	ci90	ratio	floor
boxed_strict	exact_str	11	[6, 20]	49	[30, 94]	4.45	1.75
boxed_strict	numeric	14	[7, 30]	56	[32, 104]	4.00	2.00
boxed_strict	numeric_or_str	9	[6, 17]	35	[22, 57]	3.89	2.00
boxed_then_last_num	exact_str	10	[6, 19]	37	[22, 82]	3.70	2.13
boxed_then_last_num	numeric	11	[6, 20]	29	[17, 52]	2.64	2.17
boxed_then_last_num	numeric_or_str	8	[5, 14]	23	[15, 37]	2.88	1.75
boxed_then_marker	exact_str	10	[6, 18]	48	[28, 91]	4.80	1.75
boxed_then_marker	numeric	14	[8, 31]	53	[31, 98]	3.79	2.00
boxed_then_marker	numeric_or_str	9	[5, 16]	35	[22, 55]	3.89	2.00
last_number	exact_str	41	[10, 127]	–	–	–	2.70
last_number	numeric	12	[7, 33]	39	[21, 83]	3.25	2.20
last_number	numeric_or_str	12	[7, 32]	39	[22, 78]	3.25	2.13

D THE TERMS

The protocol axes and their settings are Table 1 in §2, and the extraction and verification rules of axes A4–A5 are written out in Appendix E. Table 7 defines the terms the paper's counts are made in.

Table 7: The terms the paper's counts are made in.

term	meaning
cell	one generation coordinate (benchmark, format, T, top-p, n, seed), both arms
grid point	a cell under one of the 12 scoring rules; $k^\star$ is a property of a grid point
comparison	grid points that differ in one axis only; its ratio is max-over-min $k^\star$
local floor	the largest *valid* half-split floor among the cells of the comparison's group (§3.1)
judgeable	a comparison whose effect clears the local floor. A comparison with no local floor, and one whose effect is smaller than the floor it has, are both unjudgeable, and unjudgeable is not a pass
verdict	AHEAD, CAUGHT or UNJUDGEABLE, returned for a (pair, protocol) by `judge_boundary.py`; AHEAD is a finite-budget advantage, CAUGHT its absence within n
standard slice	the three shared benchmarks $\times$ $T \in \{0.6, 1.0\}$, seed 0, default top-p, one format

E SCORING RULES

Extraction rules, applied to the completion text:

`boxed_strict` the content of the last `\boxed{}`; no box means no answer.

`boxed_then_marker` the box if present, else the text following "answer is" or "answer:".

`boxed_then_last_num` the box if present, else the last number in the completion.

`last_number` the last number anywhere in the completion, the lenient default of several GSM8K-era harnesses.

Verification rules, applied to the extracted string against the reference:

`exact_str` normalised string equality.

`numeric` numeric equality within 10^{-6} relative tolerance; a non-parsing extraction is wrong.

`numeric_or_str` numeric equality when both sides parse, normalised string equality otherwise.

Base models rarely emit `\boxed{}` without being shown it, so extraction strictness is the axis most likely *a priori* to move the base model's curve. It does move $k^\star$ (a median of 1.9×, maximum 14.1×) but clears its local floor in only 5 of the 10 comparisons with a floor, and we therefore do not claim it (§4).

F TWO FLOOR COMPUTATIONS THAT INVERT THE VERDICT

This appendix records the two defective floor computations of §3.1 concretely, since the difference between them decides the paper.

Conditioning on definedness. A half-split ratio requires $k^\star$ to be defined in both halves. Pooling the ratios that survive this requirement over all cells, the largest per-cell 95th percentile is $12.19\times$, which exceeds every axis median in Table 2 and rejects the entire study. (That figure is a maximum *over* per-cell 95th percentiles and not itself a percentile: the 95th percentile of the same 545 per-cell values is $6.80\times$.) The surviving splits are not a noise sample: on cells where the base curve barely reaches the RLVR curve, they are the subset in which a low-probability event happened twice, and their spread describes that coincidence rather than the measurement. Marking a cell's floor invalid when more than 25% of its splits leave $k^\star$ undefined removes the artifact: of the 545 cell–rule pairs with a floor at all, 233 are valid, and going from the un-gated population to the valid one moves the median floor from $2.07\times$ to $2.00\times$ and the worst from $12.19\times$ to $5.82\times$. Every floor the verdict judges against is one of the 233.

