WHERE DOES RLVR ACT? A CONSERVED TOKEN LEDGER FOR WHAT POST-TRAINING ADDS TO A BASE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Most token-level accounts of what reinforcement learning with verifiable rewards (RLVR) changes in a base model are statistics of the two policies' next-token distributions, which have no total in accuracy and cannot say what a token is *worth*. Prefix transplantation with one permanent handoff (the trained policy writes k tokens, the base the rest) is a conserved ledger: by the performance-difference lemma its increments are the trained tokens' expected advantages under the base's own value function and sum to the measured gain; a dense check at the first four positions finds no over-count on any cell and a small pooled under-count (+0.014), as action-set truncation predicts. Across 13 cells (0.5B–8B), where front-loading is largest (0.73 of the gain in four tokens), the policy's most frequent opening, grafted into *every* problem, reproduces the whole early gain, and a better template on the base supplies about half (46–66%) of the early lift of every Qwen-Math RLVR descendant tested: the early credit is mostly format, relative to the base's template. GRPO on a *random* reward delivers 0.80 of its gain in four tokens, all of it format. Each advantage is a *room*, set by the token's surprisal under the base (its *rate*, in nats), times a *worth*, its value over the base's alternatives. Divergence statistics track the room; none predicts the worth's size beyond weakly (rank correlation at most 0.26 in magnitude). Code, raw records and scripts regenerating every table are supplementary and will be released at `https://github.com/xxx/xxx` upon acceptance.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) turns a base language model into a markedly better mathematical reasoner (DeepSeek-AI et al., 2025; Shao et al., 2024; Zeng et al., 2025), and a growing literature asks *what*, mechanically, it adds. The emerging answer is consistent: the change is sparse and concentrated early. Only a small fraction of token positions show a meaningful shift between the base and the RL policy (Meng et al., 2026; Akgül et al., 2026); those positions are high-entropy "forking" tokens (Wang et al., 2025a); the disagreement is concentrated in a few decision tokens, with a Gini coefficient above 0.9 (Shen et al., 2026); and a base model steered at a small fraction of positions recovers most of the gap to its thinking sibling (Venhoff et al., 2026). The natural reading: RLVR teaches *when* to deploy reasoning the base has, at a few early decision points.

Most of these statements are made in *distribution space*: what is localised is a statistic of the two next-token distributions (the trained policy's entropy, Jensen–Shannon divergence, cross-entropy disagreement, a log-probability shift), and the interventions among them hand control back and forth, so their gains do not add up to the total (Section 2). Such statistics have no total in accuracy, so "the gain is front-loaded" means only that the *difference* is, and none can identify what a token bought: each is compatible, in some token MDP, with the token having bought anything from the most it could to a loss of the same size (Proposition 2).

This paper answers that question (what a token bought, in accuracy) with an accounting identity rather than a proxy (Figure 1). The instrument, prefix transplantation, is familiar (Qi et al., 2025; Venhoff et al., 2026; Meng et al., 2026), and the policy that follows one model for k steps and another thereafter is the roll-in/roll-out of learning to search (Ross & Bagnell, 2014); we use it as a *single*

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.

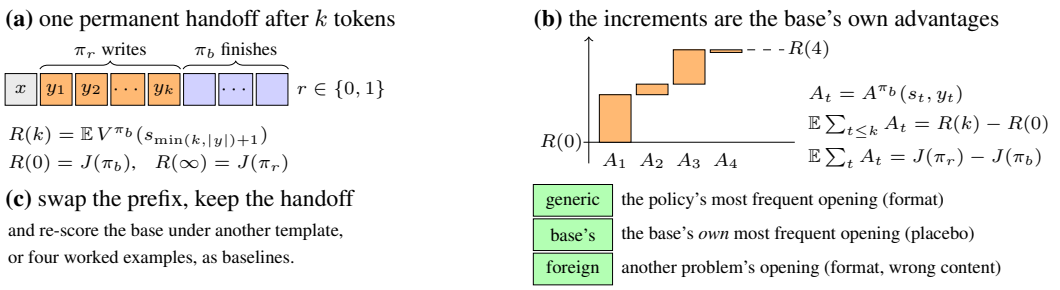


Figure 1: The probe, the ledger and the controls. (a) The trained policy writes k tokens, then the base finishes; $R(k)$ is the base's expected value at the handoff. (b) The increments of R are the base's expected advantages along the trained trace and sum to the measured gain. (c) Other prefixes separate format from content; templates or worked examples give the base other baselines.

permanent handoff after which the trained policy is unavailable, and make one observation: $R(k)$ is by construction the base's value averaged over trained k-token prefixes, so by the performance-difference lemma of Kakade & Langford (2002) $R(k) - R(0)$ is the expected sum of the first k per-token advantages of the trained trajectory *under the base's own value function*, and at $k = T$ it equals the measured gain $J(\pi_r) - J(\pi_b)$. The increments are therefore an additive, signed credit assignment in accuracy units whose parts sum to the whole, fixed once the baseline policy and the direction of the handoff are. We call it the *ledger*. Its prefix controls (a generic opening, the base's own opening, another problem's opening) could run with any transplant, but the identity puts their readings in the ledger's accuracy units against the same $R(0)$, and lets the summed increments be checked against the measured lift. A pricing layer follows: each token's price is its *rate*, $\varepsilon_t = -\log \pi_b(y_t \mid s_t)$ nats; its advantage factorises into the *room*, $1 - \pi_b(y_t \mid s_t)$, a function of the rate alone, and the *worth*, its value over the alternatives, which no divergence statistic identifies; and by a χ^2 argument it is at most $\sigma_b(s_t)\sqrt{e^{\varepsilon_t} - 1}$ in magnitude, with σ_b the dispersion of the base's own action values, so no token earns credit where those values do not vary.

We instantiate this on 8 base/trained pairs (13 cells; 0.5B–8B; GSM8K and MATH-500; Section 3), at their core three post-training runs of `Qwen2.5-Math-7B` (two RLVR recipes, one distillation) and two GRPO reward controls, all scored against the *same* value function. Every cell carries the two prefix controls at $k = 4$ (11 also against a few-shot baseline) and a dense check of the first four positions, and ledger positions are sampled over whole traces; ten extension runs (Appendix F) add bases, a recipe crossing, GSM-Symbolic, the reverse handoff, two template baselines and base temperatures. The main findings are:

(i) *The ledger conserves* (Section 4.1). At the first four positions a two-sided check resolves no over-count on any cell and a pooled under-count of +0.014, as action-set truncation predicts.

(ii) *Where front-loading is largest, the early gain is format that a template also supplies* (Sections 4.1–4.2). With $\rho(k)$ the share of the gain that k tokens deliver, the SimpleRL-Zero, Oat-Zero and SimpleRL-Zoo descendants of `Qwen2.5-Math-7B` reach $\rho(4) = 0.62$–0.73 on GSM8K (0.59 on MATH-500) and the Oat-Zero descendant of `Qwen2.5-Math-1.5B` 0.68, but every other RLVR pair ≤ 0.28, the same SimpleRL-Zoo recipe on `Qwen2.5-7B` and on `DeepSeek-Math-7B` (0.03) included: in the checkpoints tested, front-loading goes with the Qwen-Math bases and their template, not with the recipe. On the primary cell the policy's most frequent opening, grafted into every problem, reproduces 100% of the four-token lift, also on GSM-Symbolic, where memorised answers cannot help. A better template on the base supplies 46–63% of that lift on every RLVR descendant of this base (66% on the Oat-Zero descendant of the 1.5B Qwen-Math base), and with no template at all the first four trained tokens add at most +0.028 on the three descendants tested. GRPO on a *random* reward delivers 0.80 of its gain in four tokens, all of it format.

(iii) *Divergence statistics see room, not worth* (Section 4.3). They track the room factor; none predicts the size of the value term beyond weakly (Spearman $|r_s| \le 0.25$ at m=8 rollouts, ≤ 0.26 at m=32). The base's value dispersion does, weakly (+0.03 to +0.18 above estimator noise, mostly through the position's scale at m=8); whether it grades the trained token our nulls leave open.

(iv) *What a nat buys* (Section 4.4). 62–89% of the rate is spent where no value dispersion is detected at m=8, 48–64% on the whole-trace runs at m=32; the main RLVR cells fill 8–47% of a loose

per-position ceiling; and no run resolves a positive margin for the trained token over best-of-e^{ε_t} base sampling with a token-level value oracle, a statement of resolution, not of equivalence.

Contributions. The mathematics is classical. The ledger identity is a direct application of Kakade & Langford (2002) that appears in the LLM token MDP in Setlur et al. (2025), and Corollary 1 and Proposition 2 are local facts. Using the permanent-handoff probe through that identity to audit post-training is, to our knowledge, new, and the contribution is a measurement framework.

2 A CONSERVED LEDGER FOR POST-TRAINING GAINS

2.1 SETTING

A prompt x is drawn from $\mathcal{D}$; a *policy* π maps a state $s = (x, y_{<t})$ to a distribution over a finite vocabulary with an end symbol, and $s \cdot a$ appends a to s. Transitions are deterministic; an episode ends at the end symbol or at a token cap T shared by every policy (a shorter cap is a forced end symbol); the reward is terminal and verifiable, $r(x, y) \in \{0, 1\}$. A policy is the *whole* inference procedure (weights, prompt template, decoding parameters), so any two policies on one state space are comparable; π_b is the base and π_r the trained policy. Write $V^\pi(s) = \mathbb{E}_{y_{\geq t} \sim \pi(\cdot|s)}[r(x, y)]$ for a policy's success probability from s ($V^\pi = r$ at a terminal state), $Q^{\pi_b}(s, a) = V^{\pi_b}(s \cdot a)$, $A^{\pi_b}(s, a) = Q^{\pi_b}(s, a) - V^{\pi_b}(s)$, $A_t := A^{\pi_b}(s_t, y_t)$ along a trajectory, and $J(\pi) = \mathbb{E}_x\, V^\pi(x)$ for pass@1. Two facts carry the ledger: the Bellman identity $V^{\pi_b}(s) = \mathbb{E}_{a \sim \pi_b} Q^{\pi_b}(s, a)$ and its corollary $\mathbb{E}_{a \sim \pi_b} A^{\pi_b}(s, a) = 0$: gain comes only from moving mass relative to π_b, where Q^{π_b} varies.

2.2 THE TRANSPLANT CURVE IS THE LEDGER

Lemma 1 (performance difference, token MDP). *For any two policies π_r, π_b on the same token MDP, $J(\pi_r) - J(\pi_b) = \mathbb{E}_x\, \mathbb{E}_{y \sim \pi_r(\cdot|x)} \big[\sum_{t=1}^{|y|} A^{\pi_b}(s_t, y_t) \big]$.*

The proof (Appendix A) is a telescoping sum: along any trajectory $\sum_t [V^{\pi_b}(s_{t+1}) - V^{\pi_b}(s_t)] = r(x, y) - V^{\pi_b}(x)$, the reward being terminal, so the identity holds pathwise and on any selected set of traces sums to that set's own gain, $\mathbb{E}[r - V^{\pi_b}(x) \mid \text{set}]$. It is the finite-horizon, undiscounted analogue of Lemma 6.1 of Kakade & Langford (2002) and Lemma F.1 of Setlur et al. (2025).

Definition 1 (permanent handoff). *For $k \geq 0$ let H_k be the policy that follows π_r for the first k tokens and π_b thereafter, and let $R(k) = J(H_k)$. Then $R(0) = J(\pi_b)$ and $R(k) = J(\pi_r)$ for $k \geq T$.*

Proposition 1 (the transplant curve is the cumulative advantage). *$R(k) - R(0) = \mathbb{E}_x \mathbb{E}_{y \sim \pi_r(\cdot|x)} \big[\sum_{t \leq \min(k, |y|)} A^{\pi_b}(s_t, y_t) \big] =: G(k)$, and $G(T) = J(\pi_r) - J(\pi_b)$.*

By definition of H_k and the tower property, $R(k) = \mathbb{E}\, V^{\pi_b}(s_{\min(k, |y|)+1})$, the base's value at the trained policy's k-token prefix, and the truncated telescoping sum gives the claim. The consequence we build on: the Monte-Carlo estimate of $R(k)$ (prefill k trained tokens into π_b, sample completions, average the reward) is unbiased for $R(k)$ (on a length-selected set of traces, for that set's own ledger, with $R(0)$ weighted by its traces), so $\hat{R}(k) - \hat{R}(0)$ is unbiased for $G(k)$, and $\rho(k) = G(k)/G(T)$, the (signed) share of the gain the first k tokens deliver, has $\rho(0) = 0$ and $\rho(T) = 1$ whenever the gain is non-zero. The permanent handoff is essential: an intervention that alternates control (a single-token takeover, steering at selected positions) telescopes only to its *own* gain.

2.3 RATE, ROOM AND WORTH

Along a trained trajectory the *rate* is the surprisal of y_t under the base, $\varepsilon_t = -\log \pi_b(y_t \mid s_t) = \mathrm{KL}(\delta_{y_t} \,\|\, \pi_b(\cdot \mid s_t))$ nats. Fix s, write $q_b = \pi_b(\cdot \mid s)$, $Q(a) = Q^{\pi_b}(s, a)$, $\bar{V} = \mathbb{E}_{q_b} Q$.

Corollary 1 (room and worth). *For a realised token y_t with $\pi_b(y_t \mid s_t) < 1$, writing $V_b^{\mathrm{alt}}(s_t) = \mathbb{E}_{a \sim \pi_b(\cdot|s_t)}[Q^{\pi_b}(s_t, a) \mid a \neq y_t]$ for the base's value averaged over the alternatives to y_t, $A_t = (1 - \pi_b(y_t \mid s_t))\, [Q^{\pi_b}(s_t, y_t) - V_b^{\mathrm{alt}}(s_t)]$: the* room, *$1 - e^{-\varepsilon_t}$, a function of the base's surprisal alone, times the* worth, *which is where the base's value function enters; in particular $|A_t| \leq 1 - e^{-\varepsilon_t}$.*

Proposition 2 (divergence statistics do not identify worth). *Let g be any function of the two next-token distributions at s and of the token, $g(\pi_b(\cdot \mid s), \pi_r(\cdot \mid s), y)$—entropy, either KL, JSD, the*

realised log-ratio, a rank change. For every state s at least two steps before the cap, token y with $\pi_b(y \mid s) < 1$, and c with $|c| \le 1 - \pi_b(y \mid s)$, there is a reward function and a base continuation below s, leaving both next-token distributions at s unchanged, under which $A^{\pi_b}(s, y) = c$. Hence g determines A_t only up to the room bound of Corollary 1, and does not determine its sign.

The proof (Appendix A) reassigns the subtree below s, which the next-token distributions at s do not see; it is a worst-case statement, and whether such statistics *predict* worth is the question of Section 4.3. It concerns the *base's* value function, whereas under a KL-regularised objective the realised log-ratio is a soft advantage under the trained policy's soft values (Rafailov et al., 2024).

2.4 THE MOST A NAT CAN BUY

Definition 2 (rate–value frontier). *$F_s(\varepsilon) = \max\{\mathbb{E}_{a\sim q}Q(a) - \bar{V} : \mathrm{KL}(q\|q_b) \le \varepsilon\}$ is the most any next-token distribution at s can gain within ε nats of the base, and $F_s^{\det}(\varepsilon) = \max\{Q(a) - \bar{V} : -\log q_b(a) \le \varepsilon\}$ its restriction to one token (evaluated only at ε_t, where y_t is feasible).*

Proposition 3 (Gibbs attainment, linear response, gating). *Let $\sigma_b(s)^2 = \mathrm{Var}_{a\sim q_b}[Q(a)]$ be the* base value dispersion *and $\varepsilon_{\mathrm{sat}}(s) := -\log q_b(A^\star)$, $A^\star$ the maximisers of Q on the support of q_b. (i) F_s is concave, non-decreasing, and constant for $\varepsilon \ge \varepsilon_{\mathrm{sat}}$. (ii) If $\sigma_b(s) > 0$ and $0 < \varepsilon < \varepsilon_{\mathrm{sat}}$, $F_s(\varepsilon)$ is attained by $q_\beta(a) \propto q_b(a)e^{Q(a)/\beta}$ with $\mathrm{KL}(q_\beta\|q_b) = \varepsilon$. (iii) If $\sigma_b(s) > 0$, $F_s(\varepsilon) = \sigma_b(s)\sqrt{2\varepsilon} + O(\varepsilon)$ as $\varepsilon \to 0$ (coefficient in Appendix A): a first-order value, not a bound (no realised rate is in its regime; Appendix B). (iv) (gating) For any $\pi \ll q_b$ at s, $|\mathbb{E}_{a\sim\pi}A^{\pi_b}(s, a)| \le \sigma_b(s)\sqrt{\chi^2(\pi\|q_b)}$; at the point mass on y_t, $\chi^2 = e^{\varepsilon_t} - 1$, so $|A_t| \le \sigma_b(s_t)\sqrt{e^{\varepsilon_t} - 1}$.*

Part (ii) is the Gibbs principle of relative-entropy policy search (Peters et al., 2010), part (iv) the Hammersley–Chapman–Robbins inequality; what is not standard is the object: the gain *extractable* at each state of a *finished* policy, gated by the base's own dispersion (Appendix B places it against Setlur et al. (2025) and the action gap). Three quantities follow, defined there: the *wasted rate*, the share of a trace's rate spent at states where $\sigma_b = 0$, where by (iv) no token in the base's support earns credit at that step (estimated as the share where the bias-corrected $\hat{\sigma}_b^2$ is not positive, which falls with the rollout count); the *ceiling fill* $\Lambda = [J(\pi_r) - J(\pi_b)]/\mathbb{E}\sum_{t\le|y|} F_{s_t}^{\det}(\varepsilon_t)$, a ratio of expectations in $(0, 1]$ whenever the gain is positive, because $A_t \le F_{s_t}^{\det}(\varepsilon_t)$ pointwise and Lemma 1 equates the numerator with $\mathbb{E}\sum_t A_t$; and the *search margin*, A_t minus what e^{ε_t} base draws of the next token would buy, keeping the one of highest base value (a token-level value oracle).

3 EXPERIMENTAL DESIGN

Pairs. The design hinges on one base with three main descendants. `Qwen2.5-Math-7B` (Yang et al., 2024a) is the base of `Qwen-2.5-Math-7B-SimpleRL-Zero` (Zeng et al., 2025), rule-reward RL (PPO) from the base on 8K MATH problems with no supervised stage (the later SimpleRL-Zoo series is GRPO on per-base GSM8K/MATH subsets); of `Qwen2.5-Math-7B-Oat-Zero` (Liu et al., 2025b), Dr. GRPO on the same base; and of `DeepSeek-R1-Distill-Qwen-7B` (DeepSeek-AI et al., 2025), supervised distillation of R1 traces—two RLVR recipes and one distillation read off one value function; two GRPO checkpoints trained from the same base on a random and on the ground-truth reward (Shao et al., 2026) are scored against it as controls (Appendix D). `Qwen2.5-7B`, not math-specialised, has two RLVR descendants from different groups, `Qwen-2.5-7B-SimpleRL-Zoo` and `Open-Reasoner-Zero-7B` (Hu et al., 2025), each scored against the base under its own prompt (two pairs); `Qwen2.5-0.5B`, `Qwen2.5-1.5B` and `Llama-3.1-8B` (Grattafiori et al., 2024) with their `SimpleRL-Zoo` descendants add scale and a second family (hub identifiers in Appendix D). Within every RLVR pair, base and policy share template and cap, so weights are the only difference; only the distilled model writes under its own template. Every handoff prefills the trained tokens into the base under the *base's* template. Templates were chosen per pair by a 12-problem pilot on the trained model and frozen (Table 1). The choice is part of the policy and, on the chat-template pairs, disadvantages the base: on 50 held-out problems every base piloted scores 0.07–0.24 higher under the Question/Answer form and the primary base's three piloted descendants 0.15–0.24 higher under the chat form (Table 42), which Section 4.2 measures. Main-cell caps are 2048 tokens at 7B/8B and 1024 below.

Data. GSM8K test (Cobbe et al., 2021) (100–250 problems per cell) and MATH-500 (Lightman et al., 2024; Hendrycks et al., 2021) (100–200), graded by boxed-answer extraction (Appendix D); sampling at temperature 0.8, top-p 0.95, vLLM (Kwon et al., 2023).

Estimators. *Stage 0* draws 16 samples per problem from each policy: $J(\pi_b)$, $J(\pi_r)$, and a pool of trained traces. *Stage 1* estimates $R(k)$, $k \in \{4, 16, 32, 64, 128\}$, by prefilling the first k tokens of two traces per problem into the base and sampling six completions under the shared cap; traces shorter than the curve's largest k are dropped, so every point of a curve averages the same traces (a design choice; Appendix D). *Stage 2* estimates the ledger: a teacher-forced pass of the base over each trace yields ε_t and the base's top-20 distribution everywhere (read at temperature 1, while rollouts sample at 0.8 with nucleus 0.95; Appendix D checks both); at five positions per trace, drawn over the *whole* trace and stratified by ε-tercile with inverse-probability weights, the action set is the base's top four tokens plus the realised one, each with $m = 8$ base rollouts whose outcomes are stored individually. From these come $\hat{A}_t$, a bias-corrected $\hat{\sigma}_b^2$ (Appendix C), F^{det} and F^{srch} (Appendix B), a maximum and an expected order statistic of noisy estimates, reported with their estimators' known bias direction (Appendix D). Intervals are 95% cluster bootstraps over problems.

Controls. Two prefix controls separate format from content. The *generic prefix* is the trained policy's most frequent opening (the 4-token string it produces for the largest number of distinct problems), grafted identically into every problem (the top three such strings run as separate arms and pooled), so it is well-formed and problem-agnostic by construction, though not free of problem-type content; the *foreign prefix* is the policy's opening for a different problem (a fixed derangement), which keeps format but substitutes wrong content; a *placebo* grafts the base's own most frequent openings instead (5 cells; Appendix D). Both prefixes run at $k = 4$ on every cell; the foreign prefix also at $k = 16$ and 32 and the generic prefix at $k = 8$ and 16 on subsets (Tables 15, 31). A *few-shot base* arm scores the same weights with four worked examples prepended (Wei et al., 2022), an alternative baseline policy under the convention above, on 11 cells, repeated under greedy decoding on 8 and with two further prompt formats on the primary cell (Table 30). A *dense ledger* measures A_t at every one of the first four positions on every cell ($m = 6$–8), so $\sum_{t \le 4} \hat{A}_t$ can be compared with $R(4) - R(0)$ directly. A *resolution series* re-runs Stage 2 at $m = 32$ over whole traces and at $m = 32$–64 over the first 256 tokens (Table 18 lists every run); a further pass records both policies' top-20 distributions at the window positions of nine cells, where divergence statistics are scored. Ten *extension runs* (Appendix F) add a recipe crossing, two math bases, a base that already writes the answer format, GSM-Symbolic (Mirzadeh et al., 2025), the reverse handoff, template pilots, Question/Answer and no-template baselines and a base-temperature sweep.

4 Results

Table 6 lists the 13 cells with every policy's accuracy: gaps against the zero-shot base run from 0.16 to 0.46, and the same weights under a 4-shot prompt in the pair's template close up to 56% of them, more under other prompts and greedy decoding (Section 4.2). Cells are labelled base[-descendant]/dataset (QM7B: Qwen2.5-Math-7B; Q7B: Qwen2.5-7B; Q0.5B, Q1.5B; L8B: Llama-3.1-8B; no suffix: the SimpleRL descendant, -Zero on QM7B and -Zoo elsewhere; -Oat, -SFT, -ORZ: Oat-Zero, the distilled model, Open-Reasoner-Zero; G: GSM8K; M: MATH-500); the *primary cell* is QM7B/G, chosen after the curves were measured because front-loading is largest there, and the controls run deepest on it.

4.1 The ledger, and that it conserves

Figure 2(a) gives Section 2's $\rho(k)$ on the length-selected traces, $[R(k) - R(0)]/[R(\infty) - R(0)]$ with $R(\infty)$ their own accuracy (Table 11) and $R(0)$ the base's on the same problems, averaged per problem (weighting it by the traces, as unbiasedness requires, moves $\rho(4)$ by at most 0.032; Appendix D), so $\rho(0) = 0$ and $\rho(\infty) = 1$ by construction. The shapes differ sharply: on the 7B Qwen-Math pair the GSM8K curve rises to $\rho(4) = 0.73$ and is then flat (on MATH-500 0.59, then 0.71 by $k = 128$); on the 1.5B and 0.5B pairs it climbs steadily to $\rho(128) = 0.76$–0.89; and on the Llama pair it starts at 0.01 and reaches 0.73 only at $k = 128$ (on the 34% of its traces longer than 128 tokens). At face value only the Qwen-Math-7B pairs front-load (in the extension runs, a Qwen-Math-1.5B pair does too). The *rate* is front-loaded on every cell too, the gain furthest ahead of

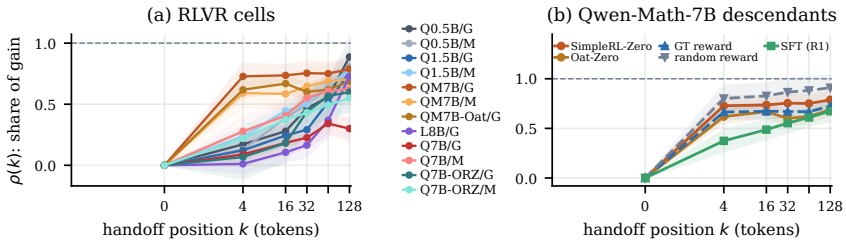


Figure 2: The conserved ledger. (a) $\rho(k)$ for the RLVR cells, 95% cluster-bootstrap bands. (b) Five descendants of one base against one value function (dashed: GRPO on the correct and on a random reward): raw $\rho(k)$ suggests RLVR front-loads and distillation less so; Section 4.2 shows that on this base, against its chat template, the RLVR lift at k=4 is mostly format.

it on the Qwen-Math-7B pairs (Table 24). On the general-purpose `Qwen2.5-7B` base $\rho(4)$ is 0.09 $[-0.01, 0.17]$ for SimpleRL-Zoo, under the same chat template, and 0.07 for Open-Reasoner-Zero (0.28 and 0.22 on MATH-500). Reversed on the primary cell, with the base writing the first k tokens and the policy finishing, the curve front-loads too, less: $\rho_{\text{rev}}(4) = 0.48$ [0.41, 0.55] (Appendix F).

We measured A_t at *every* one of the first four positions on every cell and compared $\sum_{t\le 4} \hat{A}_t$ with $R(4) - R(0)$ on the same traces, a check of the estimator, not of Proposition 1. The action set is the base's top four tokens plus the realised one; if the excluded tail is worth no more than the measured set's renormalised mean, $\hat{A}_t$ *under*-reports, so the shortfall should lie between zero and what that truncation costs (the excluded mass valued at zero). Its interval overlaps that band in 13 of 13 cells, with point estimates from -0.034 to +0.086 of which 5 are negative but none distinguishably so (Table 8); on Q1.5B/G the upper edge sits at zero and the verdict holds in 91 of 100 bootstrap streams (Appendix D). The check is one-sided. Where the allowance exceeds the lift (12 cells) even a ledger that recorded nothing would pass, so it rules out over-counting where it has power (an over-count of 0.041–0.100 would be detected with 80% power, more than the lift itself on the cells whose lift is small), not under-counting, which QM7B/G and QM7B-SFT/G resolve. A two-sided check, each position's value estimate against the direct one from its predecessor's realised rollouts, resolves no over-count on any cell and, pooled over the 13 cells, an under-count of +0.014 $[+0.002, +0.025]$ over four positions, as truncation predicts; per-cell intervals reach ± 0.11 (Appendix G).

Panel (b) is the designed comparison: against the *same* V^{π_b} and cap, the SimpleRL-Zero descendant of `Qwen2.5-Math-7B` delivers $\rho(4) = 0.73$, the Oat-Zero descendant (another recipe, same base and template) 0.62 $[0.55, 0.68]$, and the distilled one 0.37, climbing to 0.67 while both RLVR curves stay near their k=4 level. In the checkpoints tested, front-loading goes with the base and its template, not with the recipe or the reward: every RLVR descendant of this base front-loads, GRPO on a random or the correct reward included (dashed; 0.80, 0.66), and so does the SimpleRL-Zoo recipe, $\rho(4) = 0.64$, which gives 0.09 on `Qwen2.5-7B` (no less handicapped by the chat template, Table 42) and 0.03 on `DeepSeek-Math-7B`, while Oat-Zero on `Qwen2.5-Math-1.5B` gives 0.68 (Appendix F).