Taking the floor globally. The $12.19\times$ figure above originates in a cell where $k^\star$ is defined under 0 of 12 rules. Used as a global threshold it vetoes comparisons made on cells where $k^\star$ is stable to within 10% across seeds. We therefore judge each comparison against the largest valid floor among its own group's cells; for the temperature axis those local floors run from $1.40\times$ to $5.16\times$.

Which side supplies the floor. The group is wider than the two cells compared, and 8 of the 30 standard-pair comparisons with a floor have a valid one on neither of their own two cells. The gate is not symmetric between the two temperatures: both sides survive in 18 comparisons, the low-temperature side alone in 4, the high-temperature side alone in 0, because a near-censored cell is where the undefined fraction runs high. Scored instead against the ungated maximum of the two compared cells' own floors, the count is 19 of 30.

Both corrections were committed before the grid they are applied to was generated. We state this because both were found after an earlier verdict, which is the circumstance in which changing a criterion is least defensible, and the only adequate answer is a checkable timestamp rather than an assurance.

How much of the verdict is the resampling. Every floor is a 95th percentile over 20000 half-splits, so it is an estimate, and a comparison lying near its floor falls on one side or the other on the RNG alone. We rebuilt the whole chain, grid and then verdict, on the grid reported here under $three$ further RNG seeds at three resolutions (`eval/seed_stability.py`, CPU only). Each artifact records the sha256 of all 114 per-problem record files it read, so a stability result computed on some other grid is detectable and not merely described. At 20000 splits, the resolution every grid here was built at, the temperature axis reads 61 of 68 under all 3 seeds and the other three axes' counts are likewise the same under every seed (the study counts every verdict row, seed replicates included, so its denominators are larger than the seed-0 rows of Table 2). At 2000 the temperature axis reads 61 of 68 and 60 of 68 across the seeds. At 200 it reads 61 of 69, 60 of 70 and 60 of 68: the count moves with the seed at that resolution, which is why the grids are not built there. Raising the split count raises the precision of one criterion rather than changing the criterion, and we did not change the criterion. Two quantities still move at 20000: the number of cell–rule pairs with a floor *before* the validity gate, 545 on the reported grid and 544–547 across these seeds, and the median valid floor, 2.00 on the reported grid and 1.96–2.00 across them. The gated count 233, the maximum valid floor and the seed floor are identical under all 3.

G EVERY COMPARISON AGAINST ITS FLOOR, AND THE LADDER IN FULL

Figure 9 is Figure 1(c,d) at full size.

Rule-level counts overstate replication: the 30 standard-pair rows of Table 2 are 20 distinct outcomes over 3 cell pairs on 2 benchmarks, and the 11 top-p probe rows re-measure one of those pairs. Whole-ladder ratios are a different population: the top rung's truncation makes every ladder-spanning

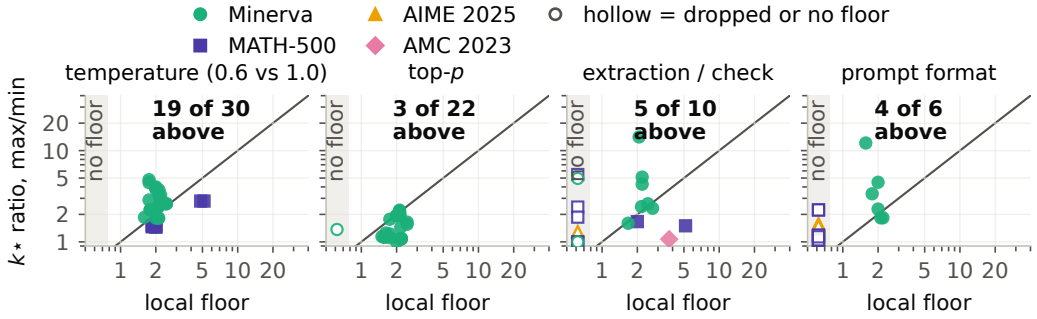

Figure 8: Every comparison against its floor, one panel per axis in the order of Table 2. A point is a comparison at (local floor, $k^\star$ ratio); above the diagonal it moved $k^\star$ by more than the measurement moves it on its own. Hollow points are dropped as truncation-suspect or have no valid floor (left strip). The temperature panel is Figure 3's population (0.6 vs 1.0), repeated here so the four axes are read on one scale.

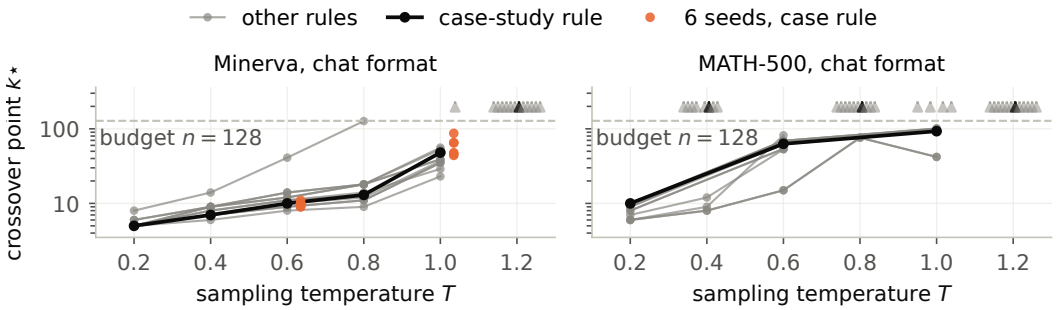

Figure 9: Figure 1(c,d) at full size: $k^\star$ against temperature, chat format, one line per scoring rule, log scale; the case-study rule is black, the others grey; orange dots are the case-study rule's $k^\star$ at the six seeds of the two reseeded Minerva cells, T=0.6 and T=1.0. Left, Minerva: all $12/12$ rules are monotone nondecreasing and censored at T=1.2 (arrows; one already at T=1.0, and the T=0.8 rung reaches $k^\star$=127 of n=128, a censoring boundary in all but name, which still counts toward the tally and contributes the widest span to the median). Right, MATH-500: 8 rules fail the monotonicity test on $k^\star$ undefined at an interior rung, 2 on a wiggle, and the passing $2/12$ are censored early enough to be monotone over 3 points.

comparison truncation-suspect and the verdict drops all 24; what survives still decides the axis (8 of 11 past their floors, median 2.3×).

H Conclusions per budget, in counts

Table 8 gives the counts behind Figure 4: per sampling budget, how many of the 684 grid points support each conclusion. Across the grid $k^\star$ spans 5–291 where defined, and is undefined at 262 of 684 points, censoring we report as such, since "did not catch up within n" and "catches up at k=10^6" are the same observation. Throughout, a $k^\star$ at k=n with both curves numerically equal is a tie and counted as censored (35 points), and n is the cell's own, so definedness compares only within a budget.

I Benchmark headroom

A benchmark both models eventually solve everywhere cannot exhibit a crossover: both curves converge to 1 and $k^\star$ degenerates (GSM8K (Cobbe et al., 2021), our first attempt, leaves both arms ever-solving 97.5–100.0%; Appendix K). We therefore report the headroom left at the maximum budget, $1 - \max(\text{pass@}n^{\text{base}}, \text{pass@}n^{\text{rlvr}})$: Minerva 67.6%, AIME 2025 43.3%, MATH-500 36.0%, AMC 2023 5.0%.