4.2 Where front-loading is largest, the early gain is format that a template also supplies

A k-token prefix carries the *format* of an answer to this kind of question and *content* specific to this question. The generic prefix, the same well-formed opening for every problem (like a fixed trigger phrase; Kojima et al., 2022), *estimates* what committing to an opening is worth, $R_{\text{gen}}(4) - R(0)$ (which depends on the opening, Appendix D); the foreign prefix keeps format but substitutes wrong content and only lower-bounds it (Table 14, Figure 3). The prefill path itself adds nothing detectable: handed back its own traces, the base moves by -0.020 [-0.068, +0.025] (Appendix D).

The generic prefix accounts for 23–100% of the early lift on 9 of the 10 Qwen cells with a resolved early lift (on Q7B-ORZ/M the generic opening buys -0.010 of a +0.075 lift) and, on the primary cell, all of it: $+0.264$ against a real lift of $+0.264$, leaving $+0.000$ $[-0.023, +0.022]$—not by construction: only 22% of the transplanted traces open with a grafted string, and on the others content is -0.003 [-0.030, +0.024] (Appendix D). Nor is it memorised answers: on GSM-Symbolic,

whose re-instantiated problems make a memorised GSM8K answer wrong, the pair front-loads as before ($\rho(4) = 0.68$) and the generic opening again takes the whole early lift (+0.250 against +0.232, content -0.018 $[-0.041, +0.005]$), the base's own opening less of it (+0.180; Appendix F). Nor is it the base's sampling temperature: sampled at $T = 0.3$ to 1.0, the base leaves four-token lifts of 0.218–0.327, and the grafted openings carry 95–106% of them (Appendix F). Under the second RLVR recipe on that base, content does not quite vanish: 83% of a $+0.278$ lift is generic, leaving $+0.049$ $[+0.025, +0.073]$. The arm also has a placebo: grafting the *base's own* most frequent opening (`To solve this problem`, 59% of problems, also the policy's third; pooled with the base's next two) gains +0.267 [+0.231, +0.304] against +0.264 for the policy's, a difference (policy minus base) of -0.003 [-0.022, +0.018], +0.006 [-0.019, +0.034] on the two arms the pools do not share (Appendix D): committing the base to its own most frequent opening, a sharpening of the base (Huang et al., 2025) rather than anything the policy taught, buys what the clearest case of "RLVR front-loads" buys; elsewhere the two openings are not interchangeable (the policy's is worth more on GSM-Symbolic and the 1.5B Qwen-Math pair, the base's on Q1.5B/G, L8B/G, Q7B-ORZ/G; Appendices D, F). Trained on either reward, the GRPO recipe opens most often with the base's own most common opening; on a random reward the generic arm it leads accounts for the whole four-token lift (+0.250 against +0.233, content -0.017 $[-0.044, +0.010]$); on the correct one the lift (+0.286) exceeds it by resolved content, +0.034 $[+0.015, +0.054]$ (Appendix D). Nor is it mainly the handoff teaching the base to stop: truncation falls from 22% to 10% over the handoff, at most 45% of the lift even fully credited (0–57% across cells, 83% on the 1.5B Qwen-Math extension; Appendices D, F), and graded on first answers the four-token lift rises, from +0.279 to +0.339 on the first 100 problems, with the generic arm at +0.300 and small resolved content, +0.039 [+0.004, +0.073] (Appendix G). The Llama policy has no shared opening (none covers over 1% of problems) and little front-loading, an asymmetry Gandhi et al. (2025) trace to what the base does.

Three things follow. First, against the chat-template base, Figure 2(b)'s ordering inverts once format is removed: the distilled descendant's early lift is $+0.070$ format and $+0.082$ content, more than the SimpleRL-Zero recipe's ($+0.000$) and at least as much as Oat-Zero's ($+0.049$; overlapping intervals, on a subset of the problems; Appendix D); against the Question/Answer base the distilled descendant's early content vanishes (Appendix F). Second, a better template on the base does much of the same work. Under the *distilled* policy's chat template the primary cell's base starts +0.350 [+0.294, +0.403] higher (99 problems) and the four-token lift falls from +0.266 [+0.215, +0.321] to +0.002 [-0.023, +0.028]. Under the Question/Answer form every piloted base prefers, the base starts +0.164 [+0.115, +0.214] higher and, across its 5 RLVR descendants, the four-token lift falls from 0.23–0.29 to 0.09–0.15 while the curves still front-load ($\rho(4) = 0.39$–0.57 where the gap allows; on the 1.5B Qwen-Math pair the lift falls from +0.311 to +0.107, $\rho(4) = 0.59$ [0.45, 0.73]); the random-reward gap on those problems falls from 0.262 to 0.096, and what remains is mostly generic (59–80%) except on the primary cell, where it is mostly content (+0.074 [+0.034, +0.113]), although content is resolved on the random-reward and correct-reward policies too (+0.032, +0.060; Appendix F). With no template at all, the base's strongest form (Liu et al., 2025b), it starts +0.280 above the chat form and four trained tokens add nothing on the primary cell (-0.005) and at most +0.028 on two more descendants (Appendix F). Format credit is relative to the baseline's template; the room–worth pattern and the wasted rate are not (Appendix F). Third, four worked examples close 56% of that cell's gap in the chat form and 96–97% in the Question/Answer and Problem/Solution forms (sampled; Tables 6, 30), and greedy four-shot prompting recovers 89–96% of the greedy gaps of that base's three descendants against 9–33% on the 4 small-Qwen cells. The controls compose: those four tokens lift the *few-shot* base by $+0.063$ where the zero-shot base gained $+0.264$. Trained tokens, worked examples and templates are substitutes, each telling the base how to start.

4.3 Room and worth

Corollary 1 reads $|A_t| = (1 - \pi_b(y_t \mid s_t))\,|Q^{\pi_b}(s_t, y_t) - V_b^{\mathrm{alt}}(s_t)|$, and Proposition 2 says no statistic of the two next-token distributions can determine the second factor. Empirically they track the first: on the window positions of 9 cells, entropy, both KLs, JSD, the trained policy's entropy and the cross-entropy correlate with $|A_t|$ at most 0.02 above the rate's coefficient (and up to 0.18 below it), and so does the realised log-ratio except on the primary base's SimpleRL-Zero cells (Tables 23, 52).

What predicts the *size* of the second factor (its sign is the ledger's own reading)? Against the value term, rate and entropy have $|r_s| \le 0.24$ on the 13 whole-trace runs (10,050 positions), running from -0.24 to +0.01 (-0.26 on a window run at m=32), so what they predict they predict inversely (0.35

at most after correcting for the value term's split-half reliability, Appendix G); the four two-policy divergences have $|r_s| \leq 0.11$, of mixed sign, on the window runs, and ≤ 0.21 on the 10 re-scored at m=32–64 (Tables 22, 23, 53). The base's value dispersion has $r_s = 0.26$–0.56 raw, but much of that is the estimator: $\hat{\sigma}_b^{\text{alt}}$ and the value term both scale with the noise of the position's success rate and co-vary under a null in which every action at a position shares one success rate, even when measured on independent rollouts (Table 32). That null alone produces $r_s = 0.16$–0.43, 51–91% of the raw figure; the excess over it is +0.03 to +0.18, significant (one-sided) in 12 of 13 cells, 9 after Holm correction; conditioning also on the position's success-rate variance $\hat{V}(1-\hat{V})$, which scales noise and spread alike, leaves it significant in 2 of them at $m = 8$ and in 6 of 15 runs at $m = 32$–64 (1 and 3 after Holm; Table 25): at $m = 8$ the excess is not told apart from scale.

Attenuated association or residual noise? On nested subsamples of one high-resolution draw the excess pooled over the 15 RLVR runs of 8 cells rises with m': +0.07, +0.12 and +0.19 at $m' = 8$, 16, 32, and from +0.14 to +0.23 between $m' = 32$ and 64 on three window runs (Table 27); noise-manufactured correlation does the opposite. Out of sample, with the value term measured on fresh-seed rollouts, dispersion predicts it at $r_s = +0.33$ over 14 runs of 9 cells against a null of +0.21, significant in 9 of them, while rate and entropy come out at -0.05 and -0.05 (Table 32).

Is the association specific to the trained token? Redrawing the realised token from the base's own distribution over the measured actions gives a coefficient the observed one does not detectably exceed (excess -0.01 to +0.01 on the 12 RLVR cells, +0.03 on the distilled one, $p \geq 0.07$), but that redraw returns the trained token itself at most positions; redrawn among the alternatives only, the observed coefficient is higher on all 8 cells of the two 7B bases (+0.04 to +0.16, $p \leq 0.01$ over redraws) and lower on the 5 others (-0.06 to -0.01), a comparison confounded by the alternatives' lower base probabilities and, resampling problems, resolved on only 6 of the 8 (3 against the nearest-probability alternative; 1 of 15 runs after Holm; Tables 26, 51). Whether dispersion grades the trained token our nulls leave open. Selection agrees (Table 29): only the held-out $|\hat{A}_t|$ oracle beats a random pick at every budget, and no criterion separates from rate. Divergence statistics, then, rank tokens by room, which bounds $|A_t|$; at our rollout counts they resolve little about worth.

4.4 What a nat buys

Table 17 reports the frontier quantities of Appendix B, a maximum and an expected maximum over noisy estimates (F^{det} inflated 4.4-fold by selection at $m = 8$ on a synthetic ledger, Table 20), so Λ is a floor against that selection and the search margin a bracket (both orderings checked there; Appendix D). *Where the rate goes.* 62–89% of a whole-trace run's total rate is spent where the bias-corrected dispersion estimate is not positive at m=8, a share that falls as the estimator sharpens (48–64% on the whole-trace runs at m=32; from 60–72% to 41–50% on the three windows re-run from m=8 to 64, Table 18) and is not a bound: at a flat position with an interior success rate the unbiased $\hat{\sigma}_b^2$ is positive about half the time. Within one draw it falls with m' on 13 of 16 runs, and at $m \geq 32$ only 0–26% sits where every rollout failed or succeeded (Appendix G). *How much of the ceiling is filled.* Λ, the measured gain over the summed *per-position* ceilings (at most one, since $A_t \leq F^{\text{det}}(\varepsilon_t)$), is 8–47% on the main RLVR cells (52% for the random-reward policy) and 1% for the distilled sibling, a level set mostly by how loose the ceiling is: position-wise maxima do not compose, and summed they exceed the headroom $1 - J(\pi_b)$ by 1.5–66$\times$ on all 13 cells (plug-in; debiased, resolved on 7; Table 19); at m=32 it is 14–90% on 5 cells, with wide intervals (Table 18). The ceiling falls after the 256th token on 8 of 12 cells, so a leading-window extrapolation would misstate Λ (Table 28); whole-trace sampling avoids that. *Search.* No run resolves a positive margin: the naive estimate lower-bounds it and the cross-fitted one upper-bounds it in expectation; on the whole-trace m=8 runs that upper bound lies within ± 0.0062 of zero on every cell and covers zero in 11 of 13, on the primary cell 0.0000 [-0.0017, +0.0014]. It lies entirely above zero on QM7B-Oat/G (+0.0017, lower edge +0.0001) and Q7B/M (+0.0016, lower edge +0.0002), naive bounds -0.0019 and -0.0014; elsewhere the cross-fit is above zero on 3 window runs, the naive bound not, and below it on Q0.5B/G (whole trace, m=32), which resolves a *negative* margin (Table 18). The bracket is 1.3–3.3 times the mean realised advantage on the 4 cells where that is resolved above zero (Table 16), and at the RLVR cells' mean rates the comparator draws $e^{\bar{\varepsilon}_t} = 1.08$–$1.42$ tokens, so this is a statement of resolution (80% power needs a margin 2.2–4.4 times the mean realised advantage; Appendix G), not of equivalence; in each run's top rate tercile, where it draws a median of 1.2–3.3 (16 on the distilled run), the bracket still contains zero on all runs but one, where it is negative (Appendix G).

5 Related work

Yue et al. (2025) showed that base pass@k with a verifier overtakes RLVR at large k, Karan & Du (2026) that verifier-free sampling from a sharpened base distribution matches GRPO, Wu et al. (2025a) that RLVR stays within the base's support, and Shao et al. (2026) that Qwen2.5-Math's gains are largely elicitation, which contamination may inflate (Wu et al., 2025b); Wen et al. (2026) and Liu et al. (2025a) dispute the support-boundary reading; minimal or reward-free training elicits much of the gain (Wang et al., 2025b; Agarwal et al., 2025); Zhao et al. (2025) find RL converging on a format pretraining already favours; and Liu et al. (2025b) find Qwen2.5 bases reasoning strongly with no prompt template, an instance of the prompt-format sensitivity Sclar et al. (2024) quantify, which our template baselines price in the ledger's units. Matsutani et al. (2026) and Venhoff et al. (2026) contrast RL with distillation, which we price against one value function. Token-level localisation (high-entropy tokens, token-level JSD, re-ranked top-5 positions, cross-entropy disagreement, early and critical tokens, entropy collapse; Wang et al., 2025a; Meng et al., 2026; Akgül et al., 2026; Shen et al., 2026; Ji et al., 2025; Lin et al., 2025; Cui et al., 2025) mostly works with statistics of *difference*, and where it intervenes it returns control, so the gains do not add; the ledger measures *worth*, and its prefix controls expose the format confound, the superficial-alignment hypothesis (Zhou et al., 2023; Lin et al., 2024; Qi et al., 2025) transposed to RLVR. Rollout estimation of $V(\text{prefix})$ underlies automated process supervision (Wang et al., 2024; Luo et al., 2024), pivotal-token search (Abdin et al., 2024) and critic-free credit assignment (Kazemnejad et al., 2025); Setlur et al. (2025) use the performance-difference lemma to turn a *prover* policy's advantages into process rewards; we use it to audit a finished policy against its base (further ancestors in Appendix D).

6 Discussion and limitations

Limitations. Value functions are estimated by rollout ($m = 8$ on every whole-trace ledger, 6–8 on the dense one, 32–64 on subsets; the base's top four tokens plus the realised one, 69–100% of its mass per cell; Tables 8, 16), so $\hat{\sigma}_b = 0$ means only "no dispersion detectable here", the wasted-rate figures fall with resolution, and Λ is a floor only against selection. Weights and rates are read at temperature 1 while rollouts sample at 0.8 with nucleus 0.95; re-weighting at 0.8 moves Λ by at most 5 percentage points, the dispersion–worth coefficient by at most 0.02, and no dense conservation verdict, and the realised token lies outside the nucleus at 1.1–2.3% of RLVR positions at most, which carry up to 21–40% of the rate; letting the base sample at temperature 1, so that weights and values are one policy's, keeps the qualitative pricing pattern on two cells but moves the Oat-Zero cell's wasted share from 74% to 52% and its Λ from 47% to 64% (Appendix D). Conservation is tested at $k \leq 4$; whole-trace sums test it only coarsely: pooled over the 13 cells they bound a common per-position bias at $\pm 0.8 \times 10^{-3}$ without detecting one, while the primary cell's sum runs high on its 46 traces cut at the cap (not under that temperature-1 re-run) and the DeepSeek-Math extension's misses its target, so per-position magnitudes at the 10^{-3} scale of the search margins are not resolved from estimator bias (Appendix D). The primary cell was chosen after the curves were measured, and much of its gap is template: against the base under the distilled policy's template, the Question/Answer template or none its guarded gap on 99 problems is 0.049, 0.234 or 0.119, not 0.398. Front-loading and the format result are observed on the Qwen-Math bases, whose benchmark contamination has been raised (Shao et al., 2026; Wu et al., 2025b); on GSM-Symbolic, where a memorised answer is wrong, the format result holds for the three checkpoints run (memorised procedures could still help; MATH-500 has no such control). The baseline controls use one exemplar set and three alternative templates, so they understate what prompting the base can elicit; each descendant is one public checkpoint, so recipe and seed are not separated; the distilled model runs at our temperature and 2048-token cap (24% truncated); all tasks are mathematical with binary rewards at ≤ 8B parameters (more: Appendix D).

What the ledger changes. Whole-trace comparisons show templates and elicitation carrying much of Qwen-Math's RLVR gain (Liu et al., 2025b; Shao et al., 2026). The ledger locates it along the trace: where front-loading is largest, the early credit is what a problem-agnostic opening buys, and it vanishes against the base with no template. Divergence statistics measure a token's room, not its worth.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, other than drafting the eight worked examples used in the few-shot baseline prompts, we did not use them to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyses publicly released open-weight models on public mathematics benchmarks. It involves no human subjects, no personal data collected by us, and no safety-relevant capabilities; the released records are raw model output, which can contain fragments memorised from public web text, and we have not filtered them. The models and benchmarks are used under their released licences, which the released records inherit. The paper's findings concern how post-training gains should be measured and attributed; we see no direct path to misuse.

REPRODUCIBILITY STATEMENT

All theory is proved in Appendix A. The estimators are checked on an enumerable synthetic MDP with known value functions (Appendix C); the identity checks are exact and need no GPU. The experimental pipeline (vLLM drivers, resumable job runners, graders with unit tests, analysis and figure scripts) and every raw generation record (77,430 records in 285 gzipped JSONL files, one record per problem and arm with its per-rollout outcomes, 369 gzipped files in all including manifests and pilots; shipped as one xz archive that a script restores and checksums) are included in the supplementary material, and every table entry is produced by the released scripts from those records, except that both policies' top-20 next-token distributions at the sampled positions are large and are released as the per-position statistics they feed instead, with the scripts that recompute Tables 23 and 52 from them, and the compute hours in Appendix D are read from run logs. Model checkpoints are public and named in Appendix D; templates, decoding parameters and token budgets are specified in Section 3 and Appendix D, and every random seed is fixed in the released code (batched vLLM generation is not bit-exact, so regenerated records reproduce the numbers up to sampling noise).

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in LLM reasoning. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Ömer Faruk Akgül, Rajgopal Kannan, Willie Neiswanger, and Viktor Prasanna. Rethinking RL for LLM reasoning: It's sparse policy selection, not capability learning. *arXiv preprint arXiv:2605.06241*, 2026.

Dan Amir and Ofra Amir. HIGHLIGHTS: Summarizing agent behavior to people. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2018.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*, 2024.

Marc G. Bellemare, Georg Ostrovski, Arthur Guez, Philip S. Thomas, and Rémi Munos. Increasing the action gap: New operators for reinforcement learning. In *AAAI Conference on Artificial Intelligence (AAAI 2016)*, 2016.

Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking paths in neural text generation. In *International Conference on Learning Representations (ICLR 2025)*, 2025.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which LLM reasoning steps matter? *arXiv preprint arXiv:2506.19143*, 2025.

Kai-Wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daumé III, and John Langford. Learning to search better than your teacher. In *International Conference on Machine Learning (ICML 2015)*, 2015.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *International Conference on Machine Learning (ICML 2025)*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Łukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

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

Amir-massoud Farahmand. Action-gap phenomenon in reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Yu Fei, Yasaman Razeghi, and Sameer Singh. Nudging: Inference-time alignment of LLMs via guided decoding. In *Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, 2025.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective STaRs. In *Conference on Language Modeling (COLM 2025)*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML 2023)*, 2023.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track*, 2021.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *arXiv preprint arXiv:2504.07086*, 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 preprint arXiv:2503.24290*, 2025.

Audrey Huang, Adam Block, Dylan J. Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *International Conference on Learning Representations (ICLR 2025)*, 2025.

Ke Ji, Jiahao Xu, Tian Liang, Qiuzhi Liu, Zhiwei He, Xingyu Chen, Xiaoyuan Liu, Zhijie Wang, Junying Chen, Benyou Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. The first few tokens are all you need: An efficient and effective unsupervised prefix fine-tuning method for reasoning models. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning (ICML 2002)*, 2002.

Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. In *International Conference on Learning Representations (ICLR 2026)*, 2026.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *International Conference on Machine Learning (ICML 2025)*, 2025.

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

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 *ACM Symposium on Operating Systems Principles (SOSP)*, 2023.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Baohao Liao, Yuhui Xu, Hanze Dong, Junnan Li, Christof Monz, Silvio Savarese, Doyen Sahoo, and Caiming Xiong. Reward-guided speculative decoding for efficient LLM reasoning. *arXiv preprint arXiv:2501.19324*, 2025.

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 (ICLR 2024)*, 2024.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *International Conference on Learning Representations (ICLR 2024)*, 2024.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Lin, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhances LLM's reasoning capability. In *International Conference on Machine Learning (ICML 2025)*, 2025.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. Tuning language models by proxy. In *Conference on Language Modeling (COLM 2024)*, 2024.

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. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 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. In *Conference on Language Modeling (COLM 2025)*, 2025b.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. DeepScaleR: Surpassing O1-Preview with a 1.5B model by scaling RL. Notion blog post and model release, `https://huggingface.co/agentica-org/DeepScaleR-1.5B-Preview`, 2025.

Ang Lv, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Rui Yan. The climb carves wisdom deeper than the summit: On the noisy rewards in learning to reason. *arXiv preprint arXiv:2505.22653*, 2025.

Kohsei Matsutani, Shota Takashiro, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. RL squeezes, SFT expands: A comparative study of reasoning LLMs. In *International Conference on Learning Representations (ICLR 2026)*, 2026.

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

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations (ICLR 2025)*, 2025.

Eric Mitchell, Rafael Rafailov, Archit Sharma, Chelsea Finn, and Christopher D. Manning. An emulator for fine-tuning large language models using small language models. In *International Conference on Learning Representations (ICLR 2024)*, 2024.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, Jilin Chen, Alex Beutel, and Ahmad Beirami. Controlled decoding from language models. In *International Conference on Machine Learning (ICML 2024)*, 2024.

Rui Pan, Yinwei Dai, Zhihao Zhang, Gabriele Oliaro, Zhihao Jia, and Ravi Netravali. SpecReason: Fast and accurate inference-time compute via speculative reasoning. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Jan Peters, Katharina Mülling, and Yasemin Altun. Relative entropy policy search. In *AAAI Conference on Artificial Intelligence (AAAI 2010)*, 2010.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations (ICLR 2025)*, 2025.

Tian Qin, Core Francisco Park, Mujin Kwun, Aaron Walsman, Eran Malach, Nikhil Anand, Hidenori Tanaka, and David Alvarez-Melis. Decomposing elements of problem solving: What "math" does RL teach? *arXiv preprint arXiv:2505.22756*, 2025.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to Q^*: Your language model is secretly a Q-function. In *Conference on Language Modeling (COLM 2024)*, 2024.

Stéphane Ross and J. Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014.

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 (ICLR 2024)*, 2024.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations (ICLR 2025)*, 2025.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. In *International Conference on Machine Learning (ICML 2026)*, 2026.

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.

Changshuo Shen, Leheng Sheng, Yuxin Chen, An Zhang, and Xiang Wang. Reasoning can be restored by correcting a few decision tokens. In *International Conference on Machine Learning (ICML 2026)*, 2026.

Shannon Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag. Learning to decode collaboratively with multiple language models. In *Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Base models know how to reason, thinking models learn when. In *International Conference on Machine Learning (ICML 2026)*, 2026.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025a.

Xuezhi Wang and Denny Zhou. Chain-of-Thought reasoning without prompting. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*, 2024.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Reinforcement learning for reasoning in large language models with one training example. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. In *International Conference on Learning Representations (ICLR 2026)*, 2026.

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 preprint arXiv:2507.14843*, 2025a.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, Yanwei Fu, Qin Liu, Songyang Zhang, and Qi Zhang. Reasoning or memorization? Unreliable results of reinforcement learning due to data contamination. *arXiv preprint arXiv:2507.10532*, 2025b.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024a.

Joy Qiping Yang, Salman Salamatian, Ziteng Sun, Ananda Theertha Suresh, and Ahmad Beirami. Asymptotics of language model alignment. *arXiv preprint arXiv:2404.01730*, 2024b.

Wang Yang, Xiang Yue, Vipin Chaudhary, and Xiaotian Han. Speculative thinking: Enhancing small-model reasoning with large model guidance at inference time. *arXiv preprint arXiv:2504.12329*, 2025.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In *International Conference on Machine Learning (ICML 2025)*, 2025.

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 (NeurIPS 2025)*, 2025.

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)*, 2025.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track (NeurIPS 2024)*, 2024.

Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. Echo chamber: RL post-training amplifies behaviors learned in pretraining. *arXiv preprint arXiv:2504.07912*, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, 2023.