Table 8: Conclusions available per budget across the 684 grid points. *tied*: curves numerically equal, supporting neither position; *out of budget*: k exceeds the cell's n. Reading these off $k^\star$ would misreport every tie as "the base model caught up".

budget	base ahead	RLVR ahead	tied	out of budget
k=4	0	684	0	0
k=8	41	643	0	0
k=16	134	550	0	0
k=32	183	501	0	0
k=64	322	360	2	0
k=128	384	271	29	0
k=256	59	127	42	456
k=512	28	82	10	564

J THE WITHIN-CELL TRUNCATION BRACKET

The truncation asymmetry can be large. On the `Qwen2.5-Math-1.5B`/`Oat-Zero` pair (§2), on `gsm8k` at a 3840-token budget, the base arm is truncated on 24.9% of completions and the RLVR arm on 0.23%, a hundredfold asymmetry. Both that benchmark and that budget are ones we discard.

Could the budget explain the gap *within* a cell? We answer with a deliberately absurd repair: count every truncated-and-wrong RLVR sample as a win, leave the base arm alone, recompute $k^\star$. The result brackets the truth under any assumption about the truncated samples, and we report it without gating on it, because it is vacuous: near a crossover a two-point boost erases the ordering, so the bound collapses to a guaranteed ratio of 1.0 at 581 of 684 grid points while the *measured* between-arm truncation gap has a median of 1.78 pp. A bound that rejects everything has no power, which is why §3.2 gates on the gap *moving* instead.

K NEGATIVE AND ABANDONED CONFIGURATIONS

We report the configurations that did not work, because two of them are traps that a reader repeating this study would otherwise fall into.

A saturated benchmark makes the floor unmeasurable in principle. Our first grid used GSM8K, where every arm of every cell ever-solves 97.5–100.0% of its problems. Both curves converge to 1, $k^\star$ degenerates to the largest k at which the two estimates happen to tie, and the noise floor cannot be measured, not because the experiment was too small but because the quantity has no stable value to measure. We report benchmark headroom in §2 for this reason.

A generation budget that clips one arm hides in the scoring. An earlier grid at a 768-token budget did not record `finish_reason`, so truncation could not be bounded at all and every verdict from it was provisional regardless of the ratios. Recording per-completion termination reason costs nothing and should be standard.

A math-specialised population answers a different question. On the 1.5B math-specialised pair the crossover does not occur on the hardest benchmark under any rule (§4). We report this as a scope condition rather than a failed replication: it is the same measurement returning "the phenomenon is not present here".

L THE POPULATION TABLE IN FULL

Table 9 gives per-benchmark rows: definedness, and the standard-pair temperature contrast (median $k^\star$ ratio and fraction of rules moving later at T=1.0) where both endpoints are defined. Its definedness column is *not* a decomposition of Table 3's: that one is computed on the standard slice so that levels are comparable across pairs, while this one is computed over each pair's whole grid, which for the main pair also holds AMC 2023, the 0.2–1.2 ladder, the extra seeds, the top-p probe and both prompt formats. Two readings of the compact table's min-$k^\star$ column belong here. It is computed on that standard slice, so it is not the grid-wide minimum of 5 quoted in §4; and a min-$k^\star$ of 1, which

Mistral-7B and DeepSeek-Math both show, does not mean the crossover arrives immediately: it means the base arm already matches the RLVR arm at pass@1, so RLVR was never ahead on that slice at all. That is a different statement from the 14B pair's k=6, where RLVR does lead first. The DeepSeek-Math AIME row's $188.50\times$ is the near-coincident-curves case discussed in §4: both arms sit at $\leq 0.2\%$ pass@1, so the crossover point there is floor-dominated and we read no direction from it.