A PROOFS

Lemma 1. Fix x and a trajectory $y = (y_1, \ldots, y_L)$ with states $s_1 = x$, $s_{t+1} = s_t \cdot y_t$. Transitions are deterministic and the reward terminal, and π_r's terminal states are terminal for π_b too (the end symbol and cap are shared; Appendix D gives how the implementation stops), so $V^{\pi_b}(s_{L+1}) = r(x, y)$. Then $\sum_{t=1}^{L} A^{\pi_b}(s_t, y_t) = \sum_{t=1}^{L} [V^{\pi_b}(s_{t+1}) - V^{\pi_b}(s_t)] = r(x, y) - V^{\pi_b}(x)$. Taking $\mathbb{E}_{y\sim\pi_r}$ gives $V^{\pi_r}(x) - V^{\pi_b}(x)$ and then $\mathbb{E}_x$ gives the claim. □

Proposition 1. Write $k' = \min(k, |y|)$. $R(k) = \mathbb{E}_x \mathbb{E}_{y\sim\pi_r} V^{\pi_b}(s_{k'+1})$ by definition of H_k and the tower property (after the handoff H_k is π_b, and the summand depends on the first k tokens only, whose law is the same under H_k and π_r); the telescoping sum truncated at k' gives $V^{\pi_b}(s_{k'+1}) - V^{\pi_b}(x) = \sum_{t\le k'} A^{\pi_b}(s_t, y_t)$, and $G(T) = J(\pi_r) - J(\pi_b)$ is Lemma 1. □

Corollary 1. Write $q = \pi_b(y_t \mid s_t) \in [0, 1)$ (at $q = 0$, $V_b^{\mathrm{alt}} = V^{\pi_b}$ and the room is one). By the Bellman identity $V^{\pi_b}(s_t) = q\, Q^{\pi_b}(s_t, y_t) + (1 - q)\, V_b^{\mathrm{alt}}(s_t)$, so $A_t = Q^{\pi_b}(s_t, y_t) - V^{\pi_b}(s_t) = (1 - q)[Q^{\pi_b}(s_t, y_t) - V_b^{\mathrm{alt}}(s_t)]$, and $|A_t| \le 1 - q$ because both values lie in $[0, 1]$. □

Proposition 2. Fix s at least two steps before the cap, the two next-token distributions at s, and a token y with $q = \pi_b(y \mid s) < 1$. Transitions are deterministic and rewards terminal, so $Q^{\pi_b}(s, a) = V^{\pi_b}(s \cdot a)$ is the base's success probability from the child state $s \cdot a$, a function of the rewards on the subtree below $s \cdot a$ and of the base's continuation policy there; the next-token distributions at s (and g, which also sees y) do not depend on either. Give every non-terminal child $s \cdot a$ a next step at which the base puts mass p_a on a token whose subtree carries terminal reward 1 and mass $1 - p_a$ on tokens whose subtrees carry reward 0; then $Q^{\pi_b}(s, a) = p_a$ exactly, for any $p_a \in [0, 1]$, while the end-symbol child is terminal and $Q^{\pi_b}(s, \mathrm{EOS}) = r \in \{0, 1\}$ is set by its reward. By Corollary 1 (or directly, $A = Q^{\pi_b}(s, y) - V^{\pi_b}(s)$, when $q = 0$), $A^{\pi_b}(s, y) = (1 - q)\,[Q^{\pi_b}(s, y) - V_b^{\mathrm{alt}}(s)]$; with $d = c/(1 - q) \in [-1, 1]$, set $Q^{\pi_b}(s, y) = d$ and every alternative's value to 0 when $d \ge 0$, or $Q^{\pi_b}(s, y) = 1 + d$ and every alternative's value to 1 when $d < 0$ (an end-symbol alternative gets these endpoint values from its reward); if y is itself the end symbol, set instead its reward to 1 and every alternative's value to $1 - d$ when $d \ge 0$, or its reward to 0 and every alternative's value to $-d$ when $d < 0$, the alternatives then being non-terminal. The construction re-specifies the base below s; if both policies are held fixed at every state and only the reward is free, the three values $\pm(1 - q)$ and 0 are still attained (reward 1 on every completion through $s \cdot y$ and 0 elsewhere, the reverse, and a constant reward), so the sign is undetermined for any statistic of the two policies at all states. □

Proposition 3. (i) For q_0, q_1 feasible at $\varepsilon_0, \varepsilon_1$, the mixture $\lambda q_1 + (1-\lambda)q_0$ is feasible at $\lambda\varepsilon_1 + (1-\lambda)\varepsilon_0$ by convexity of $\mathrm{KL}(\cdot\|q_b)$, and its value is linear in λ, so F_s is concave; the feasible sets are nested, so it is non-decreasing; q_b conditioned on the maximisers, $q_b(\cdot \mid A^\star)$, has $\mathrm{KL} = -\log q_b(A^\star) = \varepsilon_{\mathrm{sat}}$ and value $\max_{\mathrm{supp}\, q_b} Q - \bar{V}$, the largest possible, so F_s is constant from $\varepsilon_{\mathrm{sat}}$ on (a point mass on one maximiser has the larger KL $-\log q_b(a^\star)$ when there are ties). (ii) For any $q \ll q_b$ the Gibbs variational identity $\log \mathbb{E}_{q_b} e^{Q/\beta} = \sup_q [\mathbb{E}_q Q/\beta - \mathrm{KL}(q\|q_b)]$, attained uniquely at $q_\beta \propto q_b e^{Q/\beta}$, gives $\mathbb{E}_q Q \le \mathbb{E}_{q_\beta} Q - \beta[\mathrm{KL}(q_\beta\|q_b) - \mathrm{KL}(q\|q_b)]$ for every $\beta > 0$; so among distributions with $\mathrm{KL}(q\|q_b) \le \mathrm{KL}(q_\beta\|q_b)$ none beats q_β. With $\lambda = 1/\beta$, $\mathrm{KL}(q_\lambda\|q_b)$ has derivative $\lambda \,\mathrm{Var}_{q_\lambda} Q > 0$ when $\sigma_b > 0$, and runs from 0 at $\lambda = 0$ to $\varepsilon_{\mathrm{sat}}$ as $\lambda \to \infty$ (the tilt concentrates on the maximisers), so every $\varepsilon \in (0, \varepsilon_{\mathrm{sat}})$ is hit by exactly one β, and F_s is strictly increasing there. (iii) Assume $\sigma_b > 0$. Let $X = Q - \bar{V}$, $\kappa_3 = \mathbb{E}_{q_b} X^3$ its third central moment, and $\kappa(\lambda) = \log \mathbb{E}_{q_b} e^{\lambda X} = \lambda^2\sigma_b^2/2 + \lambda^3\kappa_3/6 + O(\lambda^4)$, analytic because the support is finite. The tilt q_λ gains $\kappa'(\lambda) = \lambda\sigma_b^2 + \lambda^2\kappa_3/2 + O(\lambda^3)$ at $\mathrm{KL}(q_\lambda\|q_b) = \lambda\kappa'(\lambda) - \kappa(\lambda) = \lambda^2\sigma_b^2/2 + \lambda^3\kappa_3/3 + O(\lambda^4)$; inverting, $\lambda = \sqrt{2\varepsilon}/\sigma_b - 2\varepsilon\kappa_3/(3\sigma_b^4) + O(\varepsilon^{3/2})$, and substituting into the gain gives $F_s(\varepsilon) = \sigma_b\sqrt{2\varepsilon} + \kappa_3\varepsilon/(3\sigma_b^2) + O(\varepsilon^{3/2})$ by (ii); the constants are not uniform in s as $\sigma_b \to 0$. (iv) For $\pi \ll q_b$ at s, $\mathbb{E}_\pi[Q - \bar{V}] = \mathbb{E}_{q_b}[(\pi/q_b - 1)(Q - \bar{V})]$, and Cauchy–Schwarz with $\|\pi/q_b - 1\|_{q_b}^2 = \chi^2(\pi\|q_b)$ and $\|Q - \bar{V}\|_{q_b}^2 = \sigma_b^2$ gives $|\mathbb{E}_\pi[Q - \bar{V}]| \le \sigma_b\sqrt{\chi^2(\pi\|q_b)}$, with equality iff $\pi/q_b - 1$ is a real multiple of $Q - \bar{V}$ (a non-negative one for the one-sided bound $\mathbb{E}_\pi[Q - \bar{V}] \le \sigma_b\sqrt{\chi^2}$), or $\sigma_b = 0$; $\mathbb{E}_\pi[Q - \bar{V}] = \mathbb{E}_{a\sim\pi} A^{\pi_b}(s, a)$ by the Bellman identity, and for $\pi = \delta_{y_t}$, $\chi^2 = 1/q - 1 = e^{\varepsilon_t} - 1$. □

Allocation rule. Writing $\sigma_t = \sigma_b(s_t)$, the rule is stated for the linear-response surrogate $\tilde{F}_{s_t}(\varepsilon) = \sigma_t\sqrt{2\varepsilon}$, concave in ε, so among positions with $\sigma_t > 0$ equalising marginal returns $\tilde{F}'_{s_t} = \sigma_t/\sqrt{2\varepsilon}$ is necessary and sufficient (KKT; positions with $\sigma_t = 0$ get $\varepsilon_t^\star = 0$) and gives $\varepsilon_t^\star \propto \sigma_t^2$; substituting $\varepsilon_t^\star = E\sigma_t^2/\sum_u \sigma_u^2$, with $E = \sum_t \varepsilon_t$ the realised total rate, gives $\sum_t \sigma_t\sqrt{2\varepsilon_t^\star} = \sqrt{2E\sum_t \sigma_t^2}$. □

B THE RATE–VALUE FRONTIER IN FULL

Section 2 defines F_s and F_s^{det}. Setlur et al. (2025) lower-bound one *training* step by the variance of a *prover* policy's advantages; Proposition 3 instead gates the gain extractable at each state of a finished policy by the base's own dispersion (zero where it is zero; at most $\sigma_b\sqrt{e^\varepsilon - 1}$ for a single token, though F_s itself admits no bound of the form $C(\varepsilon)\sigma_b$), a base-weighted analogue of the action gap of classical RL (Farahmand, 2011; Bellemare et al., 2016). Every estimate of them, like $\hat{A}_t$, is computed on the measured action set (the base's top four tokens plus the realised one) with q_b renormalised over it, at the realised token's unrenormalised rate ε_t (the point mass at y_t stays feasible, since renormalising only lowers its surprisal), so moves of mass to unmeasured tokens are not counted. Two effects of this truncation have opposite signs: the budget ε_t exceeds the realised token's renormalised surprisal, $\varepsilon_t - |\log q_b(\mathcal{M}_t)|$, which enlarges the feasible set and raises $\hat{F}$, while dropping unmeasured tokens that could be feasible lowers it; so Λ is a floor against selection only (Appendix D), not against truncation. The remaining frontier quantities are these.

Budget-matched search. $F_s^{\mathrm{srch}}(\varepsilon) = \sum_a Q(a)\,[\Phi(a)^N - \Phi(a^-)^N] - \bar{V}$ with $N = e^\varepsilon$, Φ the cumulative base mass along the tokens sorted by increasing Q and a^- the token before a in that order (ties broken by a fixed order): the gain from *sampling* N base tokens and keeping the one of highest base value, the expected maximum of N draws for integer N and its continuous extension otherwise, since e^ε is rarely an integer. The null selects by the exact Q (a token-level value oracle, not a verifier over completions). Its budget is matched by sample count, not by divergence: best-of-N's KL from the base is at most $\log N - (N-1)/N = \varepsilon - (1 - e^{-\varepsilon}) \le \varepsilon$ (Beirami et al., 2024; Gao et al., 2023), strictly for $\varepsilon > 0$ (for non-integer N the continuous extension obeys the same bound: its KL is exactly $\log N - (N-1)/N$ for a continuous base, and data processing gives $\le$ for a discrete one), 3.6–15.6% of ε_t at the RLVR cells' mean rates of 0.07–0.35 nats and more at higher ones (37% at one nat), so the comparator spends less divergence than the trained token's surprisal: conservative for the margin under that accounting (not against the KL of π_r's full next-token distribution, which excludes its entropy). The *search margin* is $A_t - F_{s_t}^{\mathrm{srch}}(\varepsilon_t)$, positive only where the trained token bought more than e^{ε_t} base draws with an oracle over their values would have.

The point-mass bound. Proposition 3(iv) at the point mass at y_t gives $|A_t| \le \sigma_b(s_t)\sqrt{e^{\varepsilon_t}-1}$: $\sigma_b\sqrt{\varepsilon_t}$ to first order, a factor $\sqrt{2}$ below the linear-response term $\sigma_b\sqrt{2\varepsilon}$ (a point mass converts KL into χ^2 at half the rate of a small smooth tilt), and below it only for $\varepsilon_t < 1.26$; the linear-response term itself is neither an upper nor a lower bound at finite ε, and a realised token always has $\varepsilon_t \ge \sigma_b(s_t)^2$ (taking $c = Q(y_t)$, $\sigma_b^2 \le \mathbb{E}_{q_b}(Q-c)^2 \le 1 - q_b(y_t) \le \varepsilon_t$), whereas the expansion of Proposition 3(iii) needs $\varepsilon \ll \sigma_b^2$: no realised rate lies in the linear-response regime. The factorisation is an identity, the bound an inequality; Table 21 shows how often the estimates violate the bound, which tests the estimator, not the bound.

Ceiling fill and allocation. $A_t \le F^{\det}_{s_t}(\varepsilon_t)$ pointwise because y_t is itself feasible, so the sequence form $\Lambda = [J(\pi_r) - J(\pi_b)]/\mathbb{E}\sum_{t\le|y|} F^{\det}_{s_t}(\varepsilon_t)$ lies in $(0, 1]$ whenever the gain is positive, Lemma 1 summing the numerator; it is a ratio of expectations, not a mean of per-position ratios ($F^{\det}$ is negative when every token at least as probable as the realised one is worse than the base's mean). For the linear-response surrogate $\tilde{F}_{s_t}(\varepsilon) = \sigma_b(s_t)\sqrt{2\varepsilon}$, $\max \sum_t \tilde{F}_{s_t}(\varepsilon_t)$ subject to $\sum_t \varepsilon_t \le E$, E the realised total rate, is attained by $\varepsilon_t^\star \propto \sigma_b(s_t)^2$ (the *allocation rule*), and the *allocation gap* $\Gamma = \sum_t \tilde{F}_{s_t}(\varepsilon_t^\star)/\sum_t \tilde{F}_{s_t}(\varepsilon_t) \ge 1$ prices the re-allocation, a counterfactual that holds the states fixed and treats the surrogate as the objective. Since $E \ge \sum_t \sigma_t^2$, every re-allocated rate also has $\varepsilon_t^\star \ge \sigma_t^2$, outside the surrogate's regime, and $\Gamma \ge 1/\sqrt{1-w}$ when a share w of the rate sits where $\hat{\sigma}_b = 0$; so Γ (Table 17) largely restates the wasted rate and is not interpreted in the main text. The *wasted rate* is the inverse-inclusion-weighted share of a trace's total rate $\sum_t \varepsilon_t$ spent at positions where the bias-corrected dispersion estimate is at or below zero (the clipped $\hat{\sigma}_b = 0$), an estimate that shrinks as m grows (Section 4.4) but is not a bound: dead and saturated positions are always counted, while at a flat position with an interior success rate the unbiased $\hat{\sigma}_b^2$ is positive about half the time; it is a one-step quantity, credit at that step under the base's continuation, and says nothing about whether the token positioned the policy for later credit.

C Synthetic validation of the estimators

Before any language model was run, the identities, the Gibbs solution and the dispersion estimator of Section 2 were computed exactly on a depth-6, branching-6 token MDP with a Dirichlet base policy and leaf success probabilities drawn once (each rollout a Bernoulli draw at its leaf's), and each estimator was checked against the exact value. The identity of Lemma 1 holds to $2 \cdot 10^{-16}$; Proposition 1 to 10^{-16}; the Monte-Carlo transplant estimate shows no bias under a cluster-robust z-test pooled over 42 configurations (mean $z = -0.24$, so $\sqrt{42}\,\bar{z} = -1.6$); one configuration reaches $|z| = 3.7$, beyond a Bonferroni threshold of 3.2, but re-drawn with 20 fresh seeds it gives mean $z = -0.25$ (largest $|z|$ 2.4), and a fresh seed set for all 42 gives mean $z = +0.02$ and largest $|z|$ 2.2 (`code/formal/synthetic_redraw.py`), so it was a chance draw; no random feasible policy among 41,430 beats the Gibbs solution, and its KL constraint is met to 10^{-15} (the constraint is inactive above the saturation rate, as the proof predicts); the ratio $F_s(\varepsilon)/\sigma_b\sqrt{2\varepsilon}$ is 1.001 (range 0.98–1.03) at $\varepsilon \le 10^{-3}$; the gating bound has zero violations; the allocation rule reaches 99.1% of a numerically optimised allocation of the exact F_s and 124% of uniform; and the bias-corrected dispersion estimator reduces the relative bias of the naive one from 0.44 to 0.006 (clipping at zero re-introduces 0.11; the clipped estimate is used for ranking and for the zero rule of the wasted rate, the unclipped $\hat{\sigma}_b^2$ wherever a mean is pooled). These checks cover the identities, the Gibbs solution and the dispersion estimator on a full-support synthetic base; the frontier estimators are checked on a synthetic ledger (Table 20), and the action-set truncation and the temperature and nucleus of the sampler are treated in Appendix D. The correction subtracts $\sum_a q_a(1-q_a)\hat{Q}_a(1-\hat{Q}_a)/(m-1)$ from the plug-in variance: $\mathbb{E}[\sum_a q_a(\hat{Q}_a - \hat{V})^2] = \sigma_b^2 + \sum_a q_a(1-q_a)Q_a(1-Q_a)/m$ for independent Bernoulli rollouts, and $\hat{Q}_a(1-\hat{Q}_a)/(m-1)$ is unbiased for $Q_a(1-Q_a)/m$.

D Setup details

Models. Hugging Face identifiers, base first in each group:

Qwen/Qwen2.5-Math-7B	hkust-nlp/Qwen-2.5-Math-7B-SimpleRL-Zero
	sail/Qwen2.5-Math-7B-Oat-Zero
	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
	stellalisy/rethink_rlvr_reproduce-random-qwen2.5_math_7b-lr5e-7-kl0.00-step150
	stellalisy/rethink_rlvr_reproduce-ground_truth-qwen2.5_math_7b-lr5e-7-kl0.00-step150
Qwen/Qwen2.5-7B	hkust-nlp/Qwen-2.5-7B-SimpleRL-Zoo
	Open-Reasoner-Zero/Open-Reasoner-Zero-7B
Qwen/Qwen2.5-0.5B	hkust-nlp/Qwen-2.5-0.5B-SimpleRL-Zoo
Qwen/Qwen2.5-1.5B	hkust-nlp/Qwen-2.5-1.5B-SimpleRL-Zoo
meta-llama/Llama-3.1-8B	hkust-nlp/Llama-3.1-8B-SimpleRL-Zoo
extension runs (Appendix F)	
Qwen/Qwen2.5-Math-7B	hkust-nlp/Qwen-2.5-Math-7B-SimpleRL-Zoo
deepseek-ai/deepseek-math-7b-base	hkust-nlp/DeepSeek-Math-7B-SimpleRL-Zoo
Qwen/Qwen2.5-Math-1.5B	sail/Qwen2.5-Math-1.5B-Oat-Zero
deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	agentica-org/DeepScaleR-1.5B-Preview

The Llama base is served from the ungated re-upload `unsloth/Llama-3.1-8B` (identical weights); the two `rethink_rlvr_reproduce` checkpoints are the released reproductions of Shao et al. (2026), one recipe and step count trained on a random and on the ground-truth reward. Every main-cell base is the pretrained checkpoint its descendant was initialised from; none is an instruct variant (the R1-distilled start of the extension runs is itself post-trained). All are served in bf16 with vLLM (the Open-Reasoner-Zero checkpoint is stored in fp32 and served in bf16); the base's next-token distribution at every trace position is read from vLLM's prompt log-probabilities of a teacher-forced pass, raw temperature-1 values, top-20 plus the realised token.

Templates. Qwen-0.5B/1.5B and Llama pairs use `Question:\n{q}\nAnswer:\nLet's think step by step.\n`; the Qwen-Math-7B pair and the Qwen2.5-7B→SimpleRL-Zoo pair use the chat form `<|im_start|>system\nPlease reason step by step, and put your final answer within \boxed{}.<|im_end|>\n<|im_start|>user\n{q}<|im_end|>\n<|im_start|>assistant\n`; the Open-Reasoner-Zero pair uses that model's released prompt (a `User:`/`Assistant: <think>` conversation with the answer-format instruction its training prompts carry), and the base is scored under the same prompt; the distilled model uses its own chat template, with the same instruction appended to the user turn and the assistant turn opened with `<think>\n`; every prompt is given verbatim in `code/formal/fcommon.py`. The pilot that fixed these choices (12 GSM8K problems, 8 samples each, run on the trained model of each pair; the MATH-500 cells inherit their pair's template) is Table 1; it also scores two bases, and on `Qwen2.5-7B` the base alone would have preferred the Question/Answer form, so the pair's template favours the trained model there; the few-shot baseline, which keeps the pair's template, does not remove this (it elicits -1% and 4% of the gap on that base's two pairs), and no Question/Answer baseline was run on it.

Table 1: Template pilot: accuracy on the first 12 GSM8K test problems (which the cells also score), 8 samples each, temperature 0.8. Each RLVR pair uses the trained model's best form for both policies; the distilled model uses its own chat template; the two reward controls keep the base's template (Appendix D).

model	Question/Answer	chat, boxed	R1 chat	ORZ prompt	best
Qwen-2.5-0.5B-SimpleRL-Zoo	0.46	0.25	–	–	Question/Answer
Qwen-2.5-1.5B-SimpleRL-Zoo	0.67	0.09	–	–	Question/Answer
Qwen2.5-1.5B (base)	0.46	0.03	–	–	Question/Answer
Qwen-2.5-Math-7B-SimpleRL-Zero	0.62	0.72	–	–	chat, boxed
Qwen2.5-Math-7B-Oat-Zero	0.77	0.85	0.77	–	chat, boxed
DeepSeek-R1-Distill-Qwen-7B	–	–	0.83	–	R1 chat
Llama-3.1-8B-SimpleRL-Zoo	0.78	0.45	–	–	Question/Answer
Qwen-2.5-7B-SimpleRL-Zoo	0.65	0.86	–	–	chat, boxed
Qwen2.5-7B (base)	0.66	0.53	–	–	Question/Answer
Open-Reasoner-Zero-7B	0.89	0.88	–	0.92	ORZ prompt

Frontier estimators. F^{det} and F^{srch} are a maximum and an expected order statistic over $\hat{Q}$, so the plug-in estimate selects on Monte-Carlo noise and is biased *upward*. Two debiased estimates remove that selection (a two-fold split-half estimate that selects and orders actions with one half of the rollouts and values them with the other, and a leave-one-out estimate (the "cross-fitted" one) that selects with $m-1$ and values with the held-out rollout), but both choose on noise independent of the valuation, with fewer rollouts, and are therefore biased *downward*. Table 20 checks that ordering where the answer is known: on a synthetic ledger with exact action values the plug-in estimate is 4.4× the truth at m=8, 2.0× at m=64, and the debiased ones 0.34–0.62× and 0.72–0.80×; on the 3 cells re-run at m=64, with the 64-rollout estimate as reference, the m=8 plug-in is 1.5–2.2×. The bracket $[\min(\text{split}, \text{cross-fit}),\ \text{naive}]$ contains the reference in 12 of 13 checks (for the search margin the same two checks, `code/search_bracket_validation.py`: $[\text{naive}, \text{cross-fit}]$ brackets the true margin, and $\hat{F}^{\mathrm{srch}}$'s two estimates the true F^{srch}, on the synthetic ledger at every m, and on the 3 cells re-run at m=64 the m'=8 bracket contains both 64-rollout bounds on 3). Two consequences. First, Λ computed from the plug-in ceiling is a *floor* against selection (truncating the action set has no fixed sign; Appendix B), and the right way to raise it is more rollouts, not debiasing. Second, a debiased ceiling that recovers a third of the truth inflates Λ threefold, which is why 22 debiased Λ entries in Table 18 exceed 100%—an impossible value for a fraction, and a property of the estimator at m=8–32, not of any policy. Those entries are printed rather than suppressed; they should be read as "the debiased ceiling is too noisy here to divide by". Λ's denominator $\sum_t F^{\mathrm{det}}(\varepsilon_t)$ is estimated per trace as that trace's length times its own inverse-inclusion-weighted mean sampled ceiling, then averaged over traces; the pooled alternative (mean length times mean ceiling) ignores the covariance between the two and gives the smaller Λ, i.e. the larger ceiling, in 25 of 39 rows; both are in Table 18.

Dispersion and the advantage share rollouts. Since $\hat{A}_t = (1-\tilde{q}_b(y_t))(\hat{Q}(s_t,y_t) - \hat{V}_b^{\mathrm{alt}})$ exactly, with $\tilde{q}_b$ the base's probabilities renormalised over the measured set (so the estimated room is $1 - e^{-\varepsilon_t}/q_b(\mathcal{M}_t)$, a function of the rate up to the measured mass), the dispersion used in every ranking comparison is computed from the *alternative* actions only, whose rollouts are disjoint from the *realised* action's—but $\hat{\sigma}_b^{\mathrm{alt}}$ and $\hat{V}_{\mathrm{alt}}$ are built from the same alternative rollouts, so noise in those rollouts inflates both statistics together and does manufacture correlation. That is not a defect to be argued away but a null to be measured: redrawing every outcome at its position's pooled rate reproduces $r_{\mathrm{s}} = 0.16$–0.43 with no real dispersion at all, 51–91% of the observed 0.26–0.56. Only the excess over that null, and its behaviour as m grows, is evidence. Shared rollouts are not the whole mechanism: on independent rollouts (Table 32) the null still produces +0.21, because both statistics scale with the position's success rate and tie at zero where every rollout fails or succeeds. For exact values $\{\sigma_b = 0\} \subseteq \{\sigma_b^{\mathrm{alt}} = 0\}$; the exact sets differ where the alternatives share one value and the realised token's differs (so $A_t \neq 0$ there), and the estimated sets (Table 16) differ through that and through noise.

Limitations continued. Per-cell problem counts are 80–250 on the dense conservation check and up to 250 on the transplant curves, and the four-token lifts are below 0.05 on 6 cells (Table 14), so single-cell comparisons are at the resolution where benchmark instability is itself a concern (Hochlehnert et al., 2025); every interval here is a cluster bootstrap over problems, the counts of significant cells are one-sided and uncorrected unless stated (Holm-corrected counts are in Table 54), and several claims rest on the primary cell alone: the distilled-template crossing (Table 4; 99 problems), the placebo's headline value, the reward controls (one public checkpoint each), the few-shot-base ledger and the 45% truncation bound. Templates were chosen by a pilot on the first 12 GSM8K test problems, which the cells also score, and three checkpoints (the distilled model and the reward controls) were not chosen by it; MATH-500 answers are compared as normalised strings, without symbolic equivalence. The template pilot has 12 problems per model, the best-of-three prompt figure is selected on the evaluation problems, and the generic openings are ranked on the policy's own samples of those problems. Format credit counts the handoff teaching the base to stop: truncation accounts for at most 45% of the primary cell's four-token lift (0–57% across cells, 83% on the 1.5B Qwen-Math extension), and grading first rather than last answers moves the gaps by up to 0.082 (Appendix G); the two are bounded separately. The conservation check is one-sided: its allowance values the excluded mass at zero, exceeds the lift on 12 cells, and so cannot flag an under-count as a violation there (the shortfall is nonetheless resolved above zero on QM7B/G and QM7B-SFT/G); the over-count it would detect with 80% power (Section 4.1) is $1.43 = (1.96 + 0.84)/1.96$ times the half-width of each cell's shortfall interval, a normal approximation; extended to sixteen positions on Q7B/G

(60 problems, $m = 4$) it gives $\sum_{t \le 16} \hat{A}_t = +0.004$ against $R(16) - R(0) = +0.065$, a shortfall of +0.060 [−0.034, +0.163] against an allowance of +0.167 (consistent), which on a cell with little early lift is not informative. The whole-trace ledger's summed advantages agree with the gain only within intervals of order ± 1 per cell (± 9 on the distilled cell; Table 2); pooled over the cells they bound a per-position bias common to them at about $\pm 0.8 \times 10^{-3}$ without detecting one (Appendix D), which is still the scale of the mean realised advantage and of the search margins, so the per-position estimates that feed Sections 4.3–4.4 are not resolved from estimator bias at that scale. Λ's level is set largely by the ceiling: Λ is the gap over the headroom times the largest Λ the headroom allows (Table 19; against that largest value the RLVR policies reach 28–81% and the distilled one 73%, which is the normalised gain $[J(\pi_r) - J(\pi_b)]/[1 - J(\pi_b)]$ and carries no token-level information), and the distilled cell's small Λ comes from a summed ceiling 66 times the headroom, driven by its tokens' higher rate (each widens the set of feasible alternatives the plug-in maximises over) more than by its twofold length; the random-reward policy's Λ is 52% (Appendix D; the 8–47% of the main text covers the main RLVR cells only). The ledger is one of many conserved decompositions (any handoff schedule, or an average over orders, gives one), and one of the two single-switch ones: the reverse handoff (base first, then π_r) telescopes to the negative of the same gain under π_r's value function at base-visited states and would credit positions differently; ours prices trained tokens under the base's values, which is the question asked. The trained-token specificity tests (Tables 26, 51) run mostly at m=8 (16 runs also at m=32–64, on fewer problems) and the selection test (Table 29) at m=4 per half, where the section's own resolution series says associations are attenuated. The distilled cell's dispersion–worth excess is significant over its whole trace, most of whose positions lie beyond token 256 (Table 28), and absent on its window run at every m' (Tables 25, 27); it is reported, not resolved. The distilled policy is run at temperature 0.8 with the 2048-token cap, which truncates 24% of its samples, and its handoff is the only one that crosses chat templates and tokenizer special tokens (Appendix D). The two-policy divergences are computed from both policies' top-20 lists on the window positions of 9 cells and at the whole-trace positions of the `Qwen2.5-7B` cells (Appendix G). The Llama pair contributes no format/content decomposition (its policy has no shared opening) and its curve rests on the 34% of its traces longer than 128 tokens, a harder subset. The guard is a design choice, not a requirement of Proposition 1: counting a trace shorter than k with its own reward and keeping every trace moves $\rho(4)$ by at most 0.10 (on L8B/G, whose guard drops 66% of its traces), and leaves it at 0.55–0.72 on the Qwen-Math-7B RLVR cells and at most 0.28 on the other RLVR cells (`code/unguarded_curve.py`). On the guarded traces $R(k)$ and $R(\infty)$ are averaged over traces and $R(0)$ over their problems; weighting $R(0)$ by the traces instead, as the pathwise identity on a selected set would, moves it by at most 0.015 and $\rho(4)$ by at most 0.032 (`code/rho_weighting.py`). The generic prefix is problem-agnostic by construction (the same string for every problem), not free of problem-type content: the Llama policy's most common opening names a quantity, which is part of why it is out of distribution; and different in-distribution openings are worth different amounts (Table 38), so the format/content split is relative to the arm. A 16-token shared opening is also a less typical sample of what the policy writes than a 4-token one, and coverage falls with k (Table 31). Λ rests on five stratified positions per trace, each standing in for $L/9$ (upper rate tercile) or $L/3$ (lower two) of the trace's L positions, 27–285 on average per cell and up to 683 on a trace at the cap, and every frontier quantity is a bracket rather than a point estimate (Tables 18, 20). Γ is an untrained counterfactual on a surrogate evaluated outside its regime (Appendix B). Each RLVR recipe is one public checkpoint, so recipe and seed are not separated. Cells differ in problem count and in how many traces survive the composition guard (Tables 7, 9), so cross-cell comparisons of interval width are not comparisons of effect size.