The 14B temperature ladder. Registered before any row (`PREREG_B3_14B_LADDER.md`), the 14B pair was re-run on a different box (two RTX 3090s, tensor-parallel 2, the same vLLM version and snapshots) at six temperatures on Minerva, chat format, n=128, with the two published temperatures re-generated there as same-box anchors. The anchors sit within 0.26 points of the A800 values on both arms, so the box term for this ladder is below the registered 2.0-point bound. The base arm's pass@1 falls monotonically with temperature (Spearman $\rho = -1.00$, a 15.4-point span); the RLVR arm moves within 0.5 points across the six rungs ($\rho = -0.77$), which fails the registered monotonicity test only because it is flat, and the "reproduced at 14B" sentence is therefore scoped to the base arm's dose response and the two-point RLVR comparison. RLVR leads at every rung, by 7.8 to 22.7 points (6/6 rungs above the box term); $k^\star$ is defined at the five lower rungs, between 6 and 15, and never crosses at T=1.2, where the base arm truncates 29.9% of its samples. A second draw of the same ladder (seed 1, on a third box: one RTX 4090 D per arm, tensor-parallel 1, byte-identical snapshots) was registered before its first row and is read by the same rule, separately. The base arm's dose response holds again ($\rho = -1.00$, a 15.5-point span); the RLVR arm is again flat (0.5 points) and this time passes the rank test ($\rho = -1.00$); RLVR leads at 6/6 rungs, by 7.8 to 22.8 points. The largest per-cell gap between the two draws is 0.26 points; seed and box change together, so we call it a seed-and-box spread and read no rung-to-rung step below it. No base step falls under it; 4 of the 5 RLVR steps do, which is the flatness already stated. A third draw (seed 2) on the same box as seed 1 separates the two on the four anchor cells registered as readable on their own (both arms at T=0.6 and 1.0): the largest seed-only gap there is 0.24 points, against 0.29 points from the same cells to the seed-0 draw on the first box, so the box adds nothing visible above the seed spread.

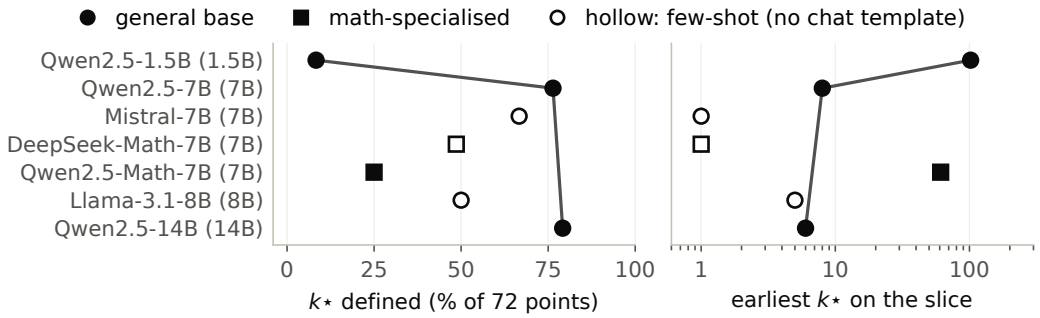

Figure 10: Table 3 by scale. Left: how often $k^\star$ is defined on the standard slice, one row per pair sorted by parameter count; the Qwen2.5 general ladder is connected. Right: the earliest $k^\star$ on that slice. Hollow markers are the pairs run few-shot. Every pair is seed 0.

Composition of the population contrasts. The population contrasts get the same judgement as the main axes: each rule's ratio against the local half-split floor of its own two cells. Of the 17 measurable standard-pair contrasts (default top-p), 12 of them the main pair's own cells and the rest from 3 of the 6 further pairs, 8 clear their floors in a majority of the rules with a floor, among them *both* benchmarks of the 14B pair, while 5 have no valid floor and remain direction-only observations, and 4 are judged and do not clear. Direction is mostly later: 12 of 17 measurable contrasts move later at T=1.0 under *every* rule, and five further cells of the standard slice, defined under every rule at T=0.6, are censored entirely at T=1.0, with the caveat that 6 of the contrasts are one MATH-500 cell pair under 6 seeds, where "every rule" holds for 3 of the 6 (the replicates give $0.60, 1.00, 0.00, 0.00, 1.00$), so the distinct-pair count is smaller than the row count and the MATH-500 direction is not stable across seeds (Table 9). The one contrast that moves the other way, DeepSeek-Math on AIME 2025, is judged and fails its floor, fittingly: both arms score $\leq 0.2\%$ pass@1, so the curves nearly coincide and their crossover is exactly the threshold statistic whose noise §3.1 measures. We read no direction from it.