Related work, continued. The placebo of Section 4.2 echoes the base's own reasoning paths that branching on its first token elicits (Wang & Zhou, 2024). State importance (Amir & Amir, 2018), the KL accounting of best-of-N (Beirami et al., 2024; Gao et al., 2023; Yang et al., 2024b) and value-guided decoding (Mudgal et al., 2024) are ancestors of σ_b and the search comparator (Appendix B). Emulated fine-tuning (Mitchell et al., 2024) and proxy-tuning (Liu et al., 2024) compose base and tuned distributions into one next-token distribution at every position, so no token of the composite is the tuned model's alone and its gain telescopes only along its own trajectories (Section 2); forking-token and sentence-resampling analyses (Bigelow et al., 2025; Bogdan et al., 2025) locate positions where resampling changes the outcome, whereas the ledger prices what the trained token bought under a fixed baseline. Hybrid decoders (Shen et al., 2024; Pan et al., 2025; Yang et al., 2025; Liao et al., 2025) route between a cheap and a strong model at the token or step level; for any such router π' the ledger bounds its gain over the weak model by $\mathbb{E}_{y\sim\pi'} \sum_t F_{s_t}(\mathrm{KL}(\pi'(\cdot \mid s_t) \| \pi_b(\cdot \mid s_t)))$, the

frontier evaluated at the router's own states (Lemma 1 applied to π' with each advantage bounded by the frontier at its KL; it needs $\pi' \ll \pi_b$ and is as loose as the summed ceiling of Table 19). Prefix prefilling and transplantation are used descriptively by Lin et al. (2024); Qi et al. (2025); Meng et al. (2026); Venhoff et al. (2026); what Proposition 1 adds is that with a *permanent* handoff its increments are conserved, which is what makes "this token was worth x points" well-formed. Learning to search analyses the same roll-in/roll-out policy to *train* a learner (Ross & Bagnell, 2014; Chang et al., 2015); we use it once, after training, to audit. An order-symmetric, Shapley-style attribution would average over handoff orders; the permanent handoff fixes one order, and the reverse handoff of Appendix F shows how much the order matters. Pivotal-token search (Abdin et al., 2024) estimates by rollouts how much a token moves the model's *own* success probability, an advantage under the model's value function rather than the base's; implicit process rewards (Yuan et al., 2025) score tokens by the realised log-ratio of a tuned to a reference model, the statistic whose association with worth Section 4.3 measures; nudging (Fei et al., 2025) lets a small aligned model write the tokens where the base is uncertain, an intervention that returns control and so telescopes only along its own trajectories; truncation and early-answering interventions test chain-of-thought faithfulness (Lanham et al., 2023); Qin et al. (2025) attribute GRPO's mathematical gains to execution rather than planning; and rewarding only reasoning-pattern phrases, without checking answers, matches verifier-based RL (Lv et al., 2025), the training-side counterpart of the format credit Section 4.2 measures. Chu et al. (2025) contrast SFT and RL on generalisation; the RL–distillation contrast here is priced against one value function instead. GSM-Symbolic removes memorised answers but not memorised procedures; a freshly written GSM8K analogue (Zhang et al., 2024) would test the format result against both.

Whole-trace conservation. Lemma 1 holds pathwise, so along a trace $\sum_t A_t = r(x, y) - V^{\pi_b}(x)$, and the whole-trace ledger's five inverse-inclusion-weighted positions give a Horvitz–Thompson estimate of that sum per trace. Table 2 compares its mean with the selected traces' accuracy minus the base's on the same problems: the difference covers zero on 13 of 13 cells, but with half-widths of 0.37–1.43 accuracy on the RLVR cells (one sampled position stands in for 285 positions on average on the distilled cell, up to 683, and $\hat{A}_t$ is an $m{=}8$ estimate), so the check has no power per cell; its point estimates are positive on 6 of 13 cells (-1.3 to +1.8 per thousand positions), the largest on QM7B/G, Q7B-ORZ/G and Q7B/M. On QM7B/G 90% of the excess comes from the 46 traces cut at the 2048-token cap (+3.89 [+0.21, +8.50] per trace, standard deviation 15, against +0.12 [-0.45, +0.72] on the 154 that finish, standard deviation 3.7): along a trace that runs on to the cap the base's finite-horizon value swings between zero and one, so single positions carry large advantages, and five of them standing in for up to 683 each leave the sum very noisy. The subgroup is post hoc, but its interval excludes zero, so noise may not be all of it: a per-position bias of 2×10^{-3}, which the pooled check below cannot exclude for one cell, would produce it over 2048 positions. The 16-token floor on rollout length (Appendix D) is not the cause: it binds at 0 of this cell's positions (6 of the 11,050 whole-trace positions of all 15 runs). Pooled, the check detects no bias common to the cells: per trace, the weighted mean advantage minus the mean that conservation implies averages $-0.1\ [-0.8, +0.6] \times 10^{-3}$ per position over the 13 cells (equal weights, problems resampled jointly within each benchmark), 1 of the 13 per-cell intervals of this per-position statistic excludes zero (Q7B-ORZ/G, $z = +2.05$; dividing each trace's difference by its length weights short traces more than Table 2 does, which is why the two can disagree on a cell), and the per-cell z-scores give $\chi^2 = 12.0$ on 13 degrees of freedom ($p = 0.53$; `code/wholetrace_pooled.py`). Valuing every position with the sampler's own weights (tempered to 0.8, then cut to the 0.95 nucleus of the measured actions) moves each RLVR cell's difference by at most 0.15 and leaves the primary cell's at +0.94 [+0.04, +2.03], now resolved, while the pooled figure is $-0.5\ [-1.1, +0.2] \times 10^{-3}$ ($\chi^2 = 14.3$, $p = 0.35$; `code/wholetrace_temperature.py`). The whole-trace check therefore bounds a common per-position bias at about $\pm 10^{-3}$ and resolves one or two per-cell discrepancies of 13 depending on the weights, about what chance gives, though the primary cell's is concentrated in its capped traces (above); it cannot resolve biases at the scale of the search margins, which is why conservation is tested densely, at $k \leq 4$ (and at $k \leq 16$ on Q7B/G), and the whole-trace ledger is used for population statistics rather than for sums of advantages (Λ's summed ceiling is such a sum, which is one reason Table 18 brackets it rather than quoting a point).

Ties in the rank correlations. $|A_t|$ and the value term are differences of Monte-Carlo means over m rollouts and so take few distinct values, and at a dead position both are zero; every Spearman coefficient in the paper therefore uses average ranks for ties (values rounded to 10^{-9} first, so floating-

Table 2: Whole-trace conservation. HT: the Horvitz–Thompson sum of the sampled advantages, trace length times the inverse-inclusion-weighted mean $\hat{A}_t$, averaged over the ledger traces (95% cluster bootstrap over problems); $R(\infty) - R(0)$: those traces' accuracy minus the base's 16-sample accuracy on the same problems; difference: HT minus target, paired; per-trace SD of the HT sum.

cell	traces	mean length	$\widehat{\sum_t A_t}$ (HT)	$R(\infty) - R(0)$	difference	per-trace SD
Q0.5B/G	250	261	−0.10 [−0.64, +0.39]	+0.206	−0.31 [−0.85, +0.19]	4.1
Q0.5B/M	200	477	+0.05 [−0.83, +0.77]	+0.251	−0.20 [−1.08, +0.52]	5.8
Q1.5B/G	250	294	+0.38 [−0.08, +0.83]	+0.271	+0.11 [−0.36, +0.56]	3.8
Q1.5B/M	200	538	−0.03 [−0.72, +0.66]	+0.341	−0.37 [−1.07, +0.34]	5.0
QM7B/G	200	732	+1.39 [+0.37, +2.60]	+0.401	+0.99 [−0.01, +2.19]	7.9
QM7B/M	100	483	+0.64 [−0.10, +1.49]	+0.322	+0.32 [−0.43, +1.19]	4.0
QM7B-Oat/G	150	399	+0.63 [+0.09, +1.18]	+0.446	+0.18 [−0.36, +0.72]	3.3
QM7B-SFT/G	100	1425	−0.59 [−9.17, +8.33]	+0.406	−1.00 [−9.63, +7.93]	45.5
L8B/G	200	137	+0.26 [−0.14, +0.61]	+0.429	−0.17 [−0.56, +0.18]	2.7
Q7B/G	100	303	+0.24 [−0.59, +1.12]	+0.316	−0.08 [−0.90, +0.79]	4.2
Q7B/M	80	544	+0.97 [+0.02, +2.03]	+0.281	+0.69 [−0.25, +1.75]	4.7
Q7B-ORZ/G	100	460	+0.98 [+0.06, +1.99]	+0.126	+0.85 [−0.06, +1.87]	4.9
Q7B-ORZ/M	80	963	+0.05 [−1.37, +1.50]	+0.322	−0.28 [−1.68, +1.18]	6.5

point noise does not split a tie), the standard convention. Breaking ties in array order instead (ordinal ranks) and recomputing every coefficient of Table 23, together with the dispersion coefficients on the same 9 cells (`code/tie_robustness.py`), moves none by more than 0.04 (the largest movement, downward, is the dispersion–value-term coefficient on Q0.5B/G), and no coefficient other than dispersion's by more than 0.01. On the whole-trace runs the raw dispersion coefficient moves by up to 0.06 under the other convention (Table 26, right block), the size of the smaller shared-rate excesses, so those excesses are quoted under the paper's convention only. The divergence-side statistics stay within Table 23's bound against the value term under either convention ($|r_s| \leq 0.10$), and the exchangeable-token null reaches the same verdict under both (Table 26). Every null in the paper is computed with the same estimator as the statistic it calibrates.

The distilled cell's resolution series. The nested-subsample series of Section 4.3 pools the RLVR runs, the question being what RLVR policies' tokens carry; the distilled cell's m=32 draw, the one non-RLVR run, is reported on its own (pooled with the others, the excess is +0.06, +0.12 and +0.18 at m'=8, 16, 32). Its null-calibrated excess is −0.22 at m'=8, −0.01 at 16 and −0.09 at 32, never significant, on the 120 positions that draw covers—against a significant excess on the same cell's 500 whole-trace positions at m=8 (Table 25). We report the disagreement rather than pool it away; with 120 positions the subsample is the weaker of the two measurements, but we do not treat that as settling it.

Coincident openings. Where a transplanted trace's first four tokens *are* one of the grafted generic openings, the real arm and that one generic arm hand the base the same state, so that arm's third of the pooled comparison is zero there by construction while the other two openings differ from the trace's; the pooled content on those traces mixes a mechanical zero with the other arms' content, so we re-estimate it on the traces that do not open with a grafted string. The share of such traces is 1–62% across cells, 22% on the primary cell (`code/format_coincidence.py`); re-estimating the decomposition on the rest gives, on the primary cell, real +0.259 against generic +0.261 (content −0.003 [−0.030, +0.024], 181 problems), so its zero content is measured, not mechanical. On the other two cells of that base the content that survives lives on the uncommon openings: +0.093 [+0.058, +0.132] for Oat-Zero (62% coincident) and +0.143 [+0.103, +0.183] for the distilled sibling (42% coincident).

Placebo for the format arm. The generic-prefix arm grafts the policy's most frequent opening into the base; if what it measures were something the policy *taught*, grafting the base's own most frequent opening should buy nothing. (Grafting it everywhere is itself an intervention, a sharpening of the base's opening distribution, not a zero-effect null.) On the primary cell it buys the same (Table 38): +0.267 [+0.231, +0.304] against +0.264, a difference of −0.003 [−0.022, +0.018], with the two openings covering 59% and 58% of problems in their own model's samples (the base's, `To solve this problem`, is also the policy's third, so one of the three pooled arms is the same on both sides; on the other two arms of each pool the difference is +0.006 [−0.019, +0.034]); on the other cells where the placebo was run the base's opening buys more (Q1.5B/G, L8B/G, Q7B-ORZ/G) or as much (Q7B/G). The arm therefore measures what committing to a well-formed in-distribution

opening is worth, not what this policy's opening taught—the stronger reading of Section 4.2, since the base can supply the opening itself; but openings differ (on Q1.5B/G, L8B/G, Q7B-ORZ/G the base's is worth more), so the split is relative to the arm. The Llama cell shows the estimator behaving as it should in the other direction: its policy has no shared opening (1% coverage) and the arm is strongly harmful there (-0.197), while the base's own opening, at 32% coverage, costs little (-0.021, small but resolved). Released as the `_placebo` records.

Reward controls. The two `rethink_rlvr_reproduce` checkpoints (Shao et al., 2026) are one GRPO recipe and step count applied to `Qwen2.5-Math-7B` with a random and with the ground-truth reward. They are scored like the other descendants of that base, on the primary cell's GSM8K problems against the same value function, template and cap (the ground-truth policy's 12-problem pilot preferred the Question/Answer form and its 50-problem held-out pilot the chat form (Table 42); both controls keep the base's, so that all five share one value function), and they enter no pooled statistic; Table 3 sets them beside the other three descendants. Trained on either reward the policy opens most often with the base's own most common opening, `To solve this problem` (on 64% and 52% of problems; the base 59%), and the two policies' three generic openings share two strings, so their generic arms are worth nearly the same, +0.250 and +0.252. What differs is how much of each policy's own four-token lift that opening accounts for. Without a correctness signal, all of it: +0.233 against +0.250, content -0.017 $[-0.044, +0.010]$, with $\rho(4) = 0.80$ $[0.69, 0.93]$, the highest point estimate of the five. With the ground-truth reward the lift, +0.286, exceeds the opening's by resolved content, +0.034 $[+0.015, +0.054]$, and the larger gap (0.431 against 0.298) arrives later, $\rho(128) = 0.72$ $[0.67, 0.78]$ against 0.91 $[0.81, 1.03]$. On GSM8K, then, the correct reward adds early content and later-trace value, and the early format is the base's own opening, which needs no correctness signal; on GSM-Symbolic the content split replicates but the timing does not (Appendix F). Four worked examples close 71% of the random-reward gap and 49% of the ground-truth one, and conservation at $k \leq 4$ holds on both (shortfalls +0.030 and +0.029). The dispersion excess over the shared-rate null is +0.10 and +0.16 (both $p < 0.01$), and -0.01 ($p = 0.55$) and +0.09 ($p = 0.01$) conditioned also on the success-rate variance. Λ is 52% for the random-reward policy; the ground-truth policy's summed ceiling is not resolved from zero at this resolution, so its Λ is not estimable here (the population ratio is defined, since the gain is positive).

Table 3: The five descendants of `Qwen2.5-Math-7B` on GSM8K against one value function, including one GRPO recipe trained on the correct (ground-truth) and on a random reward (Shao et al., 2026). ρ and content at $k = 4$ with 95% cluster-bootstrap intervals; the gap is over all problems (Table 6), ρ over the length-guarded traces (Table 9) and the lifts over Table 14's traces, so gap times $\rho(4)$ need not equal the lift; generic prefix and content as in Table 14; conservation: the $k \leq 4$ dense check; last column: share of the gap closed by four sampled worked examples.

descendant	$J(\pi_r)$	gap	$\rho(4)$	$\rho(128)$	$R(4) - R(0)$	generic prefix	content at k=4	conservation $k \leq 4$	4-shot base closes
SimpleRL-Zero	0.813	+0.379	0.73 [0.64,0.83]	0.79 [0.71,0.87]	+0.264	+0.264	+0.000 [−0.02,+0.02]	consistent	56%
Oat-Zero	0.893	+0.459	0.62 [0.55,0.68]	0.69 [0.63,0.74]	+0.278	+0.229	+0.049 [+0.03,+0.07]	consistent	46%
GRPO, correct reward	0.865	+0.431	0.66 [0.60,0.74]	0.72 [0.67,0.78]	+0.286	+0.252	+0.034 [+0.01,+0.05]	consistent	49%
GRPO, random reward	0.732	+0.298	0.80 [0.69,0.93]	0.91 [0.81,1.03]	+0.233	+0.250	−0.017 [−0.04,+0.01]	consistent	71%
R1-distilled (SFT)	0.846	+0.422	0.37 [0.27,0.48]	0.67 [0.57,0.79]	+0.152	+0.070	+0.082 [+0.04,+0.12]	consistent	52%

The ledger against another baseline. Table 4 gives the primary cell's transplant curve on 99 problems against four baseline policies with the same weights: the base under its own template, under the distilled policy's chat template and under the Question/Answer form, and the four-shot base (on 200 problems). The second starts +0.350 [+0.294, +0.403] higher and the four-token lift falls from +0.266 [+0.215, +0.321] to +0.002 [−0.023, +0.028], while $R(\infty)$, the transplanted traces' own accuracy, is the same 0.822 in both. The ledger is defined against whichever baseline is chosen, and this is the sense in which the early credit is a property of the baseline's prompting.

A null for the prefill path. Stage 1 prefills token ids into the base under the base's template. Handing the base its *own* Stage-0 traces is a null with a known answer: with $\pi_r = \pi_b$ every advantage has zero mean, so $R_{\text{self}}(k) = R(0)$ in expectation (a trace shorter than k counts with its own reward, so no length selection enters). On the 3 bases run (`Qwen2.5-Math-7B`, `Qwen2.5-7B` and `Qwen2.5-Math-1.5B`, each under the template its pairs use; first 100 GSM8K problems, two traces each; `code/formal/run_r9h.sh`, `code/self_transplant.py`) $R_{\text{self}}(k) - R(0)$ covers zero at every $k \in \{4, 16\}$, at most 0.020 in magnitude and +0.045 at the highest upper edge;

Table 4: The primary cell's transplant curve against four baseline policies and the same trained traces: the base under its own template, under the distilled policy's chat template and under the Question/Answer template on the same 99 problems (under the distilled policy's template the traces' whole gain over the base falls from 0.398 to 0.049, under Question/Answer to 0.234), and the base under a 4-shot prompt on 200; 95% cluster-bootstrap interval on the four-token lift. The four-token format/content split under each template: Table 43.

baseline	$R(0)$	$R(4)$	$R(16)$	$R(32)$	$R(64)$	$R(128)$	$R(\infty)$	$R(4) - R(0)$
base under its own template	0.424	0.689	0.706	0.703	0.705	0.717	0.822	+0.266 [+0.215, +0.321]
base under the distilled policy's template	0.773	0.776	0.757	0.783	0.784	0.784	0.822	+0.002 [−0.023, +0.028]
base under the Question/Answer template	0.588	0.688	0.729	0.709	0.723	0.728	0.822	+0.100 [+0.054, +0.145]
base under a 4-shot prompt (200 problems)	0.646	0.709	–	–	–	–	–	+0.063 [+0.034, +0.091]

on the primary base it is -0.020 [-0.068, +0.025] at $k = 4$, against the policy's real lift of +0.264 and the foreign prefix's +0.181. At this resolution the prefill path adds no detectable lift of its own: that bounds it well below the Qwen-Math lifts, though not below the smallest four-token lifts of Table 14.

Template mismatch, measured. The distilled policy writes under its own chat template while every RLVR pair is template-matched, and the distilled cell's rate statistics are outliers (Table 16: $\bar{\varepsilon}_t = 0.79$ nats against 0.07–0.35, top-k mass 0.899 against 0.908–0.997). Is that the template or the policy? We ran the crossing on a cell where the answer is known: the *SimpleRL-Zero* policy's traces handed to the same base reading the *distilled* policy's template, 500 positions over the whole trace, every statistic below computed the same way on both arms. The mismatch moves $\bar{\varepsilon}_t$ from 0.131 to 0.141, covering 1.4% of the distance to 0.786 (from unrounded values); it leaves the base's measured mass at 0.999 against 0.999 (weighted by inverse inclusion, so not Table 16's unweighted top-k mass) and the share of rate above one nat at 62% against 63%. Crossing chat templates therefore accounts for essentially none of the distilled cell's rate profile: those statistics are a property of that policy, not of the handoff's format. Released as the `_btmm` records. The tokenizers are the other crossing: the two models share a vocabulary, but the distilled model's end-of-thinking marker is an id the base's tokenizer assigns to a different special token, and handoffs pass ids. The marker occurs in 1,301 of 1,600 distilled samples, never before token 283 (median 1042), so it touches no transplant point up to $k = 256$, is the scored token at 0 of the 500 ledger positions and lies in the context of 63 of them, where the base conditions on an id it reads as a different special token; it costs 14.3 nats where it occurs, 8% of those traces' rate at or after it, and the mean rate *before* it is 0.85 nats against 0.78 over whole traces, so the distilled cell's rate is not that token.

Truncation. A policy scored wrong only because it hit the token cap would put a termination artefact inside the gap the ledger distributes. The 7B Qwen-Math cell is where this bites: at a 2048-token cap the base truncates 22% of its GSM8K samples and its SimpleRL-Zero descendant 22%—the same rate, so that cell's GSM8K gap is not a difference in how often the two policies truncate (it can still include what the handoff does to the base's stopping, below). The descendant's truncations are run-ons, not a stop-token failure: 81% of its capped samples have already written a boxed answer (72% two or more) and restart the solution instead of stopping, under the same stop set with which its Oat-Zero sibling truncates 0.1%. Its Oat-Zero descendant truncates only 0.1%, and on `Qwen2.5-7B` the base truncates 11.5% against its SimpleRL-Zoo descendant's 0.4%, so on those cells part of the gap is finishing inside the budget. On MATH-500 the base truncates 20% against the descendant's 9%, so part of that cell's gain *is* learning to finish inside the budget, and we say so rather than net it out; the distilled descendant, whose traces are the longest, truncates 24%. Every other policy of the main cells truncates at most 14%; Table 34 gives all of them (for the extension cells, where the 1.5B Qwen-Math base truncates 39%, see Appendix F). A related question is whether the $k = 4$ lift is the base learning to *stop* rather than to reason. Truncation does fall along the transplant curve (on the primary cell from 22% at $R(0)$ to 10% at $R(4)$), but even granting every recovered completion a correct grade, that accounts for at most 45% of that cell's lift, 0–57% across cells, 4% on the Llama pair (Table 35) and up to 83% on the 1.5B Qwen-Math extension (Appendix F); what the handoff teaches the base about stopping is counted as format by the generic arm. Whatever the mechanism, the placebo above shows the base's own opening supplies it.

Estimation details. Rate and all next-token distributions are read from the model's unmodified next-token distribution at temperature 1; rollouts use the sampling policy (temperature 0.8, top-p

0.95), whose value function is what the ledger measures. The two are therefore not the same policy: the weights q_b that enter V_b^{alt}, $\hat{\sigma}_b$ and the frontier are the temperature-1 probabilities, while the value function is the sampler's. Table 36 re-reads every whole-trace ledger with q_b tempered to 0.8 and renormalised over the measured actions, which changes the rate, the room, the weights and the feasible sets at the positions scored (the nucleus is not modelled): Λ moves by at most 5 points, the wasted-rate share by at most 4 points, the cross-fitted search margin by at most 0.0005, the dispersion–worth coefficient by at most 0.02, and the share of rate above one nat rises by 2–10 points because tempering sharpens the base; the dense conservation check stays consistent in 13 of 13 cells, with $\sum_{t\le 4}\hat{A}_t$ moving by at most 0.007 (last column). The nucleus is not modelled by the estimators (only the whole-trace variant of Appendix D applies it, over the measured actions), but its reach can be bounded from the stored top-four probabilities and tail mass (`code/nucleus_share.py`): tempered to 0.8, the realised token lies surely outside the base sampler's 0.95 nucleus at no more than 1.1% of the inverse-inclusion-weighted positions of any RLVR cell (2.3% if the unmeasured tail is as favourable as it can be), positions that carry up to 21% (40%) of the rate, and at 2.6% (9.1%) of the distilled cell's positions, carrying 15% (53%) of its rate. The sampler would never emit the token there: its rate under the sampler is infinite and its room one. There $\hat{A}_t$ is not a sampler advantage at all; elsewhere it depends on the weights only through V_b^{alt} and the room factor (the dense sums move by at most 0.007 when the weights are tempered, Table 36). The rates, the room and the rate shares at those positions are the temperature-1 model's, and Proposition 3(iv)'s gating needs the token in the sampler's support. The action set at a position is the base's top four tokens plus the realised token when it lies outside them, so a position has four or five measured actions; Table 37 counts the five-action positions, at most 2.8% of positions on the RLVR cells (1.6% inclusion-weighted) and 12.8% on the distilled one, whose tokens surprise the base most. Unless a caption says otherwise, every interval is a percentile bootstrap over problems (B=2000 resamples of problems with replacement, all of a problem's positions and rollouts moving together), the shared-rate null draws 300 redraws and the exchangeable-token null 200 (the resolution series 400, Table 27), and all seeds are fixed in the released code; the conservation verdict, which depends on an interval edge, is re-drawn under 100 independent bootstrap streams per cell in `code/conservation_seeds.py`: it holds in every stream on 12 of 13 cells and in 91% of them on Q1.5B/G, whose shortfall point estimate (-0.034) is the largest over-count in the study and whose interval's upper edge averages +0.001. Run to run, an independent second draw of the primary cell's $k = 4$ transplant and generic arms on its first 100 problems (fresh completions, the same traces; `code/formal/run_r9j.sh`) moves $R(4)$ by +0.011 [-0.008, +0.029] and the generic arm by -0.010 [-0.027, +0.007] (`code/temperature_sweep.py`), within the problem-level intervals; the distilled cell's $k = 8$ point (Table 31) is the one larger run-to-run difference we saw. The dense conservation ledger scores every one of the first four positions with m rollouts per action, $m = 6$–8 by cell (Table 8), on the same traces the transplant curve uses. Generation stops at the model's end-of-sequence or chat end token, at a new `Question:` (so a completion under the Question/Answer template cannot run on into a new problem), or at the cap; a completion after a k-token prefix is given $T - k$ new tokens, so prefix and completion share the cap, except that at least 16 are always given: that floor lets a rollout run past the cap only from the last 16 positions, which holds at 6 of the 11,050 whole-trace ledger positions and at none of the transplant or dense ones. Problems are the first N of each test split in index order (`openai/gsm8k`, `HuggingFaceH4/MATH-500`), and the traces transplanted are two of a problem's 16 Stage-0 samples drawn by a seed hashed from the policy and problem id, the first of which the dense and whole-trace ledgers walk. A completion cut off by the token cap is graded like any other, from its last boxed answer or, failing that, the last number in the text, so a truncated trace is usually but not automatically wrong. All population-level ledger statistics are inverse-inclusion weighted (below).

A consistent-policy re-run. The ledger reads its weights at temperature 1 but values positions with a sampler at 0.8 and nucleus 0.95 (above). Letting the base *sample* at temperature 1 with no nucleus removes that mismatch: the weights, the rates, the room and the frontier are then the sampling policy's own, and every token has a finite rate. We re-ran the whole-trace ledger that way on the primary cell (200 traces) and on the Oat-Zero cell (150), with the same trained traces, fresh rollouts at m=8 and the base's Stage 0 at the same settings (`code/formal/run_r9l.sh`, `run_r9m.sh`; `code/run_extensions.sh t1 t1oat`; Table 5). The base is weaker at temperature 1, so the primary cell's gap is larger (0.519 against 0.379) and Λ with it (19% against 16%), but the pricing reads the same: 72% of the rate sits at zero estimated dispersion (72% under the sampler at 0.8), the rate tracks $|\hat{A}_t|$ at $r_{\mathrm{s}} = 0.72$ and the value term at -0.05, dispersion tracks the value term at +0.30

with an excess of +0.08 over the shared-rate null ($p = 0.003$), and the cross-fitted search margin, +0.0010 [−0.0007, +0.0027], covers zero. The Oat-Zero cell reads the same way qualitatively, but with less of its rate at zero estimated dispersion (52% against 74%) and a higher Λ (64% against 47%), the value term against the rate at −0.06 and against dispersion at +0.22, and a cross-fitted margin that no longer resolves above zero (+0.0001 [−0.0019, +0.0021], against +0.0017 [+0.0001, +0.0032] under the sampler at 0.8). Over whole traces the sums cover their targets on both (+0.39 [−0.99, +2.09] and −0.38 [−1.04, +0.29]), and the primary cell's capped-trace excess, +3.89 [+0.21, +8.50] under the mixed weights, is no longer detected: +1.00 [−4.46, +7.26] on its 46 capped traces, which suggests that much of Appendix D's excess comes from the mismatch between weights and sampler rather than from the ledger.

Table 5: The whole-trace ledger with the base sampling at temperature 0.8 and nucleus 0.95 (the paper's runs, weights read at temperature 1) and at temperature 1 with no nucleus (weights and value function of one policy), same trained traces. Gap: the policy's accuracy minus the base's on the ledger's problems; wasted: share of the rate at zero estimated dispersion; Spearman correlations over positions; excess of the dispersion–worth coefficient over the shared-rate null, with its p; sum − target: whole-trace Horvitz–Thompson sum minus the trace's reward minus the base's accuracy (per trace), and the same on the traces cut at the cap. 95% cluster-bootstrap intervals over problems.

					Spearman r_s						
run	traces	gap	Λ (%)	wasted (%)	$\lvert\hat{A}_t\rvert$, rate	value, rate	value, $\hat{\sigma}_b^{\text{alt}}$	excess (p)	cross-fit margin	sum − target	capped traces (n)
QM7B/G, sampler T=0.8, top-p 0.95	200	0.379	16	72	0.62	−0.12	+0.41	+0.15 (<0.003)	−0.0000 [−0.0017,+0.0014]	+0.99 [−0.01,+2.19]	+3.89 [+0.21,+8.50] (46)
QM7B/G, sampler T=1	200	0.519	19	72	0.72	−0.05	+0.30	+0.08 (0.003)	+0.0010 [−0.0007,+0.0027]	+0.39 [−0.99,+2.09]	+1.00 [−4.46,+7.26] (46)
QM7B-Oat/G, sampler T=0.8, top-p 0.95	150	0.459	47	74	0.74	−0.15	+0.28	+0.08 (0.013)	+0.0017 [+0.0001,+0.0032]	+0.18 [−0.36,+0.72]	–
QM7B-Oat/G, sampler T=1	150	0.613	64	52	0.80	−0.06	+0.22	+0.06 (0.033)	+0.0001 [−0.0019,+0.0021]	−0.38 [−1.04,+0.29]	–

Grading. Boxed-answer extraction with brace matching; normalisation of thousands separators (but not tuple commas), fractions, roots, degree and currency marks; numeric comparison at 10^{-6} relative tolerance; bracketed lists (sets, and also tuples and intervals) compared order-free; other symbolic answers compared as normalised strings, with no symbolic equivalence. The grader's 27 unit tests ship in `code/formal/test_fgrade.py` (8 answer-extraction cases, 19 answer-comparison cases).

Generic prefix. For each trained policy and dataset we rank every 4-token opening (as token ids) among the policy's 16 Stage-0 samples per problem by the number of *distinct* problems it was produced for, ties broken by total count; the top three are run as separate arms (arm 0 the top-ranked) and pooled with equal weight. On the Qwen pairs the winner covers 10–79% of problems (e.g., `First, let's`); on the Llama pair no opening covers more than 1% of problems, which is itself why its early ledger carries no format credit. Coverage falls with k (no 16-token opening is shared by more than 16% of problems), so $k = 4$ is the primary control and $k = 8$, 16 are reported with their coverage in Table 31. The arms and their coverage are released.

Few-shot baseline. Four worked examples close 56% of the primary cell's gap, 15–52% on the other cells of the Qwen-Math, small-Qwen and Llama bases and −1% and 4% on the two `Qwen2.5-7B` pairs (Table 6); on the extension cells 53–69%, the most on `DeepSeek-Math-7B`, which does not front-load (Table 39). Like the gap itself, Λ depends on the baseline: against this arm the primary cell's floor falls from 6% [4, 11] to 3% [2, 10] on the identical window and resolution (overlapping intervals; mostly the smaller gap). The arm prepends four standard worked examples in the policy's own template, with step-by-step solutions ending in `\boxed{}`, fixed across problems and released with the code. Re-running the ledger against this few-shot base on the primary cell (a different baseline policy with a different value function) leaves the room pattern intact (0.78 for $|A_t|$ with rate, −0.13 for the value term with rate), while the dispersion correlation, +0.23 observed against +0.19 under a shared-rate null, leaves an excess of +0.04 ($p = 0.21$) that its 300 positions do not resolve ("4-shot base" rows of Tables 33 and 25).

Position sampling. Five positions per trace, drawn without replacement and uniformly within the terciles of that trace's own rates, one from each of the lower two ε-terciles and three from the upper, each weighted by its tercile's size over the number drawn from it and divided by the trace length (so a trace's weighted mean is its mean per-position value); all population-level ledger statistics (means, shares, Λ, margins) are weighted, while rank correlations are computed over the sampled positions unweighted, so they describe the stratified sample, which over-represents high-rate positions

(weighted back to the population they move by at most 0.08 on the RLVR runs; Appendix G). The primary runs stratify over the *whole* trace; the runs labelled "window" in Table 18 stratify over the first 256 tokens only and are kept because the divergence head-to-head needs both policies' distributions at the same positions.

Compute. One 24 GB consumer GPU (RTX 4090), shared with other jobs, over four sessions totalling about 52 hours of wall-clock, of which roughly 4 were model download. The extension runs of Appendix F took a fifth session of about 21 hours on the same GPU. These three figures are read off the run-log session spans by hand; they are the only quantities in the paper the released records cannot recount, and the counts that follow cover the main records only. The released records contain 1.90 million sampled completions and at least 478 million generated tokens, 57% of the completions being ledger rollouts; `code/count_generations.py` recounts both from the records, including the split by scale: 66% of the generated tokens come from the 7B and 8B cells and 34% from the sub-2B ones, which are several times faster per token. Requests are issued at a concurrency of 16–48 so that vLLM batches them; over the remaining 48 generation hours that is 2.8k output tokens per second in aggregate—a mix-weighted figure, not a claim about the 7B cells alone.

E ADDITIONAL TABLES

Every entry below is regenerated by the released analysis scripts from the released records (the divergence columns from the released per-position statistics); intervals are 95% cluster bootstraps over problems. The same lift $R(4) - R(0)$ can differ between tables (by up to half, on Q1.5B/G, whose lift is small) because each control is computed on its own problem set and traces: the transplant curves on the composition-guarded traces (those longer than the curve's largest k) (Tables 9, 10), the conservation check on the single trace the dense ledger walks (Table 8), the format arms on the problems the generic prefix was run on (Table 14). Table 9 gives the raw transplant curves, Table 10 the normalised ledger $\rho(k)$ of Figure 2 with intervals, Table 11 the length-selection check, Table 12 the paired increments between successive handoff points, Table 13 the $k \leq 256$ curve on the subset of traces longer than 256 tokens, Table 15 the foreign-prefix control at $k \in \{4, 16, 32\}$, Table 16 the population-level ledger statistics, Table 18 the frontier brackets with intervals, Table 21 the bound diagnostics, Table 23 the head-to-head among divergence-side statistics as predictors of $|A_t|$ and of its value term, Table 24 the cumulative rate spent against the cumulative value delivered (the data of Figure 5), and Table 33 the split-half reliabilities and raw correlations with $|A_t|$.

Table 6: Policies. J is pass@1 over 16 samples per problem, temperature 0.8, with 95% cluster-bootstrap intervals over problems (the Stage-0 problems of each cell, half the traces drawn in Table 7); gap $= J(\pi_r) - J(\pi_b)$; $\pi_b^{\text{4-shot}}$ is the same weights with four worked examples prepended (a different baseline policy under the convention of Section 2, stronger on every cell but one) $\text{gap}^{\text{4-shot}} = J(\pi_r) - J(\pi_b^{\text{4-shot}})$, and "elicited" is the share of the zero-shot gap the few-shot prompt closes. Rows are labelled base → trained / dataset.

base → trained / dataset	$J(\pi_b)$	$J(\pi_b^{\text{4-shot}})$	$J(\pi_r)$	gap	$\text{gap}^{\text{4-shot}}$	elicited
Qwen-0.5B→RLVR / GSM8K	0.274 [0.240, 0.309]	0.345	0.505 [0.454, 0.554]	**0.231**	0.160	31%
Qwen-0.5B→RLVR / MATH	0.129 [0.102, 0.156]	0.192	0.370 [0.320, 0.423]	**0.241**	0.178	26%
Qwen-1.5B→RLVR / GSM8K	0.481 [0.443, 0.518]	0.556	0.764 [0.724, 0.804]	**0.283**	0.208	27%
Qwen-1.5B→RLVR / MATH	0.214 [0.180, 0.251]	0.269	0.558 [0.500, 0.615]	**0.343**	0.289	16%
Qwen-Math-7B→RLVR / GSM8K	0.434 [0.401, 0.467]	0.646	0.813 [0.779, 0.845]	**0.379**	0.167	56%
Qwen-Math-7B→RLVR / MATH	0.443 [0.391, 0.492]	0.486	0.730 [0.671, 0.788]	**0.287**	0.244	15%
Qwen-Math-7B→RLVR (Dr. GRPO) / GSM8K	0.434 [0.401, 0.468]	0.646	0.893 [0.856, 0.928]	**0.459**	0.247	46%
Qwen-Math-7B→SFT-distill / GSM8K	0.424 [0.376, 0.471]	0.645	0.846 [0.791, 0.896]	**0.422**	0.201	52%
Llama-3.1-8B→RLVR / GSM8K	0.371 [0.336, 0.407]	0.438	0.813 [0.768, 0.855]	**0.443**	0.376	15%
Qwen-7B→RLVR / GSM8K	0.607 [0.571, 0.643]	0.603	0.894 [0.855, 0.929]	**0.287**	0.290	−1%
Qwen-7B→RLVR / MATH	0.489 [0.429, 0.548]	–	0.756 [0.681, 0.823]	**0.267**	–	–
Qwen-7B→RLVR (ORZ) / GSM8K	0.789 [0.747, 0.826]	0.795	0.945 [0.913, 0.972]	**0.156**	0.150	4%
Qwen-7B→RLVR (ORZ) / MATH	0.439 [0.367, 0.514]	–	0.786 [0.713, 0.854]	**0.347**	–	–

Table 7: Cell sizes. "Traces drawn" is two per problem from the Stage-0 pool; "after the guard" keeps only traces longer than the largest k, so that every point of a curve averages the same traces. The Llama cell keeps a third of its traces because its policy's answers are short, so its curve describes that policy's longer answers; every other cell keeps at least 87%.

base → trained / dataset	traces drawn	after the guard	kept	ledger positions	mean trace length
Qwen-0.5B→RLVR / GSM8K	500	447	89%	1250	261
Qwen-0.5B→RLVR / MATH	400	391	98%	1000	477
Qwen-1.5B→RLVR / GSM8K	500	489	98%	1250	294
Qwen-1.5B→RLVR / MATH	400	400	100%	1000	538
Qwen-Math-7B→RLVR / GSM8K	400	385	96%	1000	732
Qwen-Math-7B→RLVR / MATH	300	260	87%	500	483
Qwen-Math-7B→RLVR (Dr. GRPO) / GSM8K	400	400	100%	750	399
Qwen-Math-7B→SFT-distill / GSM8K	200	200	100%	500	1425
Llama-3.1-8B→RLVR / GSM8K	400	134	34%	1000	137
Qwen-7B→RLVR / GSM8K	300	298	99%	500	303
Qwen-7B→RLVR / MATH	200	200	100%	400	544
Qwen-7B→RLVR (ORZ) / GSM8K	300	300	100%	500	460
Qwen-7B→RLVR (ORZ) / MATH	200	200	100%	400	963

Table 8: Empirical conservation. $\sum_{t\le 4}\hat{A}_t$ measures the advantage at *every* one of the first four positions (four or five actions each, the realised token added when outside the base's top four, no position sampling; m rollouts per action); $R(4)-R(0)$ is the transplant curve on the same traces, trace 0 of each problem in both cases, on the problems listed. If the base mass outside the measured set ("excluded mass", last column, mean over positions) has mean value at most the renormalised mean over the measured set, renormalising overstates V^{π_b}, $\hat{A}_t$ is biased downward and the shortfall $[R(4)-R(0)]-\sum_{t\le 4}\hat{A}_t$ should be non-negative and no larger than that truncation costs; the "truncation can explain" column is that bound, obtained by valuing the excluded mass at 0 instead of V^{π_b}. "Consistent" asks whether a cluster-bootstrap interval for the shortfall overlaps that band; where the band's top exceeds the lift a ledger recording nothing would also be consistent, so the check detects over-counting only. Intervals are 95% cluster bootstraps over problems, paired across the two quantities.

cell	traces	m	$\sum_{t\le 4}\hat{A}_t$	$R(4)-R(0)$	shortfall	truncation can explain	consistent	excluded mass
Q0.5B/G	250	6	+0.022 [−0.002, +0.049]	+0.048 [+0.023, +0.071]	+0.026 [−0.004, +0.055]	$[0,+0.171]$	yes	0.163
Q0.5B/M	200	6	+0.058 [+0.031, +0.082]	+0.034 [+0.013, +0.056]	−0.024 [−0.053, +0.004]	$[0,+0.096]$	yes	0.170
Q1.5B/G	250	6	+0.050 [+0.021, +0.082]	+0.016 [−0.009, +0.042]	−0.034 [−0.070, +0.002]	$[0,+0.294]$	yes	0.160
Q1.5B/M	200	6	+0.052 [+0.017, +0.087]	+0.087 [+0.058, +0.115]	+0.036 [−0.006, +0.077]	$[0,+0.142]$	yes	0.165
QM7B/G	150	8	+0.226 [+0.181, +0.270]	+0.271 [+0.224, +0.316]	+0.046 [+0.006, +0.086]	$[0,+0.314]$	yes	0.170
QM7B/M	80	6	+0.133 [+0.067, +0.204]	+0.152 [+0.089, +0.215]	+0.019 [−0.045, +0.086]	$[0,+0.225]$	yes	0.129
QM7B-Oat/G	200	6	+0.264 [+0.222, +0.306]	+0.279 [+0.238, +0.319]	+0.015 [−0.022, +0.052]	$[0,+0.263]$	yes	0.144
QM7B-SFT/G	100	8	+0.094 [+0.029, +0.158]	+0.180 [+0.128, +0.231]	+0.086 [+0.016, +0.155]	$[0,+0.652]$	yes	0.314
L8B/G	200	6	+0.053 [+0.025, +0.080]	+0.046 [+0.018, +0.074]	−0.007 [−0.046, +0.032]	$[0,+0.310]$	yes	0.211
Q7B/G	100	6	+0.028 [−0.020, +0.074]	+0.024 [−0.023, +0.069]	−0.004 [−0.072, +0.062]	$[0,+0.236]$	yes	0.102
Q7B/M	80	6	+0.049 [+0.013, +0.088]	+0.079 [+0.035, +0.124]	+0.030 [−0.032, +0.091]	$[0,+0.159]$	yes	0.076
Q7B-ORZ/G	100	6	−0.009 [−0.048, +0.028]	+0.006 [−0.032, +0.042]	+0.015 [−0.034, +0.065]	$[0,+0.248]$	yes	0.079
Q7B-ORZ/M	80	6	+0.079 [+0.044, +0.114]	+0.068 [+0.030, +0.107]	−0.011 [−0.060, +0.037]	$[0,+0.154]$	yes	0.082

Table 9: Transplant curve $R(k)$ per cell, on the traces that pass the composition guard. $R(\infty)$ is the accuracy of the traces themselves.

cell	$R(0)$	$R(4)$	$R(16)$	$R(32)$	$R(64)$	$R(128)$	$R(\infty)$
Q0.5B/G	0.241 [0.21,0.28]	0.276 [0.24,0.31]	0.298 [0.26,0.34]	0.336 [0.29,0.38]	0.354 [0.31,0.40]	0.420 [0.37,0.47]	0.443 [0.39,0.50]
Q0.5B/M	0.123 [0.10,0.15]	0.154 [0.12,0.19]	0.216 [0.18,0.25]	0.243 [0.20,0.29]	0.275 [0.23,0.32]	0.315 [0.27,0.36]	0.363 [0.30,0.42]
Q1.5B/G	0.480 [0.45,0.52]	0.511 [0.47,0.55]	0.541 [0.50,0.58]	0.553 [0.51,0.59]	0.610 [0.56,0.66]	0.672 [0.63,0.72]	0.732 [0.68,0.78]
Q1.5B/M	0.214 [0.18,0.25]	0.295 [0.25,0.34]	0.366 [0.32,0.41]	0.387 [0.34,0.44]	0.438 [0.39,0.49]	0.477 [0.42,0.53]	0.555 [0.49,0.62]
QM7B/G	0.434 [0.40,0.47]	0.697 [0.66,0.73]	0.700 [0.66,0.74]	0.706 [0.67,0.74]	0.705 [0.66,0.74]	0.718 [0.68,0.76]	0.795 [0.74,0.84]
QM7B/M	0.445 [0.39,0.50]	0.598 [0.54,0.65]	0.596 [0.54,0.66]	0.612 [0.55,0.67]	0.623 [0.56,0.68]	0.628 [0.57,0.68]	0.704 [0.63,0.77]
QM7B-Oat/G	0.434 [0.40,0.47]	0.711 [0.67,0.75]	0.734 [0.70,0.77]	0.703 [0.66,0.74]	0.712 [0.67,0.75]	0.742 [0.70,0.78]	0.882 [0.84,0.92]
QM7B-SFT/G	0.424 [0.38,0.47]	0.576 [0.53,0.62]	0.622 [0.57,0.67]	0.648 [0.60,0.69]	0.672 [0.62,0.72]	0.698 [0.65,0.74]	0.830 [0.77,0.89]
L8B/G	0.217 [0.17,0.27]	0.223 [0.17,0.28]	0.265 [0.21,0.32]	0.291 [0.23,0.36]	0.382 [0.30,0.46]	0.542 [0.45,0.64]	0.664 [0.57,0.77]
Q7B/G	0.607 [0.57,0.64]	0.635 [0.60,0.67]	0.665 [0.63,0.70]	0.678 [0.64,0.72]	0.715 [0.68,0.76]	0.702 [0.66,0.74]	0.923 [0.88,0.96]
Q7B/M	0.489 [0.43,0.55]	0.568 [0.50,0.63]	0.603 [0.53,0.67]	0.644 [0.57,0.71]	0.660 [0.59,0.73]	0.664 [0.59,0.73]	0.770 [0.69,0.84]
Q7B-ORZ/G	0.789 [0.75,0.83]	0.798 [0.76,0.83]	0.814 [0.77,0.85]	0.851 [0.81,0.89]	0.869 [0.83,0.91]	0.873 [0.83,0.91]	0.930 [0.89,0.96]
Q7B-ORZ/M	0.439 [0.37,0.51]	0.514 [0.44,0.59]	0.567 [0.49,0.64]	0.586 [0.51,0.66]	0.608 [0.53,0.68]	0.627 [0.55,0.70]	0.780 [0.70,0.85]

Table 10: Normalised ledger $\rho(k) = [R(k) - R(0)]/[R(\infty) - R(0)]$ with cluster-bootstrap intervals, the normalisation of Section 4.1 and Figure 2: the denominator is the transplanted traces' own accuracy above the base's, not the policy's 16-sample gap $J(\pi) - J(\pi_b)$, which Table 11 gives separately.

cell	$k = 4$	$k = 16$	$k = 32$	$k = 64$	$k = 128$
Q0.5B/G	0.172 [0.07,0.27]	0.279 [0.17,0.39]	0.467 [0.35,0.60]	0.558 [0.43,0.69]	0.887 [0.78,1.02]
Q0.5B/M	0.129 [0.06,0.21]	0.386 [0.30,0.48]	0.500 [0.40,0.62]	0.633 [0.54,0.75]	0.801 [0.70,0.92]
Q1.5B/G	0.122 [0.04,0.21]	0.244 [0.16,0.33]	0.291 [0.19,0.39]	0.517 [0.42,0.62]	0.761 [0.68,0.84]
Q1.5B/M	0.235 [0.17,0.30]	0.445 [0.38,0.52]	0.507 [0.44,0.58]	0.655 [0.58,0.74]	0.772 [0.70,0.85]
QM7B/G	0.728 [0.64,0.83]	0.736 [0.66,0.82]	0.754 [0.68,0.85]	0.751 [0.67,0.84]	0.788 [0.71,0.87]
QM7B/M	0.592 [0.47,0.73]	0.584 [0.47,0.70]	0.646 [0.52,0.79]	0.688 [0.57,0.82]	0.705 [0.59,0.83]
QM7B-Oat/G	0.618 [0.55,0.68]	0.669 [0.60,0.74]	0.601 [0.53,0.67]	0.620 [0.56,0.68]	0.687 [0.63,0.74]
QM7B-SFT/G	0.374 [0.27,0.48]	0.487 [0.39,0.59]	0.553 [0.46,0.66]	0.610 [0.53,0.70]	0.674 [0.57,0.79]
L8B/G	0.012 [−0.07,0.08]	0.106 [0.02,0.18]	0.165 [0.07,0.26]	0.368 [0.24,0.49]	0.727 [0.60,0.85]
Q7B/G	0.090 [−0.01,0.17]	0.183 [0.09,0.27]	0.226 [0.14,0.30]	0.341 [0.26,0.42]	0.300 [0.21,0.38]
Q7B/M	0.278 [0.17,0.39]	0.406 [0.28,0.53]	0.552 [0.44,0.66]	0.608 [0.50,0.72]	0.623 [0.51,0.75]
Q7B-ORZ/G	0.068 [−0.11,0.21]	0.182 [0.03,0.33]	0.438 [0.29,0.58]	0.567 [0.41,0.71]	0.599 [0.44,0.75]
Q7B-ORZ/M	0.220 [0.13,0.32]	0.374 [0.29,0.47]	0.430 [0.33,0.54]	0.494 [0.39,0.61]	0.550 [0.44,0.68]

Table 11: Length selection on the composition-guarded traces. $R(0)$ is the base's accuracy on the problems whose traces pass the guard; $R(\infty)$ the accuracy of those traces; $J(\pi)$ the trained policy's 16-sample accuracy on the same problems. The length selection moves the selected traces' accuracy by −0.031 to +0.029 relative to the policy's mean, with the interval excluding zero on Q1.5B/G, Q7B/G.

cell	problems	$R(0)$, base on subset	$R(\infty)$	$J(\pi_r)$ on subset	$R(\infty) - J(\pi)$
Q0.5B/G	230	0.241 [0.21,0.28]	0.443 [0.38,0.50]	0.465 [0.42,0.51]	−0.023 [−0.05,+0.01]
Q0.5B/M	198	0.123 [0.10,0.15]	0.363 [0.30,0.42]	0.363 [0.31,0.42]	−0.000 [−0.02,+0.03]
Q1.5B/G	249	0.480 [0.44,0.52]	0.732 [0.68,0.78]	0.763 [0.72,0.80]	−0.031 [−0.06,−0.00]
Q1.5B/M	200	0.214 [0.18,0.25]	0.555 [0.49,0.62]	0.558 [0.50,0.62]	−0.003 [−0.03,+0.02]
QM7B/G	198	0.434 [0.40,0.47]	0.795 [0.74,0.84]	0.812 [0.78,0.84]	−0.017 [−0.05,+0.01]
QM7B/M	135	0.445 [0.39,0.50]	0.704 [0.63,0.77]	0.719 [0.65,0.78]	−0.015 [−0.04,+0.01]
QM7B-Oat/G	200	0.434 [0.40,0.47]	0.882 [0.84,0.92]	0.893 [0.85,0.93]	−0.011 [−0.03,+0.01]
QM7B-SFT/G	100	0.424 [0.37,0.47]	0.830 [0.77,0.89]	0.846 [0.79,0.89]	−0.016 [−0.05,+0.02]
L8B/G	75	0.217 [0.17,0.27]	0.664 [0.56,0.76]	0.669 [0.60,0.75]	−0.005 [−0.07,+0.06]
Q7B/G	150	0.607 [0.57,0.64]	0.923 [0.88,0.96]	0.894 [0.86,0.93]	+0.029 [+0.01,+0.05]
Q7B/M	100	0.489 [0.43,0.55]	0.770 [0.69,0.84]	0.756 [0.68,0.83]	+0.014 [−0.01,+0.04]
Q7B-ORZ/G	150	0.789 [0.75,0.83]	0.930 [0.89,0.96]	0.945 [0.92,0.97]	−0.015 [−0.03,+0.00]
Q7B-ORZ/M	100	0.439 [0.37,0.51]	0.780 [0.70,0.85]	0.786 [0.72,0.85]	−0.006 [−0.03,+0.02]

Table 12: Paired increments of the transplant curve between successive handoff points (same traces at both points). On QM7B/G and QM7B/M no increment after $k = 4$ excludes zero; on the 0.5B and 1.5B pairs the point estimates are positive throughout.

cell	$R(16) - R(4)$	$R(32) - R(16)$	$R(64) - R(32)$	$R(128) - R(64)$
Q0.5B/G	0.022 [−0.00,0.04]	0.038 [0.02,0.06]	0.018 [−0.00,0.04]	0.066 [0.04,0.09]
Q0.5B/M	0.062 [0.04,0.09]	0.027 [0.01,0.05]	0.032 [0.01,0.05]	0.040 [0.02,0.06]
Q1.5B/G	0.031 [0.01,0.05]	0.012 [−0.01,0.04]	0.057 [0.03,0.08]	0.062 [0.04,0.09]
Q1.5B/M	0.071 [0.05,0.10]	0.021 [0.00,0.04]	0.050 [0.03,0.07]	0.040 [0.02,0.06]
QM7B/G	0.003 [−0.02,0.03]	0.006 [−0.02,0.03]	−0.001 [−0.02,0.02]	0.013 [−0.01,0.04]
QM7B/M	−0.002 [−0.03,0.03]	0.016 [−0.01,0.04]	0.011 [−0.02,0.04]	0.004 [−0.02,0.03]
QM7B-Oat/G	0.022 [0.00,0.04]	−0.030 [−0.05,−0.01]	0.009 [−0.02,0.03]	0.030 [0.00,0.06]
QM7B-SFT/G	0.046 [0.01,0.08]	0.027 [−0.01,0.07]	0.023 [−0.02,0.06]	0.026 [−0.01,0.06]
L8B/G	0.042 [0.00,0.08]	0.026 [−0.01,0.06]	0.091 [0.05,0.14]	0.160 [0.09,0.23]
Q7B/G	0.030 [−0.00,0.06]	0.013 [−0.02,0.04]	0.036 [0.01,0.06]	−0.013 [−0.04,0.02]
Q7B/M	0.036 [−0.00,0.07]	0.041 [0.01,0.08]	0.016 [−0.02,0.05]	0.004 [−0.02,0.03]
Q7B-ORZ/G	0.016 [−0.00,0.04]	0.036 [0.02,0.06]	0.018 [−0.00,0.04]	0.004 [−0.01,0.02]
Q7B-ORZ/M	0.052 [0.02,0.08]	0.019 [−0.01,0.05]	0.022 [−0.01,0.05]	0.019 [−0.01,0.05]

Table 13: Transplant curve up to $k = 256$ on the subset of traces longer than 256 tokens ("problems" counts the problems with such a trace; on the cells whose traces all exceed 256 tokens the curve is Table 9's; elsewhere a smaller and usually harder subset, hence a lower $R(\infty)$). Point estimates only.

cell	problems	$R(0)$	$R(4)$	$R(16)$	$R(32)$	$R(64)$	$R(128)$	$R(256)$	$R(\infty)$
Q0.5B/G	129	0.135	0.154	0.181	0.198	0.180	0.271	0.285	0.318
Q0.5B/M	159	0.083	0.106	0.166	0.177	0.202	0.233	0.256	0.267
Q1.5B/G	182	0.414	0.436	0.478	0.489	0.548	0.622	0.649	0.681
Q1.5B/M	183	0.188	0.250	0.317	0.334	0.390	0.430	0.451	0.506
QM7B/G	167	0.432	0.661	0.665	0.663	0.671	0.689	0.696	0.748
QM7B/M	117	0.434	0.565	0.556	0.583	0.588	0.596	0.583	0.648
QM7B-Oat/G	189	0.431	0.707	0.727	0.694	0.703	0.737	0.756	0.872
QM7B-SFT/G	100	0.424	0.576	0.622	0.648	0.672	0.698	0.717	0.830
L8B/G	11	0.142	0.088	0.118	0.127	0.108	0.167	0.235	0.471
Q7B/G	109	0.579	0.589	0.639	0.638	0.673	0.678	0.713	0.898
Q7B/M	95	0.486	0.557	0.587	0.628	0.640	0.647	0.662	0.754
Q7B-ORZ/G	144	0.782	0.791	0.807	0.842	0.862	0.866	0.880	0.926
Q7B-ORZ/M	100	0.439	0.514	0.567	0.586	0.608	0.627	0.661	0.780

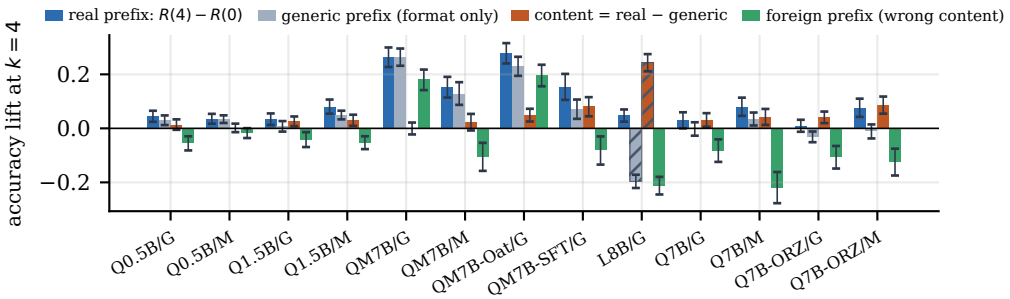


Figure 3: Format versus content at $k = 4$, 95% cluster-bootstrap intervals. A generic opening (grey) reproduces 23–100% of the real lift on 9 of the 10 Qwen cells with a resolved lift (-14% on Q7B-ORZ/M) and all of it on QM7B/G, where the content bar is zero; a foreign problem's opening (green) has a negative point estimate on 11 of 13 cells (10 resolved) and helps on QM7B/G, QM7B-Oat/G, so it only lower-bounds format. The Llama policy has no shared opening to graft (1% coverage), so its hatched bars are an out-of-distribution graft and its cost, not a decomposition.