Table 9: The crossover across seven released pairs, per benchmark, over each pair's whole grid rather than the standard slice of Table 3. Dashes: contrasts censored on at least one side. *Past floor* counts a different population from the 19 of 30 in Table 2 and is not that count broken out. Two differences, both deliberate: these rows are default top-p only, so the 2 top-p settings that Table 2 counts as separate comparisons appear here once; and each rule is judged against the floors of *its own two cells* rather than the largest valid floor in the comparison group. On the main pair the column therefore reads 64 of 68 where Table 2 reads 19 of 30. Both are stated; neither is a correction of the other.

pair	benchmark	$k^\star$ defined	med. 0.6→1.0	rising	past floor
Qwen2.5-7B (general, main pair)	aime25	24/72	–	–	–
	amc23	78/108	–	–	–
	math500 (seed 0)	152/240	1.48×	1.00	–
	math500 (seed 1)		1.48×	0.60	–
	math500 (seed 2)		3.36×	1.00	–
	math500 (seed 3)		1.54×	0.00	–
	math500 (seed 4)		1.31×	0.00	0/2
	math500 (seed 5)		3.15×	1.00	0/2
	minerva (seed 0)	203/264	3.79×	1.00	11/11
	minerva (seed 1)		3.33×	1.00	11/11
	minerva (seed 2)		3.75×	1.00	11/11
	minerva (seed 3)		4.62×	1.00	11/11
	minerva (seed 4)		4.44×	1.00	11/11
	minerva (seed 5)		4.67×	1.00	9/9
DeepSeek-Math-7B (specialised)	aime25	24/24	188.50×	0.00	0/3
	math500	1/24	–	–	–
	minerva	10/24	–	–	–
Llama-3.1-8B (general, few-shot)	aime25	12/24	–	–	–
	math500	12/24	–	–	–
	minerva	12/24	–	–	–
Qwen2.5-Math-7B (specialised)	aime25	0/24	–	–	–
	math500	18/24	–	–	–
	minerva	0/24	–	–	–
Mistral-7B-v0.1 (general, few-shot)	aime25	24/24	1.74×	0.50	0/12
	math500	10/24	–	–	–
	minerva	14/24	4.78×	1.00	–
Qwen2.5-14B (general)	aime25	12/24	–	–	–
	math500	21/24	4.55×	1.00	5/5
	minerva	24/24	3.02×	1.00	12/12
Qwen2.5-1.5B (general)	aime25	0/24	–	–	–
	math500	0/24	–	–	–
	minerva	6/24	–	–	–

M CALIBRATING THE VERDICT GATE

The gate g is the fraction of problem resamples in which a rule-seed judgement may disagree with its verdict. Table 10 runs the released procedure on the null control of §4, two draws of one base arm, at seven gates and under three aggregation rules: the majority over the 36 rule-benchmark judgements (the paper's), unanimity, and the primary rule alone. AHEAD on a null comparison is a false verdict; CAUGHT is the correct reading of two curves from one distribution. The registered $g{=}0.25$ gives 5 false AHEAD verdicts of 45 under the majority rule; $g{=}0.15$, the first gate under 5%, gives 2, and every tighter gate buys its lower rate with unjudgeable comparisons. Unanimity never returns a false AHEAD at any gate and is unjudgeable on a third of the null at the registered one.