Table 14: Format versus content at $k = 4$, all arms on the same problems (the numbers behind Figure 3). Generic prefix: the policy's most frequent 4-token opening (the one it produces for the largest number of distinct problems), the same string for every problem, pooled over the top three such openings run as separate arms (as in Table 31; Table 38's re-drawn intervals differ in the third decimal). Foreign prefix: its opening for a different problem. Content is real minus generic; format share is generic over real. The subtraction is a decomposition only where the grafted opening is in distribution for the policy: where the generic arm is itself harmful (the Llama cell, whose best opening covers 1% of problems) it measures the cost of an out-of-distribution graft, "content" absorbs that cost, and both columns are upper bounds of unknown looseness rather than measurements. Coverage per arm is in Table 31.

cell	n	real lift $R(4) - R(0)$	generic prefix	foreign prefix	content	format share
Q0.5B/G	250	+0.044 [+0.024, +0.065]	+0.030 [+0.012, +0.048]	−0.055 [−0.082, −0.029]	+0.014 [−0.005, +0.033]	69%
Q0.5B/M	200	+0.035 [+0.017, +0.054]	+0.033 [+0.020, +0.048]	−0.017 [−0.036, +0.001]	+0.001 [−0.014, +0.018]	97%
Q1.5B/G	250	+0.035 [+0.013, +0.055]	+0.008 [−0.012, +0.027]	−0.041 [−0.069, −0.014]	+0.027 [+0.008, +0.044]	23%
Q1.5B/M	200	+0.080 [+0.055, +0.107]	+0.050 [+0.033, +0.065]	−0.053 [−0.077, −0.029]	+0.030 [+0.010, +0.051]	62%
QM7B/G	200	+0.264 [+0.227, +0.300]	+0.264 [+0.232, +0.296]	+0.181 [+0.142, +0.218]	+0.000 [−0.023, +0.022]	100%
QM7B/M	149	+0.152 [+0.114, +0.191]	+0.128 [+0.087, +0.171]	−0.107 [−0.157, −0.054]	+0.023 [−0.008, +0.054]	85%
QM7B-Oat/G	200	+0.278 [+0.240, +0.316]	+0.229 [+0.195, +0.265]	+0.197 [+0.156, +0.236]	+0.049 [+0.025, +0.073]	83%
QM7B-SFT/G	100	+0.152 [+0.105, +0.202]	+0.070 [+0.035, +0.107]	−0.081 [−0.134, −0.030]	+0.082 [+0.045, +0.116]	46%
L8B/G	200	+0.047 [+0.025, +0.070]	−0.197 [−0.221, −0.172]	−0.211 [−0.245, −0.180]	+0.244 [+0.212, +0.275]	≤ 0
Q7B/G	150	+0.029 [−0.000, +0.060]	−0.002 [−0.027, +0.023]	−0.082 [−0.124, −0.041]	+0.031 [+0.006, +0.056]	≤ 0
Q7B/M	100	+0.078 [+0.046, +0.114]	+0.035 [+0.011, +0.059]	−0.219 [−0.277, −0.162]	+0.043 [+0.013, +0.072]	45%
Q7B-ORZ/G	150	+0.010 [−0.013, +0.032]	−0.031 [−0.051, −0.012]	−0.107 [−0.149, −0.065]	+0.041 [+0.019, +0.062]	≤ 0
Q7B-ORZ/M	100	+0.075 [+0.042, +0.110]	−0.010 [−0.037, +0.015]	−0.125 [−0.175, −0.075]	+0.085 [+0.054, +0.118]	≤ 0

Table 15: Foreign-prefix control. $R^{\text{foreign}}(k)$ is the base's accuracy after being prefilled with the trained policy's k-token prefix for a *different* problem (fixed derangement of the problem set); $\text{lift}_k = R^{\text{foreign}}(k) - R(0)$. Traces pass a 32-token guard; the $k = 4$ column is therefore on a slightly different problem set from Table 14, whose arms share one problem set, and can differ from it (by up to 0.01). "–": arm not run.

cell	traces	$R(0)$	$R^{\text{foreign}}(4)$	$R^{\text{foreign}}(16)$	$R^{\text{foreign}}(32)$	lift_4	lift_{16}	lift_{32}
Q0.5B/G	500	0.274 [0.24,0.31]	0.219 [0.19,0.25]	0.036 [0.03,0.05]	0.020 [0.01,0.03]	−0.055 [−0.08,−0.03]	−0.238 [−0.27,−0.21]	−0.254 [−0.29,−0.22]
Q0.5B/M	400	0.129 [0.10,0.16]	0.112 [0.09,0.14]	0.056 [0.04,0.07]	0.046 [0.03,0.06]	−0.017 [−0.04,+0.00]	−0.073 [−0.10,−0.05]	−0.083 [−0.11,−0.06]
Q1.5B/G	500	0.481 [0.44,0.52]	0.440 [0.40,0.48]	0.065 [0.05,0.08]	0.037 [0.03,0.05]	−0.041 [−0.07,−0.01]	−0.416 [−0.45,−0.38]	−0.444 [−0.48,−0.41]
Q1.5B/M	400	0.214 [0.18,0.25]	0.162 [0.13,0.20]	0.049 [0.04,0.07]	0.040 [0.03,0.05]	−0.053 [−0.08,−0.03]	−0.165 [−0.20,−0.13]	−0.174 [−0.21,−0.14]
QM7B/G	400	0.434 [0.40,0.47]	0.615 [0.58,0.65]	0.437 [0.40,0.47]	0.417 [0.39,0.45]	+0.181 [+0.14,+0.22]	+0.003 [−0.04,+0.04]	−0.016 [−0.06,+0.02]
QM7B/M	300	0.442 [0.39,0.49]	0.342 [0.30,0.39]	0.276 [0.24,0.32]	0.284 [0.24,0.33]	−0.100 [−0.15,−0.05]	−0.166 [−0.22,−0.11]	−0.158 [−0.21,−0.10]
QM7B-Oat/G	400	0.434 [0.40,0.47]	0.631 [0.59,0.67]	0.451 [0.41,0.49]	–	+0.197 [+0.16,+0.24]	+0.017 [−0.03,+0.06]	–
QM7B-SFT/G	200	0.424 [0.38,0.47]	0.343 [0.29,0.39]	0.141 [0.11,0.17]	0.136 [0.11,0.16]	−0.081 [−0.13,−0.03]	−0.283 [−0.33,−0.24]	−0.288 [−0.34,−0.24]
L8B/G	400	0.371 [0.34,0.41]	0.159 [0.13,0.19]	0.075 [0.06,0.09]	0.068 [0.06,0.08]	−0.211 [−0.25,−0.18]	−0.295 [−0.33,−0.26]	−0.302 [−0.34,−0.27]
Q7B/G	300	0.607 [0.57,0.64]	0.525 [0.48,0.57]	–	–	−0.082 [−0.13,−0.04]	–	–
Q7B/M	200	0.489 [0.43,0.55]	0.270 [0.22,0.32]	–	–	−0.219 [−0.28,−0.16]	–	–
Q7B-ORZ/G	300	0.789 [0.75,0.82]	0.682 [0.63,0.73]	–	–	−0.107 [−0.15,−0.07]	–	–
Q7B-ORZ/M	200	0.439 [0.36,0.52]	0.314 [0.25,0.38]	–	–	−0.125 [−0.18,−0.08]	–	–

Table 16: Population-level ledger statistics (inverse-inclusion weighted except the top-k mass and dead zone, which are means over positions). Top-k mass: base probability covered by the top-four alternatives plus the realised token; dead zone: fraction of positions where every rollout of every action failed; $\bar{\varepsilon}_t$, $\bar{A}_t$, $\overline{F^{\text{det}}}$, $\overline{F^{\text{srch}}}$: population means of rate, realised advantage, deterministic frontier and search null; last column: fraction of the total rate spent at positions with $\hat{\sigma}_b^{\text{alt}} = 0$ (the alternatives-only dispersion; with the full dispersion the figures are 62–89%, Table 17; for exact values the full set is contained in the alternatives-only one; the estimated sets differ, neither containing the other). Top-k mass is the mean over positions of the measured set's base probability; Table 37 reports its complement split by whether the realised token was in the top four.

cell	positions	top-k mass	dead zone	$\bar{\varepsilon}_t$	$\bar{A}_t$	$\overline{F^{\text{det}}}$	$\overline{F^{\text{srch}}}$	rate at $\hat{\sigma}_b^{\text{alt}}$=0
Q0.5B/G	1250	0.975	26%	0.205 [0.185, 0.225]	−0.0003 [−0.0021, 0.0014]	0.0050 [0.0035, 0.0064]	0.0068 [0.0057, 0.0080]	0.72 [0.67, 0.78]
Q0.5B/M	1000	0.983	38%	0.159 [0.138, 0.180]	0.0004 [−0.0014, 0.0021]	0.0038 [0.0023, 0.0055]	0.0051 [0.0039, 0.0063]	0.81 [0.74, 0.87]
Q1.5B/G	1250	0.985	10%	0.188 [0.168, 0.208]	0.0010 [−0.0006, 0.0027]	0.0057 [0.0044, 0.0072]	0.0066 [0.0056, 0.0077]	0.68 [0.61, 0.75]
Q1.5B/M	1000	0.987	24%	0.145 [0.127, 0.164]	0.0001 [−0.0015, 0.0016]	0.0033 [0.0019, 0.0048]	0.0050 [0.0038, 0.0061]	0.78 [0.72, 0.84]
QM7B/G	1000	0.990	10%	0.131 [0.109, 0.156]	0.0015 [−0.0001, 0.0032]	0.0044 [0.0030, 0.0059]	0.0051 [0.0040, 0.0062]	0.71 [0.63, 0.79]
QM7B/M	500	0.990	15%	0.119 [0.099, 0.142]	0.0024 [−0.0000, 0.0048]	0.0046 [0.0026, 0.0071]	0.0046 [0.0030, 0.0066]	0.73 [0.63, 0.83]
QM7B-Oat/G	750	0.997	6%	0.085 [0.072, 0.099]	0.0016 [0.0001, 0.0030]	0.0027 [0.0013, 0.0042]	0.0035 [0.0027, 0.0043]	0.70 [0.61, 0.78]
QM7B-SFT/G	500	0.899	2%	0.786 [0.712, 0.864]	−0.0007 [−0.0061, 0.0050]	0.0244 [0.0193, 0.0299]	0.0300 [0.0263, 0.0337]	0.67 [0.60, 0.74]
L8B/G	1000	0.908	8%	0.349 [0.321, 0.377]	0.0037 [0.0007, 0.0066]	0.0125 [0.0101, 0.0151]	0.0165 [0.0146, 0.0186]	0.64 [0.59, 0.69]
Q7B/G	500	0.990	5%	0.128 [0.103, 0.155]	0.0011 [−0.0014, 0.0035]	0.0055 [0.0031, 0.0082]	0.0068 [0.0054, 0.0084]	0.73 [0.63, 0.82]
Q7B/M	400	0.994	8%	0.073 [0.058, 0.088]	0.0017 [0.0001, 0.0034]	0.0024 [0.0009, 0.0041]	0.0032 [0.0021, 0.0044]	0.63 [0.51, 0.76]
Q7B-ORZ/G	500	0.988	6%	0.222 [0.182, 0.265]	0.0027 [0.0006, 0.0050]	0.0047 [0.0027, 0.0071]	0.0053 [0.0033, 0.0078]	0.84 [0.76, 0.91]
Q7B-ORZ/M	400	0.990	12%	0.153 [0.125, 0.186]	−0.0004 [−0.0020, 0.0010]	0.0015 [0.0004, 0.0027]	0.0027 [0.0018, 0.0037]	0.81 [0.69, 0.92]

Table 17: Frontier quantities from the whole-trace ledger (n positions), one weighting throughout. Λ: the fraction of the ceiling $\sum_t F^{\text{det}}(\varepsilon_t)$ that the gain fills, from the plug-in ceiling, which is selection-inflated and so makes Λ a *floor*—at m=8 and again at the highest resolution the cell was re-run at. "Debiased": the smaller of the split-half and cross-fitted values, printed even where it exceeds 100%, which is the estimator (the larger debiased ceiling, which this choice selects, recovers 0.62–0.90 of a known one at m=8–32, Table 20), not the policy. late/early: mean per-position ceiling after the 256th token over the one before, where the cell has enough late positions (Table 28). Search margin and frac. beating (the share of positions where $\hat{A}_t$ exceeds F^{srch}, naive/cross-fit) at the highest-resolution whole-trace run; rate at $\hat{\sigma}_b$=0 and Γ are the m=8 figures; Γ: enlargement of the linear-response surrogate under $\varepsilon_t^\star \propto \hat{\sigma}_b^2$, evaluated outside the surrogate's regime and bounded below by the wasted rate (Appendix B), so reported, not interpreted. Of the rate at $\hat{\sigma}_b$=0, 4–39% sits at dead positions (every rollout of every measured action failed), 0–45% at saturated ones (every rollout succeeded), and 44–91% where outcomes varied but the bias-corrected dispersion clipped to zero. Intervals (and point estimates) above 999% are shown as 999. Every resolution: Table 18.

		Λ			ceiling	search margin, highest-resolution whole-trace run					
cell	n	floor, m=8	floor, m=32	debiased	late/early	m	$\overline{A - F^{\text{srch}}}$ naive	$\overline{A - F^{\text{srch}}}$ cross-fit	frac. beating	rate at $\hat{\sigma}_b$=0	Γ
Q0.5B/G	1250	20% [15,29]	23% [13,56]	84%	0.21	32	−0.0064 [−0.0098,−0.0039]	−0.0034 [−0.0066,−0.0006]	0.42/0.47	71%	2.5×
Q0.5B/M	1000	19% [13,35]	–	71%	0.18	8	−0.0046 [−0.0067,−0.0027]	−0.0004 [−0.0025,+0.0014]	0.32/0.35	79%	3.1×
Q1.5B/G	1250	18% [14,23]	24% [14,48]	54%	0.23	32	−0.0029 [−0.0051,−0.0011]	−0.0008 [−0.0032,+0.0012]	0.53/0.60	66%	2.5×
Q1.5B/M	1000	27% [18,49]	–	119%	0.18	8	−0.0049 [−0.0067,−0.0031]	−0.0015 [−0.0032,+0.0002]	0.39/0.42	74%	2.8×
QM7B/G	1000	16% [11,28]	14% [7,117]	35%	0.51	32	−0.0025 [−0.0058,+0.0002]	−0.0003 [−0.0031,+0.0024]	0.59/0.63	72%	2.8×
QM7B/M	500	20% [13,39]	–	32%	0.34	8	−0.0022 [−0.0044,−0.0004]	+0.0021 [−0.0003,+0.0045]	0.50/0.55	71%	3.2×
QM7B-Oat/G	750	47% [31,95]	90% [32,999]	110%	1.43	32	−0.0022 [−0.0051,+0.0002]	−0.0008 [−0.0038,+0.0018]	0.51/0.56	74%	2.9×
QM7B-SFT/G	500	1% [1,1]	–	3%	0.69	8	−0.0307 [−0.0366,−0.0249]	−0.0062 [−0.0124,+0.0002]	0.41/0.57	62%	2.2×
L8B/G	1000	37% [29,52]	55% [42,78]	84%	–	32	−0.0088 [−0.0134,−0.0050]	+0.0007 [−0.0036,+0.0046]	0.46/0.59	62%	2.1×
Q7B/G	500	18% [12,35]	–	34%	1.07	8	−0.0058 [−0.0087,−0.0031]	+0.0014 [−0.0019,+0.0044]	0.53/0.60	71%	2.6×
Q7B/M	400	20% [11,60]	–	39%	1.79	8	−0.0014 [−0.0030,−0.0001]	+0.0016 [+0.0002,+0.0031]	0.52/0.58	66%	2.3×
Q7B-ORZ/G	500	8% [5,14]	–	12%	1.21	8	−0.0027 [−0.0045,−0.0011]	+0.0007 [−0.0010,+0.0023]	0.37/0.40	84%	3.5×
Q7B-ORZ/M	400	24% [12,145]	–	–	0.94	8	−0.0031 [−0.0051,−0.0013]	+0.0001 [−0.0018,+0.0020]	0.43/0.47	89%	4.4×

Table 18: Frontier brackets and search margins at every resolution run. Naive: $F^{\det}$ from the full-rollout estimates (maxima over noisy values, biased upward, so Λ is biased downward and the margin downward); split: two-fold split-half; cross-fit: leave-one-out, biased the other way. Λ is reported with two aggregations of the ceiling $\sum_t F^{\det}(\varepsilon_t)$: per-trace, averaging each trace's length times its own mean sampled ceiling (what the main text uses), and pooled, mean length times mean ceiling, which ignores the covariance between the two and gives the smaller Λ in 25 of the 39 rows. Rows marked "window" come from runs that sample positions over the first 256 tokens only; they multiply a window mean by the full trace length and therefore extrapolate; the whole-trace rows do not. The last column is the share of the run's total rate at positions where the clipped $\hat{\sigma}_b = 0$ (the wasted-rate figures of Section 4.4). Rows marked m=32 or m=64 re-draw the rollouts at a subset of the same positions with fresh seeds (two window runs excepted, Table 32); "4-shot base" is the same ledger computed against the few-shot base policy. Intervals are cluster bootstraps, and an interval whose upper end exceeds 999% is shown as 999. A Λ entry is blank where its per-trace ceiling total is not positive; that total is a length-weighted aggregate, so its sign can differ from the mean $\overline{F^{\det}}$ columns. *22 debiased entries exceed* 100%, which is impossible for a fraction: at these resolutions the debiased ceiling recovers only 0.34–0.90× of a known one (Table 20), so dividing by it inflates Λ several-fold. They are printed rather than suppressed, and the main text reads the naive column as a floor instead.

	$\overline{F^{\det}}$			Λ (per-trace ceiling)			Λ (pooled ceiling)		$\overline{A - F^{\mathrm{srch}}}$		rate at
cell	naive	split	cross-fit	naive	split	cross-fit	naive	cross-fit	naive	cross-fit	$\hat{\sigma}_b$=0
Q0.5B/G (window)	+0.0063 [+0.0039, +0.0089]	+0.0022 [+0.0002, +0.0044]	+0.0015 [−0.0009, +0.0039]	15% [10,27]	44% [19,392]	49% [19,518]	14% [10,21]	57% [22,479]	−0.0050 [−0.0075, −0.0026]	−0.0006 [−0.0032, +0.0018]	72%
Q0.5B/M (window)	+0.0057 [+0.0035, +0.0080]	+0.0030 [+0.0005, +0.0056]	+0.0037 [+0.0013, +0.0063]	10% [7,17]	19% [10,95]	14% [8,44]	9% [6,15]	14% [9,46]	−0.0045 [−0.0071, −0.0021]	−0.0012 [−0.0039, +0.0013]	71%
Q1.5B/G (window)	+0.0061 [+0.0042, +0.0081]	+0.0013 [−0.0006, +0.0032]	−0.0001 [−0.0022, +0.0020]	14% [11,21]	61% [27,546]	221% [32,999]	16% [12,23]	–	−0.0074 [−0.0102, −0.0047]	−0.0001 [−0.0031, +0.0027]	65%
Q1.5B/M (window)	+0.0055 [+0.0034, +0.0075]	+0.0014 [−0.0004, +0.0034]	+0.0009 [−0.0017, +0.0034]	13% [9,22]	38% [18,296]	43% [15,508]	12% [8,20]	74% [18,772]	−0.0055 [−0.0085, −0.0029]	−0.0004 [−0.0034, +0.0026]	73%
QM7B/G (window)	+0.0070 [+0.0047, +0.0092]	+0.0018 [−0.0002, +0.0039]	+0.0018 [−0.0010, +0.0044]	6% [4,11]	24% [9,307]	17% [7,233]	7% [6,11]	29% [11,290]	−0.0040 [−0.0061, −0.0020]	+0.0023 [+0.0004, +0.0043]	63%
QM7B/M (window)	+0.0061 [+0.0028, +0.0098]	+0.0025 [−0.0004, +0.0057]	+0.0021 [−0.0010, +0.0056]	11% [6,34]	75% [11,747]	27% [8,398]	10% [6,21]	29% [11,272]	−0.0055 [−0.0086, −0.0027]	−0.0009 [−0.0043, +0.0022]	59%
QM7B-Oat/G (window)	+0.0031 [+0.0012, +0.0051]	−0.0002 [−0.0019, +0.0019]	−0.0004 [−0.0026, +0.0018]	34% [18,103]	385% [35,999]	543% [39,999]	37% [22,97]	–	−0.0044 [−0.0073, −0.0021]	+0.0004 [−0.0027, +0.0031]	73%
QM7B-SFT/G (window)	+0.0243 [+0.0181, +0.0308]	−0.0023 [−0.0091, +0.0047]	−0.0016 [−0.0087, +0.0051]	1% [1,2]	–	246% [3,274]	1% [1,2]	–	−0.0341 [−0.0415, −0.0269]	+0.0001 [−0.0084, +0.0083]	66%
L8B/G (window)	+0.0128 [+0.0090, +0.0168]	+0.0055 [+0.0019, +0.0095]	+0.0064 [+0.0022, +0.0106]	31% [22,49]	63% [32,215]	53% [30,151]	27% [21,39]	53% [32,134]	−0.0147 [−0.0192, −0.0105]	−0.0031 [−0.0072, +0.0012]	60%
Q0.5B/G (whole trace)	+0.0050 [+0.0035, +0.0064]	+0.0009 [−0.0006, +0.0025]	+0.0016 [+0.0000, +0.0032]	20% [15,29]	974% [47,999]	84% [36,949]	18% [14,25]	57% [28,373]	−0.0072 [−0.0093, −0.0053]	−0.0012 [−0.0033, +0.0008]	71%
Q0.5B/M (whole trace)	+0.0038 [+0.0024, +0.0054]	+0.0006 [−0.0008, +0.0021]	+0.0011 [−0.0006, +0.0027]	19% [13,35]	119% [33,999]	71% [25,771]	13% [9,21]	48% [18,571]	−0.0046 [−0.0067, −0.0027]	−0.0004 [−0.0025, +0.0014]	79%
Q1.5B/G (whole trace)	+0.0057 [+0.0044, +0.0072]	+0.0021 [+0.0007, +0.0034]	+0.0022 [+0.0008, +0.0036]	18% [14,23]	54% [32,178]	58% [31,265]	17% [13,22]	45% [26,120]	−0.0056 [−0.0073, −0.0040]	+0.0001 [−0.0018, +0.0018]	66%
Q1.5B/M (whole trace)	+0.0033 [+0.0019, +0.0049]	+0.0008 [−0.0006, +0.0023]	+0.0010 [−0.0006, +0.0026]	27% [18,49]	279% [44,999]	119% [32,999]	19% [13,35]	64% [23,658]	−0.0049 [−0.0067, −0.0031]	−0.0015 [−0.0032, +0.0002]	74%
QM7B/G (whole trace)	+0.0044 [+0.0029, +0.0060]	+0.0022 [+0.0008, +0.0039]	+0.0020 [+0.0004, +0.0037]	16% [11,28]	35% [16,210]	63% [18,677]	12% [9,18]	26% [14,101]	−0.0035 [−0.0051, −0.0021]	−0.0000 [−0.0017, +0.0014]	72%
QM7B/M (whole trace)	+0.0046 [+0.0026, +0.0070]	+0.0021 [+0.0006, +0.0037]	+0.0018 [−0.0000, +0.0037]	20% [13,39]	32% [17,120]	38% [18,322]	13% [9,23]	33% [16,290]	−0.0022 [−0.0044, −0.0004]	+0.0021 [−0.0003, +0.0045]	71%
QM7B-Oat/G (whole trace)	+0.0027 [+0.0013, +0.0041]	+0.0010 [−0.0004, +0.0025]	+0.0007 [−0.0007, +0.0023]	47% [31,95]	110% [48,921]	189% [55,999]	43% [28,91]	158% [50,999]	−0.0019 [−0.0033, −0.0007]	+0.0017 [+0.0001, +0.0032]	74%
QM7B-SFT/G (whole trace)	+0.0244 [+0.0191, +0.0298]	+0.0046 [−0.0006, +0.0098]	+0.0067 [+0.0001, +0.0129]	1% [1,1]	5% [2,32]	3% [2,9]	1% [1,2]	4% [2,21]	−0.0307 [−0.0366, −0.0249]	−0.0062 [−0.0124, +0.0002]	62%
L8B/G (whole trace)	+0.0125 [+0.0101, +0.0151]	+0.0052 [+0.0026, +0.0077]	+0.0064 [+0.0038, +0.0093]	37% [29,52]	118% [62,555]	84% [51,301]	26% [21,32]	51% [35,90]	−0.0129 [−0.0158, −0.0100]	−0.0007 [−0.0038, +0.0023]	62%
Q7B/G (whole trace)	+0.0055 [+0.0032, +0.0078]	+0.0027 [+0.0002, +0.0052]	+0.0025 [−0.0004, +0.0053]	18% [12,35]	35% [17,215]	34% [16,274]	17% [12,31]	39% [18,260]	−0.0058 [−0.0087, −0.0031]	+0.0014 [−0.0019, +0.0044]	71%
Q7B/M (whole trace)	+0.0024 [+0.0009, +0.0040]	+0.0012 [−0.0003, +0.0029]	+0.0007 [−0.0012, +0.0026]	20% [11,60]	39% [17,376]	67% [17,721]	20% [12,56]	72% [19,916]	−0.0014 [−0.0030, −0.0001]	+0.0016 [+0.0002, +0.0031]	66%
Q7B-ORZ/G (whole trace)	+0.0047 [+0.0028, +0.0070]	+0.0025 [+0.0005, +0.0048]	+0.0029 [+0.0009, +0.0054]	8% [5,14]	14% [8,54]	12% [7,35]	7% [5,12]	12% [6,38]	−0.0027 [−0.0045, −0.0011]	+0.0007 [−0.0010, +0.0023]	84%
Q7B-ORZ/M (whole trace)	+0.0015 [+0.0003, +0.0026]	−0.0001 [−0.0015, +0.0012]	−0.0008 [−0.0026, +0.0009]	24% [12,145]	–	–	24% [13,100]	–	−0.0031 [−0.0051, −0.0013]	+0.0001 [−0.0018, +0.0020]	89%
Q0.5B/G (whole trace, m=32)	+0.0035 [+0.0015, +0.0059]	+0.0014 [−0.0009, +0.0040]	+0.0020 [−0.0003, +0.0046]	23% [13,56]	47% [19,390]	35% [17,264]	25% [15,57]	45% [19,464]	−0.0064 [−0.0098, −0.0039]	−0.0034 [−0.0066, −0.0006]	60%
Q1.5B/G (whole trace, m=32)	+0.0040 [+0.0019, +0.0063]	+0.0026 [+0.0004, +0.0051]	+0.0022 [−0.0001, +0.0048]	24% [14,48]	35% [18,125]	42% [19,331]	23% [14,49]	42% [19,304]	−0.0029 [−0.0051, −0.0011]	−0.0008 [−0.0032, +0.0012]	51%
QM7B/G (whole trace, m=32)	+0.0036 [+0.0015, +0.0062]	+0.0025 [+0.0005, +0.0047]	+0.0018 [−0.0014, +0.0049]	14% [7,117]	25% [9,246]	19% [7,222]	11% [7,30]	22% [8,216]	−0.0025 [−0.0058, +0.0002]	−0.0003 [−0.0031, +0.0024]	48%
QM7B-Oat/G (whole trace, m=32)	+0.0018 [−0.0007, +0.0048]	+0.0015 [−0.0010, +0.0045]	+0.0009 [−0.0018, +0.0038]	90% [32,999]	104% [35,999]	155% [36,999]	60% [23,682]	118% [26,999]	−0.0022 [−0.0051, +0.0002]	−0.0008 [−0.0038, +0.0018]	64%
L8B/G (whole trace, m=32)	+0.0077 [+0.0054, +0.0101]	+0.0052 [+0.0030, +0.0075]	+0.0027 [−0.0006, +0.0059]	55% [42,78]	83% [55,162]	134% [65,999]	39% [30,57]	114% [52,999]	−0.0088 [−0.0134, −0.0050]	+0.0007 [−0.0036, +0.0046]	61%
Q0.5B/G (window, m=32)	+0.0034 [+0.0012, +0.0058]	+0.0015 [−0.0009, +0.0041]	+0.0005 [−0.0022, +0.0033]	31% [17,102]	88% [26,956]	>999% [26,999]	26% [15,64]	182% [21,999]	−0.0026 [−0.0045, −0.0011]	+0.0010 [−0.0012, +0.0033]	60%
Q0.5B/M (window, m=32)	+0.0031 [+0.0007, +0.0057]	+0.0018 [−0.0006, +0.0044]	+0.0016 [−0.0012, +0.0044]	24% [13,92]	42% [16,404]	55% [17,682]	18% [10,71]	36% [12,526]	−0.0025 [−0.0053, −0.0001]	−0.0003 [−0.0030, +0.0022]	65%
Q1.5B/G (window, m=32)	+0.0042 [+0.0014, +0.0077]	−0.0009 [−0.0042, +0.0017]	−0.0010 [−0.0050, +0.0031]	21% [12,66]	–	–	21% [12,68]	–	−0.0019 [−0.0039, −0.0001]	+0.0034 [+0.0010, +0.0065]	62%
Q1.5B/M (window, m=32)	+0.0027 [+0.0006, +0.0048]	+0.0007 [−0.0011, +0.0026]	+0.0001 [−0.0026, +0.0026]	31% [15,214]	183% [22,999]	–	25% [14,82]	667% [23,999]	−0.0025 [−0.0049, −0.0005]	+0.0002 [−0.0026, +0.0025]	55%
QM7B/G (window, m=32)	+0.0068 [+0.0034, +0.0107]	+0.0055 [+0.0018, +0.0099]	+0.0040 [−0.0005, +0.0088]	5% [3,11]	8% [4,39]	9% [4,90]	6% [4,12]	10% [5,79]	−0.0015 [−0.0041, +0.0013]	+0.0037 [+0.0004, +0.0074]	63%
QM7B/M (window, m=32)	+0.0062 [+0.0012, +0.0129]	+0.0043 [−0.0004, +0.0100]	+0.0035 [−0.0023, +0.0104]	12% [5,84]	16% [6,203]	16% [6,190]	9% [4,38]	16% [5,267]	−0.0026 [−0.0059, +0.0011]	+0.0006 [−0.0033, +0.0044]	53%
QM7B-SFT/G (window, m=32)	+0.0172 [+0.0135, +0.0211]	+0.0036 [−0.0006, +0.0076]	+0.0036 [−0.0027, +0.0094]	2% [1,2]	8% [4,58]	7% [3,85]	2% [1,2]	8% [3,112]	−0.0148 [−0.0212, −0.0085]	+0.0040 [−0.0040, +0.0115]	63%
L8B/G (window, m=32)	+0.0098 [+0.0066, +0.0130]	+0.0050 [+0.0020, +0.0084]	+0.0070 [+0.0037, +0.0102]	38% [26,60]	112% [49,923]	49% [31,94]	29% [22,41]	41% [27,75]	−0.0105 [−0.0165, −0.0057]	−0.0035 [−0.0090, +0.0012]	55%
Q0.5B/G (window, m=64)	+0.0029 [+0.0007, +0.0053]	+0.0014 [−0.0004, +0.0034]	+0.0026 [+0.0003, +0.0050]	32% [17,175]	58% [25,678]	36% [18,249]	28% [15,106]	31% [16,138]	−0.0044 [−0.0072, −0.0020]	−0.0023 [−0.0053, +0.0004]	43%
Q1.5B/G (window, m=64)	+0.0043 [+0.0022, +0.0070]	+0.0031 [+0.0008, +0.0059]	+0.0033 [+0.0008, +0.0061]	20% [12,45]	28% [15,119]	29% [14,207]	21% [13,42]	28% [15,96]	−0.0027 [−0.0048, −0.0009]	−0.0015 [−0.0037, +0.0004]	50%
L8B/G (window, m=64)	+0.0086 [+0.0058, +0.0116]	+0.0055 [+0.0026, +0.0085]	+0.0046 [+0.0004, +0.0085]	28% [14,69]	48% [25,129]	37% [16,250]	30% [22,44]	57% [30,305]	−0.0070 [−0.0119, −0.0030]	−0.0030 [−0.0079, +0.0013]	41%
QM7B/G (window, 4-shot base)	+0.0085 [+0.0046, +0.0128]	+0.0006 [−0.0030, +0.0043]	+0.0025 [−0.0018, +0.0070]	3% [2,10]	36% [5,314]	14% [3,178]	5% [3,9]	17% [6,226]	−0.0069 [−0.0105, −0.0035]	+0.0012 [−0.0028, +0.0051]	56%

Table 19: How loose the position-wise ceiling is. The per-trace ceiling total $\sum_t F^{\det}(\varepsilon_t)$ of the whole-trace ledger (plug-in, as in the main text) against the accuracy headroom $1 - J(\pi_b)$, which no policy can exceed. Position-wise maxima do not compose along a trajectory, so the ceiling exceeds the headroom on every cell; "debiased / headroom" uses the larger of the split-half and cross-fitted ceilings, which exceeds the headroom on 7 cells, so the excess is resolved there only; "largest possible Λ", headroom over ceiling, is the most any policy could post against this ceiling, and Λ is the gap over the headroom (last column) times it.

cell	$J(\pi_b)$	headroom $1 - J(\pi_b)$	ceiling $\sum_t F^{\det}(\varepsilon_t)$	ceiling / headroom	debiased / headroom	Λ	largest possible Λ	gap / headroom
Q0.5B/G	0.274	0.726	1.16	1.6×	0.4×	20%	63%	32%
Q0.5B/M	0.129	0.871	1.27	1.5×	0.4×	19%	69%	28%
Q1.5B/G	0.481	0.519	1.62	3.1×	1.0×	18%	32%	55%
Q1.5B/M	0.214	0.786	1.29	1.6×	0.4×	27%	61%	44%
QM7B/G	0.434	0.566	2.42	4.3×	1.9×	16%	23%	67%
QM7B/M	0.443	0.557	1.43	2.6×	1.6×	20%	39%	52%
QM7B-Oat/G	0.434	0.566	0.97	1.7×	0.7×	47%	58%	81%
QM7B-SFT/G	0.424	0.576	37.77	65.5×	23.5×	1%	2%	73%
L8B/G	0.371	0.629	1.20	1.9×	0.8×	37%	52%	70%
Q7B/G	0.607	0.393	1.61	4.1×	2.2×	18%	24%	73%
Q7B/M	0.489	0.511	1.36	2.7×	1.3×	20%	38%	52%
Q7B-ORZ/G	0.789	0.211	1.95	9.2×	6.1×	8%	11%	74%
Q7B-ORZ/M	0.439	0.561	1.44	2.6×	< 0	24%	39%	62%

		naive		split-half		cross-fit		
reference	m	$\overline{F^{\mathrm{det}}}$	/ref	$\overline{F^{\mathrm{det}}}$	/ref	$\overline{F^{\mathrm{det}}}$	/ref	brackets?
synthetic ($F^{\mathrm{det}} = 0.0244$ exact)	8	0.1068	4.38×	0.0083	0.34×	0.0152	0.62×	yes
synthetic ($F^{\mathrm{det}} = 0.0244$ exact)	16	0.0815	3.34×	0.0141	0.58×	0.0175	0.72×	yes
synthetic ($F^{\mathrm{det}} = 0.0244$ exact)	32	0.0634	2.60×	0.0179	0.73×	0.0219	0.90×	yes
synthetic ($F^{\mathrm{det}} = 0.0244$ exact)	64	0.0494	2.03×	0.0176	0.72×	0.0195	0.80×	yes
Q0.5B/G (m=64 ref. 0.0048)	8	0.0105	2.16×	−0.0014	−0.29×	0.0049	1.00×	yes
	16	0.0075	1.55×	0.0041	0.85×	0.0019	0.39×	yes
	32	0.0059	1.21×	0.0021	0.44×	0.0017	0.34×	yes
Q1.5B/G (m=64 ref. 0.0078)	8	0.0125	1.61×	0.0031	0.40×	0.0013	0.16×	yes
	16	0.0071	0.92×	−0.0013	−0.17×	−0.0025	−0.33×	no
	32	0.0079	1.01×	0.0044	0.57×	0.0032	0.41×	yes
L8B/G (m=64 ref. 0.0149)	8	0.0222	1.49×	0.0029	0.19×	−0.0015	−0.10×	yes
	16	0.0184	1.23×	0.0028	0.19×	0.0040	0.27×	yes
	32	0.0178	1.19×	0.0107	0.72×	0.0073	0.49×	yes

Table 20: Do the frontier estimators bracket a known ceiling? Top block: a synthetic ledger of 4,000 positions with exact action values (five actions, base probabilities Dirichlet(0.45), a position success level Beta$(1.2, 1.2)$ plus Gaussian action spread of standard deviation 0, 0.02, 0.06 or 0.15 with probabilities 0.45, 0.2, 0.2, 0.15, and the realised token drawn uniformly from the five, so its rates run higher than the RLVR cells'; its exact mean F^{det} equals the distilled cell's plug-in mean by coincidence), where F^{det} is computed from the true values and the three estimators from m Bernoulli rollouts drawn at those values, using the same code as the measured cells. Lower blocks: the cells re-run at 64 rollouts per action over their first 256 tokens, with the 64-rollout plug-in estimate as the reference (itself selection-inflated, by about $2\times$ on the synthetic ledger, so these blocks test the ordering of the estimators, not their level) and the estimators recomputed on nested subsamples of the first m of the *same* rollouts; means here are unweighted over positions, whereas Table 18 weights by inverse inclusion, which is why the same runs' $\overline{F^{\mathrm{det}}}$ differ between the two tables. "/ref" is the estimate over the reference; the naive estimate is expected to exceed 1 and both debiased ones to fall below it, with exceptions where the noise is largest. "Brackets?" asks whether $[\min(\text{split}, \text{cross-fit}),\ \text{naive}]$ contains the reference.

Table 21: Bound diagnostics. The per-token bound of Proposition 3(iv) is an inequality for exact quantities; on the bias-corrected, clipped $\hat{\sigma}_b$ and the plug-in $\hat{A}_t$ it fails at the rates shown without slack (a plug-in check with the uncorrected dispersion cannot fail, since the inequality is algebraic), the second column against the linear-response term $\hat{\sigma}_b\sqrt{2\varepsilon_t}$, which is not a bound (Appendix B), and beyond a two-standard-error slack of $2\sqrt{0.25/m}$ in at most 0.3% of positions (the larger of the two columns' shares; non-zero in 4 runs). That slack is 0.35 in accuracy units at $m = 8$ (larger than all but a handful of realised advantages), so the slack test has almost no power and is not used as evidence either way; the columns count one-sided violations ($\hat{A}_t$ above the bound). The last three columns locate the rate: on most runs the larger share is spent above one nat, and the linear-response regime of Proposition 3(iii) is narrower still, $\varepsilon \ll \sigma_b^2$ (Appendix B), so no realised or re-allocated rate is in the regime the allocation rule assumes. n: positions.

cell	n	m	$\hat{A}_t > \hat{\sigma}_b\sqrt{e^{\varepsilon_t}-1}$	$\hat{A}_t > \hat{\sigma}_b\sqrt{2\varepsilon_t}$	beyond 2 SE	rate at $\varepsilon_t < 0.3$	rate at $\varepsilon_t > 1$	positions $\varepsilon_t > 1$
Q0.5B/G (window)	700	8	27%	23%	0.0%	12%	59%	12%
Q0.5B/M (window)	500	8	29%	23%	0.0%	14%	55%	9%
Q1.5B/G (window)	700	8	35%	29%	0.0%	8%	63%	12%
Q1.5B/M (window)	500	8	31%	25%	0.0%	11%	59%	9%
QM7B/G (window)	600	8	38%	31%	0.0%	9%	70%	12%
QM7B/M (window)	400	8	35%	28%	0.0%	13%	59%	8%
QM7B-Oat/G (window)	600	8	39%	31%	0.0%	14%	50%	6%
QM7B-SFT/G (window)	300	8	38%	38%	0.0%	4%	85%	47%
L8B/G (window)	500	8	33%	29%	0.2%	10%	66%	23%
Q0.5B/G (whole trace)	1250	8	23%	19%	0.0%	13%	59%	11%
Q0.5B/M (whole trace)	1000	8	24%	19%	0.0%	13%	57%	8%
Q1.5B/G (whole trace)	1250	8	30%	26%	0.0%	10%	61%	11%
Q1.5B/M (whole trace)	1000	8	28%	22%	0.1%	12%	55%	8%
QM7B/G (whole trace)	1000	8	36%	28%	0.0%	10%	63%	7%
QM7B/M (whole trace)	500	8	37%	30%	0.0%	15%	54%	7%
QM7B-Oat/G (whole trace)	750	8	38%	30%	0.0%	19%	36%	3%
QM7B-SFT/G (whole trace)	500	8	35%	32%	0.0%	3%	84%	42%
L8B/G (whole trace)	1000	8	36%	32%	0.0%	12%	64%	20%
Q7B/G (whole trace)	500	8	40%	32%	0.2%	16%	54%	7%
Q7B/M (whole trace)	400	8	37%	30%	0.0%	19%	27%	3%
Q7B-ORZ/G (whole trace)	500	8	24%	21%	0.0%	8%	76%	13%
Q7B-ORZ/M (whole trace)	400	8	32%	26%	0.0%	9%	63%	10%
Q0.5B/G (whole trace, m=32)	500	32	26%	20%	0.0%	12%	58%	12%
Q1.5B/G (whole trace, m=32)	300	32	34%	25%	0.3%	11%	60%	10%
QM7B/G (whole trace, m=32)	200	32	36%	29%	0.0%	12%	59%	6%
QM7B-Oat/G (whole trace, m=32)	200	32	44%	33%	0.0%	19%	33%	2%
L8B/G (whole trace, m=32)	300	32	36%	32%	0.0%	10%	66%	23%
Q0.5B/G (window, m=32)	300	32	29%	24%	0.0%	15%	49%	9%
Q0.5B/M (window, m=32)	250	32	30%	22%	0.0%	13%	54%	9%
Q1.5B/G (window, m=32)	160	32	34%	29%	0.0%	13%	65%	8%
Q1.5B/M (window, m=32)	250	32	26%	23%	0.0%	11%	54%	8%
QM7B/G (window, m=32)	160	32	40%	35%	0.0%	12%	59%	9%
QM7B/M (window, m=32)	160	32	37%	26%	0.0%	15%	60%	9%
QM7B-SFT/G (window, m=32)	120	32	47%	45%	0.0%	4%	85%	50%
L8B/G (window, m=32)	250	32	31%	28%	0.0%	10%	67%	24%
Q0.5B/G (window, m=64)	200	64	30%	22%	0.0%	15%	52%	10%
Q1.5B/G (window, m=64)	200	64	36%	27%	0.0%	9%	68%	12%
L8B/G (window, m=64)	200	64	27%	23%	0.0%	9%	69%	26%
QM7B/G (window, 4-shot base)	300	8	42%	35%	0.0%	9%	66%	13%

Table 22: Room and worth: Spearman correlations at the 10,050 whole-trace ledger positions, with the realised advantage (left) and its value term (right)—the per-cell form of the ranges in Section 4.3. $\hat{\sigma}_b$ here is the alternatives-only dispersion $\hat{\sigma}_b^{\mathrm{alt}}$. "Under null": the same correlation with every outcome redrawn at its position's pooled success rate (mean, 95% band, 300 draws); excess: observed minus null mean (one-sided p); partial: rank-partial controlling for $\hat{V}(1-\hat{V})$, observed/null. The four two-policy divergences against both targets, on the nine cells with a divergence pass: Table 23; every resolution and variant: Table 25; Figure 4 plots these same numbers. On the 7 RLVR window runs re-scored at m=32, rate and entropy against the value term run from −0.26 (entropy, QM7B/G) to +0.05 (`results/formal/worth.json`).

		with $\lvert A_t\rvert$		with the value term $\lvert Q_b(s_t,y_t) - V_b^{\mathrm{alt}}(s_t)\rvert$					
cell	n	ε_t	$\hat{\sigma}_b$	ε_t	H_b	$\hat{\sigma}_b$	$\hat{\sigma}_b$ under null	excess (p)	partial obs/null
Q0.5B/G	1250	0.38	+0.37	−0.07	−0.06	+0.45	+0.39 [+0.35,+0.43]	**+0.06** (<0.01)	−0.07/+0.00
Q0.5B/M	1000	0.35	+0.42	−0.01	−0.02	+0.50	+0.42 [+0.38,+0.46]	**+0.08** (<0.01)	−0.01/−0.00
Q1.5B/G	1250	0.49	+0.25	−0.11	−0.11	+0.39	+0.34 [+0.29,+0.38]	**+0.06** (<0.01)	−0.06/+0.02
Q1.5B/M	1000	0.37	+0.32	−0.08	−0.08	+0.43	+0.38 [+0.34,+0.42]	**+0.05** (0.02)	−0.07/+0.02
QM7B/G	1000	0.62	+0.15	−0.12	−0.12	+0.41	+0.27 [+0.22,+0.31]	**+0.15** (<0.01)	+0.11/+0.04
QM7B/M	500	0.61	+0.08	−0.08	−0.08	+0.40	+0.29 [+0.22,+0.34]	**+0.11** (<0.01)	+0.07/+0.04
QM7B-Oat/G	750	0.74	−0.09	−0.15	−0.16	+0.28	+0.20 [+0.13,+0.26]	**+0.08** (0.01)	+0.11/+0.06
QM7B-SFT/G	500	0.66	−0.00	−0.10	−0.14	+0.26	+0.18 [+0.10,+0.25]	**+0.09** (0.02)	+0.07/+0.03
L8B/G	1000	0.63	+0.21	+0.01	−0.01	+0.28	+0.25 [+0.20,+0.30]	**+0.03** (0.14)	−0.02/+0.03
Q7B/G	500	0.76	−0.07	−0.13	−0.13	+0.32	+0.16 [+0.08,+0.23]	**+0.15** (<0.01)	+0.19/+0.07
Q7B/M	400	0.61	+0.14	−0.10	−0.09	+0.41	+0.26 [+0.17,+0.34]	**+0.15** (<0.01)	+0.10/+0.03
Q7B-ORZ/G	500	0.18	+0.36	−0.19	−0.16	+0.56	+0.43 [+0.37,+0.50]	**+0.13** (<0.01)	−0.05/−0.05
Q7B-ORZ/M	400	0.27	+0.30	−0.24	−0.23	+0.56	+0.38 [+0.31,+0.44]	**+0.18** (<0.01)	+0.06/−0.00

Table 23: Divergence statistics head to head on the 256-token-window runs of the 9 cells with a divergence pass (n positions), against the realised advantage $|A_t|$ (left) and against its value term alone (centre). $\mathrm{KL}(\pi\|\pi_b)$ and $\mathrm{KL}(\pi_b\|\pi)$ are the two directions, JSD the symmetric one, computed between the two policies' top-20 next-token distributions on the union of the two lists, a token absent from one list taking probability 10^{-12} there and each policy's remaining mass lumped into one extra symbol—an approximation to the full-vocabulary divergences, not a bound, used here only for ranks; "partial" columns rank-residualise both $|A_t|$ and the predictor on the other predictor. By Corollary 1 $|A_t|$ contains the room factor $1 - e^{-\varepsilon_t}$ (on the measured set $1 - e^{-\varepsilon_t}/q_b(\mathcal{M}_t)$, nearly a function of the rate), and the left-hand correlations track it; the realised log-ratio does so except on QM7B/G and QM7B/M. Against the value term all 6 statistics have $|r_s| \leq 0.11$ on every cell here, of mixed sign, and rate and entropy $|r_s| \leq 0.24$ on the whole-trace runs of Table 22.

		Spearman with $\|A_t\|$ (room×worth)						Spearman with the value term (worth)						partial, $\|A_t\|$	
cell	n	ε_t	H_b	$\mathrm{KL}(\pi\|\pi_b)$	$\mathrm{KL}(\pi_b\|\pi)$	JSD	$\log \frac{\pi(y_t)}{\pi_b(y_t)}$	ε_t	H_b	$\mathrm{KL}(\pi\|\pi_b)$	$\mathrm{KL}(\pi_b\|\pi)$	JSD	$\log \frac{\pi(y_t)}{\pi_b(y_t)}$	KL| ε	ε |KL
Q0.5B/G	700	0.43	0.43	0.40	0.40	0.40	0.36	−0.04	−0.04	−0.04	−0.01	−0.04	−0.03	0.09	0.09
Q0.5B/M	500	0.44	0.42	0.40	0.37	0.39	0.37	+0.04	+0.02	+0.02	+0.03	+0.02	+0.06	−0.01	0.15
Q1.5B/G	700	0.55	0.54	0.52	0.51	0.52	0.43	−0.07	−0.07	−0.08	−0.07	−0.08	−0.06	0.09	0.22
Q1.5B/M	500	0.45	0.46	0.46	0.45	0.46	0.44	−0.04	−0.04	−0.02	−0.00	−0.02	+0.01	0.13	0.10
QM7B/G	600	0.75	0.74	0.60	0.64	0.63	0.08	−0.10	−0.10	−0.09	−0.08	−0.09	+0.10	0.00	0.66
QM7B/M	400	0.73	0.74	0.65	0.66	0.67	0.29	+0.00	+0.02	+0.05	+0.08	+0.06	+0.07	0.11	0.55
QM7B-Oat/G	600	0.80	0.79	0.80	0.77	0.80	0.73	−0.07	−0.09	−0.06	−0.07	−0.06	−0.06	0.24	0.07
QM7B-SFT/G	300	0.69	0.65	0.68	0.63	0.67	0.55	−0.08	−0.03	−0.04	−0.01	−0.03	−0.06	0.20	0.26
L8B/G	500	0.61	0.60	0.60	0.58	0.60	0.59	+0.06	+0.07	+0.06	+0.10	+0.06	+0.07	0.07	0.08

Table 24: Cumulative fraction of the trace's total rate spent by position k / cumulative fraction of the gain delivered, $\rho(k)$. The first four tokens are 0.3–2.9% of a trace but carry 1–12% of its rate, so the rate is front-loaded too—but less sharply than value: at $k = 4$ the gain delivered runs ahead of the rate spent on 12 of the 13 cells, most sharply on the 7B Qwen-Math pair, where $\rho(4) = 0.73$; at $k = 4$ only the Llama pair spends its rate faster than it delivers, and at larger k rate overtakes gain on some cells (Figure 5). Point estimates, no intervals; the traces are the composition-guarded ones of Table 9.

cell	$k = 4$	$k = 16$	$k = 32$	$k = 64$	$k = 128$
Q0.5B/G	0.12 / 0.17	0.19 / 0.28	0.35 / 0.47	0.56 / 0.56	0.75 / 0.89
Q0.5B/M	0.09 / 0.13	0.17 / 0.39	0.25 / 0.50	0.38 / 0.63	0.54 / 0.80
Q1.5B/G	0.11 / 0.12	0.18 / 0.24	0.33 / 0.29	0.51 / 0.52	0.70 / 0.76
Q1.5B/M	0.09 / 0.24	0.17 / 0.44	0.24 / 0.51	0.37 / 0.66	0.53 / 0.77
QM7B/G	0.11 / 0.73	0.18 / 0.74	0.28 / 0.75	0.42 / 0.75	0.58 / 0.79
QM7B/M	0.12 / 0.59	0.20 / 0.58	0.29 / 0.65	0.41 / 0.69	0.55 / 0.71
QM7B-Oat/G	0.10 / 0.62	0.17 / 0.67	0.27 / 0.60	0.41 / 0.62	0.59 / 0.69
QM7B-SFT/G	0.01 / 0.37	0.03 / 0.49	0.05 / 0.55	0.08 / 0.61	0.14 / 0.67
L8B/G	0.11 / 0.01	0.25 / 0.11	0.46 / 0.16	0.72 / 0.37	0.93 / 0.73
Q7B/G	0.08 / 0.09	0.16 / 0.18	0.31 / 0.23	0.49 / 0.34	0.69 / 0.30
Q7B/M	0.06 / 0.28	0.12 / 0.41	0.20 / 0.55	0.34 / 0.61	0.49 / 0.62
Q7B-ORZ/G	0.05 / 0.07	0.08 / 0.18	0.17 / 0.44	0.30 / 0.57	0.46 / 0.60
Q7B-ORZ/M	0.03 / 0.22	0.06 / 0.37	0.09 / 0.43	0.17 / 0.49	0.28 / 0.55

Table 25: Null calibration of the dispersion–worth correlation. At every ledger position, every action's m outcomes are redrawn as Bernoulli at the position's pooled success rate (so no action is worth more than any other), and the Spearman correlation of $\hat{\sigma}_b^{\mathrm{alt}}$ with the value term is recomputed with the same estimators, 300 times per run. "Dead": fraction of positions where every rollout failed; n: positions. Four column blocks: the clipped $\hat{\sigma}_b$ used for ranking (null mean and 95% band, excess = observed − null mean, one-sided p); the unclipped bias-corrected $\hat{\sigma}^2$, whose null correlation is negative on most runs because at live positions its noise is mostly below the dead-zone tie at zero; the rank-partial correlation controlling for $\hat{V}(1-\hat{V})$, whose null is not zero either because the residualisation is linear in ranks. "Live": observed correlation restricted to positions with some success and some failure.