Table 11 is synthetic and is labelled so in its artifact. For a fraction f of the problems a cell's base arm never solves, one of the n completions in one arm is rewritten as correct, which moves that arm's curve by a known amount, and the procedure is run at the calibrated gate on twenty draws per (f, cell), $f \in \{0.05/0.10/0.20\}$. Minerva reads AHEAD on 20/20 draws at $f{=}0.10$ and on 0/20 at $f{=}0.05$; MATH-500 reads AHEAD on 0/20 at any f, because each rule leaves it one to fourteen never-solved problems to inject into; AMC23 has 0 never-solved problems under seed 1 and cannot be injected at all, so its rows are the unmodified null. The pooled rate is therefore diluted by cells the

Table 10: Null calibration: verdict counts (AHEAD/CAUGHT/UNJUDGEABLE) on 45 comparisons with no movement, by gate and aggregation rule. Bootstrap B=20000, three seeds.

	majority			unanimity			case-study rule		
g	ahead	caught	unj.	ahead	caught	unj.	ahead	caught	unj.
0.25	5	39	1	0	29	16	3	33	9
0.20	3	37	5	0	28	17	1	32	12
0.15	2	35	8	0	27	18	1	31	13
0.10	1	32	12	0	25	20	0	30	15
0.05	0	31	14	0	21	24	0	27	18
0.025	0	29	16	0	19	26	0	25	20
0.01	0	28	17	0	18	27	0	25	20

construction cannot move, and the power of the gate is a property of the cell's never-solved set, not of g.

Table 11: Synthetic power of the calibrated gate: AHEAD verdicts out of twenty draws per cell, by injected fraction f. Synthetic; not a measurement on any released pair.

f	amc23	math500	minerva	pooled
0.00	0/1	0/1	0/1	0/3
0.02	0/20	0/20	0/20	0/60
0.05	0/20	0/20	0/20	0/60
0.10	0/20	0/20	20/20	20/60
0.20	0/20	0/20	20/20	20/60

N LINEAGE VERIFICATION FOR EVERY PAIR

Table 12 runs the weight-space check of §3.3 on every released pair this paper touches: median relative per-tensor distance, separation from the mispaired negative control, and the verdict. All six sit in the declared-lineage regime, but not equally: Mistral-7B's median relative distance is several times the positive control's and its separation from the mispaired control is the smallest of the six by almost an order of magnitude. The check is a screen against a mispaired release, which it passes; it is not a claim that the six are equally close to their bases.

Table 12: Weight-space lineage per pair, against one shared mispaired control.

pair	median rel. dist.	vs. mispaired	verdict
Qwen2.5-7B (main)	0.000390	3653×	confirmed
DeepSeek-Math-7B	0.000519	2744×	confirmed
Llama-3.1-8B	0.000839	1697×	confirmed
Qwen2.5-Math-7B	0.000330	4319×	confirmed
Mistral-7B-v0.1	0.002890	493×	confirmed
Open-Reasoner-Zero-7B	0.001018	1399×	confirmed
Qwen2.5-14B	0.000366	3892×	confirmed
Qwen2.5-1.5B	0.000425	3348×	confirmed

O THE EIGHTH PAIR: A SECOND RLVR RECIPE ON THE SAME BASE

The seven released pairs share one training recipe (SimpleRL-Zoo), so §4 leaves one objection open: the protocol sensitivity could be a property of that recipe. The eighth pair tests it as far as one added recipe can: the RLVR arm is Open-Reasoner-Zero-7B (Hu et al., 2025), trained from Qwen2.5-7B, the base of the 7b pair. All twelve cells were generated for this pair on the 3090 box (vLLM 0.7.3, torch 2.5.1+cu121, transformers 4.49.0): the result store does not key a cell by model path and the 7b pair ran AIME 2025 at n=512, so the registered base-cell reuse proviso did not hold and all six base cells were rerun rather than reused, keeping every within-pair comparison on one box. The run cost 29.93 GPU-hours (8.43 h wall on four cards). Lineage is verified as for every other pair: median relative distance 0.00101789 against the base, 1399× below the mispaired control (1.42378), the same regime as the SimpleRL-Zoo pair (0.000389765) and the Oat-Zero positive control (0.000984935).

Under P_0 the released procedure returns AHEAD for this pair at the calibrated gate (AHEAD at the registered one), dissenting on aime25; it is kept out of the seven-pair count, which is the one-recipe population.

R1, the transfer reading: undecided, and why. $k^\star$ is defined at both temperatures for two Minerva rules, and both moves are up: 62→110 ($1.77\times$) under `boxed_then_last_num` and 75→96 ($1.28\times$) under `boxed_then_marker`, the direction the `7b` pair moves in on all thirty of its own comparisons with a floor. What the pair lacks is a floor: the 25% validity gate rejects the local floor on both sides of both comparisons, and both are truncation-suspect under the 3pp asymmetry gate, so the registered test (does the move exceed its own floor?) cannot be evaluated. We report the moves and do not claim the transfer, nor its absence; the abstract's scope sentence is untouched either way.

R2, existence. $k^\star$ is defined on 19/72 cell–rule pairs of this slice, against 18/48 on the `7b` pair's matched slice.

R3, recipe versus protocol. Exactly one (benchmark, rule, temperature) point carries both a recipe ratio and a protocol ratio on the shared slice (Minerva, `boxed_then_marker`, $T{=}0.6$), and at it the recipe effect exceeds the protocol effect ($1.45\times$ against $1.28\times$). The registered clause reads "at every budget"; evaluated on a basis of one point it does not establish dominance, so the title's claim stands and the comparison is reported descriptively with $n{=}1$ beside it. One further caveat attaches to the recipe ratio itself: the two pairs' base arms were generated on different boxes (this pair's on the 3090, the `7b` pair's on the A800), so it carries that difference as well, while the within-pair $k^\star$ comparisons do not. Table 13 gives the slice cell by cell.

Table 13: $k^\star$ on the shared slice (few-shot, n=128, seed 0), the eighth pair against the 7b pair. "–" is an undefined $k^\star$ in a cell that exists; "—" marks where no comparable 7b cell exists at all (AIME 2025 ran at n=512 in the 7b pair). A budget-edge exact tie, censored under the verdict convention, is marked †.

			orz7b $k^\star$		main-pair $k^\star$	
benchmark	extraction	check	T=0.6	T=1.0	T=0.6	T=1.0
AIME 2025	boxed_strict	exact_str	89	–	—	—
	boxed_strict	numeric	76	–	—	—
	boxed_strict	numeric_or_str	76	–	—	—
	boxed_then_last_num	exact_str	89	–	—	—
	boxed_then_last_num	numeric	76	–	—	—
	boxed_then_last_num	numeric_or_str	76	–	—	—
	boxed_then_marker	exact_str	89	–	—	—
	boxed_then_marker	numeric	76	–	—	—
	boxed_then_marker	numeric_or_str	76	–	—	—
	last_number	exact_str	89	–	—	—
	last_number	numeric	76	–	—	—
	last_number	numeric_or_str	76	–	—	—
MATH-500	boxed_strict	exact_str	–	–	–	110
	boxed_strict	numeric	–	–	–	66
	boxed_strict	numeric_or_str	–	–	–	96
	boxed_then_last_num	exact_str	–	–	–	112
	boxed_then_last_num	numeric	–	–	–	63
	boxed_then_last_num	numeric_or_str	–	–	–	98
	boxed_then_marker	exact_str	–	–	–	98
	boxed_then_marker	numeric	–	–	–	60
	boxed_then_marker	numeric_or_str	–	–	–	88
	last_number	exact_str	–	–	–	83
	last_number	numeric	–	–	–	94
	last_number	numeric_or_str	–	–	–	94
Minerva	boxed_strict	exact_str	–	–	–	–
	boxed_strict	numeric	–	–	63	–
	boxed_strict	numeric_or_str	–	128†	–	–
	boxed_then_last_num	exact_str	–	–	–	–
	boxed_then_last_num	numeric	–	110	37	–
	boxed_then_last_num	numeric_or_str	62	110	–	–
	boxed_then_marker	exact_str	–	–	–	–
	boxed_then_marker	numeric	–	103	32	–
	boxed_then_marker	numeric_or_str	75	96	109	–
	last_number	exact_str	–	–	–	–
	last_number	numeric	–	–	22	–
	last_number	numeric_or_str	–	–	22	–