				clipped $\hat{\sigma}_b^{\mathrm{alt}}$ (ranking statistic)				signed $\hat{\sigma}^2$			partial | $\hat{V}(1-\hat{V})$			live
cell	n	m	dead	obs.	null	excess	p	obs.	null	p	obs.	null	p	obs.
Q0.5B/G (window)	700	8	25%	+0.47	+0.38 [+0.32,+0.43]	+0.09	<0.003	+0.16	−0.16	<0.003	−0.06	−0.01	0.880	+0.21
Q0.5B/M (window)	500	8	31%	+0.47	+0.38 [+0.31,+0.44]	+0.09	<0.003	+0.12	−0.16	<0.003	+0.05	+0.02	0.317	+0.20
Q1.5B/G (window)	700	8	9%	+0.36	+0.31 [+0.26,+0.37]	+0.04	0.077	+0.09	−0.12	<0.003	+0.04	+0.03	0.417	+0.14
Q1.5B/M (window)	500	8	22%	+0.46	+0.36 [+0.29,+0.42]	+0.10	<0.003	+0.11	−0.15	<0.003	+0.05	+0.01	0.200	+0.25
QM7B/G (window)	600	8	4%	+0.25	+0.18 [+0.12,+0.26]	+0.07	0.033	+0.12	−0.04	<0.003	+0.07	+0.04	0.273	+0.19
QM7B/M (window)	400	8	8%	+0.22	+0.20 [+0.12,+0.28]	+0.02	0.320	+0.07	−0.06	0.003	−0.03	+0.06	0.970	+0.11
QM7B-Oat/G (window)	600	8	4%	+0.27	+0.18 [+0.11,+0.24]	+0.09	0.003	+0.15	−0.03	<0.003	+0.09	+0.05	0.180	+0.21
QM7B-SFT/G (window)	300	8	1%	+0.05	+0.11 [+0.02,+0.21]	−0.06	0.903	+0.00	−0.00	0.477	+0.01	+0.05	0.770	+0.03
L8B/G (window)	500	8	9%	+0.29	+0.27 [+0.19,+0.34]	+0.01	0.330	+0.04	−0.11	<0.003	−0.01	+0.01	0.737	+0.10
Q0.5B/G (whole trace)	1250	8	26%	+0.45	+0.39 [+0.35,+0.43]	+0.06	<0.003	+0.13	−0.16	<0.003	−0.07	+0.00	0.993	+0.14
Q0.5B/M (whole trace)	1000	8	38%	+0.50	+0.42 [+0.38,+0.46]	+0.08	<0.003	+0.10	−0.18	<0.003	−0.01	−0.00	0.663	+0.22
Q1.5B/G (whole trace)	1250	8	10%	+0.39	+0.34 [+0.29,+0.38]	+0.06	0.003	+0.09	−0.13	<0.003	−0.06	+0.02	1.000	+0.16
Q1.5B/M (whole trace)	1000	8	24%	+0.43	+0.38 [+0.34,+0.42]	+0.05	0.017	+0.07	−0.16	<0.003	−0.07	+0.02	0.987	+0.13
QM7B/G (whole trace)	1000	8	10%	+0.41	+0.27 [+0.22,+0.31]	+0.15	<0.003	+0.22	−0.09	<0.003	+0.11	+0.04	0.020	+0.27
QM7B/M (whole trace)	500	8	15%	+0.40	+0.29 [+0.22,+0.34]	+0.11	<0.003	+0.16	−0.10	<0.003	+0.07	+0.04	0.240	+0.25
QM7B-Oat/G (whole trace)	750	8	6%	+0.28	+0.20 [+0.13,+0.26]	+0.08	0.013	+0.18	−0.04	<0.003	+0.11	+0.06	0.080	+0.20
QM7B-SFT/G (whole trace)	500	8	2%	+0.26	+0.18 [+0.10,+0.25]	+0.09	0.017	+0.12	−0.05	<0.003	+0.07	+0.03	0.237	+0.20
L8B/G (whole trace)	1000	8	8%	+0.28	+0.25 [+0.20,+0.30]	+0.03	0.137	+0.05	−0.10	<0.003	−0.02	+0.03	0.913	+0.14
Q7B/G (whole trace)	500	8	5%	+0.32	+0.16 [+0.08,+0.23]	+0.15	<0.003	+0.16	−0.02	<0.003	+0.19	+0.07	<0.003	+0.26
Q7B/M (whole trace)	400	8	8%	+0.41	+0.26 [+0.17,+0.34]	+0.15	<0.003	+0.23	−0.09	<0.003	+0.10	+0.03	0.073	+0.31
Q7B-ORZ/G (whole trace)	500	8	6%	+0.56	+0.43 [+0.37,+0.50]	+0.13	<0.003	+0.17	−0.17	<0.003	−0.05	−0.05	0.500	+0.37
Q7B-ORZ/M (whole trace)	400	8	12%	+0.56	+0.38 [+0.31,+0.44]	+0.18	<0.003	+0.27	−0.15	<0.003	+0.06	−0.00	0.147	+0.37
Q0.5B/G (whole trace, m=32)	500	32	18%	+0.53	+0.32 [+0.24,+0.38]	+0.21	<0.003	+0.33	−0.16	<0.003	+0.07	+0.02	0.123	+0.39
Q1.5B/G (whole trace, m=32)	300	32	6%	+0.42	+0.25 [+0.15,+0.35]	+0.17	<0.003	+0.23	−0.11	<0.003	+0.10	+0.04	0.153	+0.33
QM7B/G (whole trace, m=32)	200	32	4%	+0.54	+0.20 [+0.07,+0.33]	+0.33	<0.003	+0.36	−0.05	<0.003	+0.32	+0.07	<0.003	+0.48
QM7B-Oat/G (whole trace, m=32)	200	32	6%	+0.40	+0.16 [+0.02,+0.28]	+0.24	<0.003	+0.25	−0.01	<0.003	+0.14	+0.08	0.187	+0.35
L8B/G (whole trace, m=32)	300	32	7%	+0.40	+0.20 [+0.11,+0.31]	+0.20	<0.003	+0.23	−0.11	<0.003	+0.20	+0.03	<0.003	+0.33
Q0.5B/G (window, m=32)	300	32	17%	+0.46	+0.31 [+0.23,+0.39]	+0.16	<0.003	+0.21	−0.14	<0.003	+0.11	+0.03	0.100	+0.31
Q0.5B/M (window, m=32)	250	32	29%	+0.58	+0.37 [+0.28,+0.47]	+0.22	<0.003	+0.31	−0.19	<0.003	+0.18	+0.04	0.020	+0.38
Q1.5B/G (window, m=32)	160	32	3%	+0.42	+0.24 [+0.10,+0.36]	+0.18	<0.003	+0.27	−0.11	<0.003	+0.05	+0.04	0.443	+0.33
Q1.5B/M (window, m=32)	250	32	19%	+0.50	+0.30 [+0.21,+0.40]	+0.19	<0.003	+0.24	−0.14	<0.003	+0.11	+0.04	0.147	+0.34
QM7B/G (window, m=32)	160	32	1%	+0.27	+0.10 [−0.05,+0.26]	+0.17	0.020	+0.23	+0.02	0.007	+0.11	+0.07	0.287	+0.27
QM7B/M (window, m=32)	160	32	9%	+0.28	+0.19 [+0.05,+0.32]	+0.08	0.113	+0.12	−0.04	0.017	+0.14	+0.08	0.203	+0.17
QM7B-SFT/G (window, m=32)	120	32	2%	+0.00	+0.09 [−0.08,+0.26]	−0.09	0.833	−0.02	+0.01	0.653	−0.05	+0.06	0.897	−0.02
L8B/G (window, m=32)	250	32	7%	+0.40	+0.19 [+0.08,+0.29]	+0.21	<0.003	+0.20	−0.10	<0.003	+0.18	+0.03	0.017	+0.34
Q0.5B/G (window, m=64)	200	64	14%	+0.50	+0.27 [+0.15,+0.38]	+0.23	<0.003	+0.30	−0.15	<0.003	+0.19	+0.04	0.010	+0.37
Q1.5B/G (window, m=64)	200	64	2%	+0.38	+0.20 [+0.07,+0.32]	+0.17	<0.003	+0.22	−0.09	<0.003	+0.10	+0.05	0.267	+0.31
L8B/G (window, m=64)	200	64	6%	+0.46	+0.17 [+0.04,+0.28]	+0.29	<0.003	+0.32	−0.10	<0.003	+0.26	+0.03	<0.003	+0.40
QM7B/G (window, 4-shot base)	300	8	4%	+0.23	+0.19 [+0.09,+0.28]	+0.04	0.207	+0.08	−0.04	0.017	+0.04	+0.05	0.603	+0.16

Table 26: Exchangeable-token null for the dispersion–worth correlation, whole-trace $m{=}8$ ledgers. At every position a pseudo-realised action is drawn from the base's own distribution over the measured actions, $\hat{\sigma}_b^{\mathrm{alt}}$ and the value term are recomputed with it in the realised token's role, and the Spearman coefficient over positions is taken; 200 draws per cell. If dispersion predicted something specific to the trained token, the observed coefficient would exceed this null's mean; it does not (excess $=$ observed $-$ null mean, one-sided p the share of draws at or above the observed value). But the null draws the trained token itself with its base probability, so on most positions the pseudo-realised token is the realised one and the test has little power; against the "alternatives only" null the observed coefficient is higher on all 8 cells of the two 7B bases and lower on the 5 others, a comparison confounded by the alternatives' lower base probabilities (the value term's structure changes with the token's probability). Neither null settles specificity. "Shared-rate null" repeats Table 25's null mean for comparison; "alternatives only" draws the pseudo-realised token from the alternatives alone, excluding the trained token (its excess is the observed coefficient minus that null's mean, with its own one-sided p). The left block uses the paper's estimator, average ranks for ties ("observed" equals Table 22's coefficient); the right block repeats observed, null and excess with ties broken in array order, and reaches the same verdict (excess -0.01 to +0.01 on the RLVR cells, -0.00 on the distilled one, $p \geq 0.10$).

		average ranks for ties (the paper's estimator)							ties broken in array order			
			shared-rate	exchangeable-token null			alternatives only				exchangeable null	
cell	n	observed	null mean	mean [95% band]	excess	p	null mean	excess (p)	observed	null	excess	p
Q0.5B/G	1250	+0.45	+0.39	+0.46 [+0.44,+0.47]	**−0.007**	0.82	+0.48	−0.034 (1.00)	+0.45	+0.44	**+0.006**	0.24
Q0.5B/M	1000	+0.50	+0.42	+0.49 [+0.48,+0.51]	**+0.006**	0.19	+0.50	−0.005 (0.71)	+0.49	+0.49	**+0.006**	0.22
Q1.5B/G	1250	+0.39	+0.34	+0.40 [+0.39,+0.41]	**−0.007**	0.85	+0.43	−0.042 (1.00)	+0.36	+0.36	**−0.008**	0.85
Q1.5B/M	1000	+0.43	+0.38	+0.43 [+0.42,+0.44]	**−0.003**	0.69	+0.48	−0.059 (1.00)	+0.46	+0.46	**+0.001**	0.45
QM7B/G	1000	+0.41	+0.27	+0.40 [+0.39,+0.42]	**+0.009**	0.07	+0.33	+0.089 (0.01)	+0.37	+0.37	**+0.009**	0.10
QM7B/M	500	+0.40	+0.29	+0.40 [+0.38,+0.42]	**−0.002**	0.62	+0.35	+0.050 (0.01)	+0.35	+0.35	**+0.001**	0.41
QM7B-Oat/G	750	+0.28	+0.20	+0.28 [+0.26,+0.29]	**+0.003**	0.34	+0.24	+0.042 (0.01)	+0.24	+0.24	**+0.001**	0.45
QM7B-SFT/G	500	+0.26	+0.18	+0.24 [+0.19,+0.27]	**+0.026**	0.09	+0.16	+0.104 (0.01)	+0.24	+0.24	**−0.003**	0.59
L8B/G	1000	+0.28	+0.25	+0.28 [+0.25,+0.30]	**+0.004**	0.37	+0.29	−0.007 (0.70)	+0.30	+0.30	**−0.001**	0.48
Q7B/G	500	+0.32	+0.16	+0.31 [+0.29,+0.34]	**+0.007**	0.27	+0.21	+0.102 (0.01)	+0.26	+0.26	**−0.000**	0.51
Q7B/M	400	+0.41	+0.26	+0.41 [+0.38,+0.43]	**+0.004**	0.40	+0.29	+0.122 (0.01)	+0.37	+0.37	**+0.003**	0.42
Q7B-ORZ/G	500	+0.56	+0.43	+0.56 [+0.54,+0.58]	**+0.004**	0.34	+0.46	+0.104 (0.01)	+0.50	+0.50	**−0.000**	0.48
Q7B-ORZ/M	400	+0.56	+0.38	+0.55 [+0.53,+0.57]	**+0.006**	0.24	+0.40	+0.159 (0.01)	+0.61	+0.61	**−0.001**	0.53

Table 27: Rollout-resolution series. From one high-resolution draw, the first m' outcomes of every action are kept, $m' \in \{8, 16, 32, 64\}$: the positions, the actions, the base probabilities and the underlying rollouts are identical across a row, so only the number of Bernoulli draws per action changes. Each block gives the observed Spearman of $\hat{\sigma}_b^{\mathrm{alt}}$ with the value term and its shared-rate null mean, the excess with its one-sided p, and the excess of the rank-partial correlation controlling for $\hat{V}(1-\hat{V})$. A correlation manufactured by estimator noise must lose its excess as m' grows; a real association attenuated by that noise must gain it. Nulls are re-drawn for this table, so at $m' = m$ they differ from Table 25's in the second decimal. The "window" distilled run covers the first 256 tokens, where fewer than a third of that cell's whole-trace positions lie (Table 28).

		m'=8			m'=16			m'=32			m'=64		
cell	n	obs./null	excess (p)	partial	obs./null	excess (p)	partial	obs./null	excess (p)	partial	obs./null	excess (p)	partial
Q0.5B/G (whole trace, m=32)	500	+0.44/+0.38	**+0.06**(0.02)	−0.08	+0.48/+0.35	**+0.14**(<0.005)	−0.05	+0.53/+0.32	**+0.21**(<0.005)	+0.04	–	–	–
Q0.5B/G (window, m=32)	300	+0.48/+0.37	**+0.11**(0.01)	+0.02	+0.49/+0.35	**+0.14**(<0.005)	+0.03	+0.46/+0.31	**+0.16**(<0.005)	+0.08	–	–	–
Q0.5B/G (window, m=64)	200	+0.38/+0.36	**+0.03**(0.27)	−0.07	+0.44/+0.32	**+0.12**(0.01)	+0.06	+0.40/+0.30	**+0.10**(0.05)	+0.04	+0.50/+0.28	**+0.22**(<0.005)	+0.15
Q0.5B/M (window, m=32)	250	+0.52/+0.41	**+0.11**(0.01)	−0.05	+0.53/+0.40	**+0.14**(<0.005)	+0.05	+0.58/+0.37	**+0.22**(<0.005)	+0.14	–	–	–
Q1.5B/G (whole trace, m=32)	300	+0.41/+0.34	**+0.06**(0.07)	−0.07	+0.38/+0.30	**+0.09**(0.03)	−0.05	+0.42/+0.25	**+0.17**(<0.005)	+0.06	–	–	–
Q1.5B/G (window, m=32)	160	+0.32/+0.30	**+0.03**(0.35)	+0.03	+0.32/+0.28	**+0.04**(0.26)	−0.08	+0.42/+0.24	**+0.18**(0.01)	+0.01	–	–	–
Q1.5B/G (window, m=64)	200	+0.22/+0.32	**−0.09**(0.95)	−0.06	+0.36/+0.28	**+0.08**(0.07)	−0.04	+0.34/+0.23	**+0.11**(0.03)	−0.02	+0.38/+0.20	**+0.18**(<0.005)	+0.05
Q1.5B/M (window, m=32)	250	+0.44/+0.35	**+0.09**(0.03)	−0.00	+0.47/+0.33	**+0.14**(<0.005)	−0.05	+0.50/+0.30	**+0.20**(<0.005)	+0.08	–	–	–
QM7B-Oat/G (whole trace, m=32)	200	+0.36/+0.20	**+0.16**(0.01)	+0.14	+0.40/+0.16	**+0.24**(<0.005)	+0.16	+0.40/+0.15	**+0.25**(<0.005)	+0.07	–	–	–
QM7B/G (whole trace, m=32)	200	+0.39/+0.24	**+0.15**(<0.005)	+0.05	+0.46/+0.22	**+0.24**(<0.005)	+0.15	+0.54/+0.19	**+0.35**(<0.005)	+0.26	–	–	–
QM7B/G (window, m=32)	160	+0.31/+0.15	**+0.16**(0.01)	+0.13	+0.18/+0.12	**+0.06**(0.21)	−0.04	+0.27/+0.11	**+0.17**(0.01)	+0.04	–	–	–
QM7B/M (window, m=32)	160	+0.21/+0.23	**−0.02**(0.63)	+0.02	+0.19/+0.21	**−0.02**(0.58)	−0.08	+0.28/+0.19	**+0.09**(0.13)	+0.06	–	–	–
QM7B-SFT/G (window, m=32)	120	−0.09/+0.13	**−0.22**(0.99)	−0.22	+0.08/+0.09	**−0.01**(0.59)	+0.00	+0.00/+0.09	**−0.09**(0.84)	−0.11	–	–	–
L8B/G (whole trace, m=32)	300	+0.39/+0.29	**+0.10**(0.02)	+0.03	+0.40/+0.25	**+0.15**(<0.005)	+0.06	+0.40/+0.20	**+0.20**(<0.005)	+0.17	–	–	–
L8B/G (window, m=32)	250	+0.24/+0.28	**−0.03**(0.75)	−0.02	+0.28/+0.23	**+0.05**(0.15)	+0.04	+0.40/+0.18	**+0.22**(<0.005)	+0.15	–	–	–
L8B/G (window, m=64)	200	+0.33/+0.30	**+0.03**(0.23)	+0.03	+0.33/+0.24	**+0.09**(0.07)	+0.02	+0.41/+0.20	**+0.21**(<0.005)	+0.15	+0.46/+0.17	**+0.28**(<0.005)	+0.23

Table 28: Where the ceiling lives along a trace, from the whole-trace ledger. Positions are split at the 256th token, the window a leading-window estimator covers. The rate falls late on 12 of the 12 cells, the base less surprised by the trained token as a trace approaches its answer. On 10 of them the dispersion of the alternatives' values falls with it; on Q7B/G, Q7B-ORZ/G (the general-purpose base on GSM8K, not on its MATH cells) the late dispersion is *higher* and the plug-in ceiling rises with it; on QM7B-Oat/G, Q7B/M the rate falls, the dispersion does not rise and the plug-in ceiling rises anyway, a movement of a maximum over noisy values that the dispersion does not support. Ratios are point estimates of selection-inflated maxima. The last two columns are Λ computed from all positions and, for comparison, from the early positions extrapolated over the whole trace: the early-window extrapolation, which understates Λ on 9 cells and overstates it on Q7B/G, Q7B-ORZ/G (the ceiling's direction predicts the sign except on QM7B-Oat/G, where the two differ by 2 points). Its size also depends on trace length: the extrapolation multiplies each trace's early mean ceiling by the trace's length, so a few long traces can dominate it, which is why Q7B-ORZ/M, the RLVR cell with the longest traces, differs most although its pooled ceiling barely falls. "–": the extrapolated early ceiling total is not positive (Q7B/M). Cells whose traces rarely reach 256 tokens (fewer than 10% of positions late) are omitted.

	positions		$\bar{\varepsilon}_t$		$\bar{\tilde{\sigma}}_b$		$\overline{F^{\det}}$			Λ naive	
cell	<256	≥ 256	early	late	early	late	early	late	late/early	whole trace	early-only
Q0.5B/G	1113	137	0.22	0.11	0.0114	0.0041	0.0055	0.0011	0.21	20%	15%
Q0.5B/M	675	325	0.19	0.10	0.0088	0.0041	0.0053	0.0010	0.18	19%	12%
Q1.5B/G	1089	161	0.20	0.11	0.0095	0.0052	0.0065	0.0015	0.23	18%	13%
Q1.5B/M	608	392	0.19	0.08	0.0095	0.0035	0.0051	0.0009	0.18	27%	18%
QM7B/G	669	331	0.16	0.09	0.0091	0.0048	0.0054	0.0027	0.51	16%	9.7%
QM7B/M	362	138	0.13	0.10	0.0091	0.0031	0.0057	0.0019	0.34	20%	8.3%
QM7B-Oat/G	555	195	0.10	0.05	0.0071	0.0061	0.0024	0.0034	1.43	47%	45%
QM7B-SFT/G	121	379	0.92	0.75	0.0256	0.0155	0.0322	0.0223	0.69	1.1%	0.5%
Q7B/G	421	79	0.14	0.09	0.0107	0.0137	0.0055	0.0058	1.07	18%	21%
Q7B/M	253	147	0.08	0.06	0.0072	0.0054	0.0019	0.0033	1.79	20%	–
Q7B-ORZ/G	316	184	0.25	0.19	0.0043	0.0094	0.0044	0.0053	1.21	8.0%	9.8%
Q7B-ORZ/M	170	230	0.18	0.14	0.0050	0.0025	0.0015	0.0014	0.94	24%	7.5%

Table 29: Is worth actionable on top of room? Positions are ranked by each criterion computed on half the rollouts (4 of 8) and scored by the inverse-inclusion-weighted mean signed A_t measured on the other half, pooled over the 12 RLVR whole-trace m=8 runs (9550 positions) with one cluster bootstrap over problems (criterion ranks are normalised within cell first, so cells with different advantage scales pool fairly); 95% intervals. A random pick scores the population mean in expectation (first column); every other entry is a paired difference on the same bootstrap resamples: a criterion's selected mean minus the population mean, and two contrasts between criteria. Rate is the exact room factor; the product is the room times dispersion (Corollary 1's room, with dispersion standing in for the worth it cannot see); the oracle ranks by $|\hat{A}_t|$ from the ranking half, an unsigned oracle, so it bounds what a noisy magnitude selector can do, not a signed one. Only the oracle beats a random pick at every budget; the product reaches the edge of its interval at 10% and 30%; no criterion separates from rate (the dispersion–rate and entropy–rate contrasts, computed on the same resamples but not printed, cover zero at every budget, as product–rate does), and the product does not separate from the oracle. The test runs at m=4 per half.

	random pick	gain over a random pick					paired contrasts	
budget	(population mean)	ε_t	H_b	$\hat{\sigma}_b^{\mathrm{alt}}$	$(1-e^{-\varepsilon_t})\hat{\sigma}_b^{\mathrm{alt}}$	$\lvert\hat{A}_t\rvert$ (half A)	product − rate	oracle − product
top 10%	+0.0009 [+0.0002,+0.0017]	+0.0012 [−0.0092,+0.0110]	+0.0071 [−0.0006,+0.0150]	+0.0012 [−0.0007,+0.0037]	+0.0070 [−0.0000,+0.0135]	+0.0107 [+0.0011,+0.0212]	+0.0058 [−0.0041,+0.0155]	+0.0038 [−0.0049,+0.0135]
top 20%	+0.0009 [+0.0002,+0.0017]	+0.0022 [−0.0031,+0.0074]	+0.0032 [−0.0016,+0.0078]	+0.0003 [−0.0014,+0.0019]	+0.0018 [−0.0002,+0.0041]	+0.0058 [+0.0008,+0.0109]	−0.0004 [−0.0053,+0.0047]	+0.0040 [−0.0007,+0.0087]
top 30%	+0.0009 [+0.0002,+0.0017]	+0.0019 [−0.0015,+0.0053]	+0.0022 [−0.0013,+0.0054]	+0.0009 [−0.0003,+0.0021]	+0.0013 [−0.0000,+0.0026]	+0.0038 [+0.0009,+0.0069]	−0.0006 [−0.0039,+0.0029]	+0.0025 [−0.0003,+0.0054]

Table 30: Decoding and prompt grid, greedy decoding except where a row says $T{=}0.8$, for the base and trained policies of the cells with such a run (model codes: base05/rl05, base15/rl15 the 0.5B and 1.5B Qwen bases and their SimpleRL-Zoo descendants; base7, rl7, oat7, sft7 Qwen2.5-Math-7B and its three descendants; the sft7 rows use 200 problems where the distilled cell elsewhere uses 100). The paper's arms are the sampled zero-shot template and the sampled 4-shot prompt in the pair's template (temperature 0.8, Table 6); here the same prompts are decoded greedily, with a 2048-token cap (the 7B and 8B cells' own, twice the 0.5B and 1.5B cells'; Table 34), and the other rows vary one factor at a time. The published few-shot accuracy for the Qwen2.5-Math-7B base is a greedy number under an evaluation harness prompt. The elicited shares in Section 4.2 are read two ways: the best greedy 4-shot arm against the *sampled* zero-shot base and policy (mixed decoding), and greedy against greedy, the same arm against the trained policy's own greedy zero-shot run (0.895 for SimpleRL-Zero, 0.905 for Oat-Zero, 0.880 for the distilled sibling), which gives 89–96% on the three 7B descendants and 9–33% on the 4 other cells with a greedy policy run. The chat-template and Question/Answer 4-shot arms both score $173/200$ on this cell; they differ on 14 problems and coincide in the total. Sampled at $T{=}0.8$, the Question/Answer and Problem/Solution 4-shot forms reach 0.798 and 0.802 on the primary cell, 96–97% of its sampled gap, against the 56% of the paper's chat-form arm, which truncates 11% of its samples.

model	data	prompt	shots	decoding	n	accuracy	truncated	mean length
base05	GSM8K	zero-shot (the paper's policy)	0	greedy	250	0.352 [0.296,0.412]	6%	186
rl05	GSM8K	zero-shot (the paper's policy)	0	greedy	250	0.492 [0.428,0.556]	1%	254
base05	GSM8K	4-shot, Question/Answer	4	greedy	250	0.364 [0.304,0.424]	1%	125
base05	MATH	zero-shot (the paper's policy)	0	greedy	200	0.145 [0.100,0.200]	22%	387
rl05	MATH	zero-shot (the paper's policy)	0	greedy	200	0.355 [0.290,0.420]	14%	491
base05	MATH	4-shot, Question/Answer	4	greedy	200	0.215 [0.160,0.275]	12%	273
base15	GSM8K	zero-shot (the paper's policy)	0	greedy	250	0.568 [0.508,0.628]	6%	180
rl15	GSM8K	zero-shot (the paper's policy)	0	greedy	250	0.760 [0.708,0.812]	1%	306
base15	GSM8K	4-shot, Question/Answer	4	greedy	250	0.612 [0.552,0.668]	0%	103
base15	MATH	zero-shot (the paper's policy)	0	greedy	200	0.275 [0.220,0.340]	20%	347
rl15	MATH	zero-shot (the paper's policy)	0	greedy	200	0.580 [0.515,0.650]	12%	553
base15	MATH	4-shot, Question/Answer	4	greedy	200	0.325 [0.265,0.390]	13%	255
base7	GSM8K	zero-shot (the paper's policy)	0	greedy	200	0.545 [0.475,0.610]	20%	644
base7	GSM8K	4-shot, chat template (the paper's arm)	4	greedy	200	0.865 [0.820,0.910]	0%	82
base7	GSM8K	4-shot, Problem/Solution	4	greedy	200	0.840 [0.790,0.885]	1%	131
base7	GSM8K	4-shot, Problem/Solution	4	$T{=}0.8$, 16	200	0.802 [0.766,0.835]	2%	229
base7	GSM8K	4-shot, Question/Answer	4	greedy	200	0.865 [0.815,0.910]	0%	85
base7	GSM8K	4-shot, Question/Answer	4	$T{=}0.8$, 16	200	0.798 [0.761,0.835]	1%	149
oat7	GSM8K	zero-shot (the paper's policy)	0	greedy	200	0.905 [0.865,0.940]	0%	391
rl7	GSM8K	zero-shot (the paper's policy)	0	greedy	200	0.895 [0.850,0.935]	4%	371
rl7	MATH	zero-shot (the paper's policy)	0	greedy	150	0.773 [0.707,0.833]	5%	516
base7	MATH	4-shot, Problem/Solution	4	greedy	150	0.593 [0.520,0.667]	2%	179
sft7	GSM8K	zero-shot (the paper's policy)	0	greedy	200	0.880 [0.835,0.925]	18%	1278

Table 31: The generic prefix control at $k = 4$, 8 and 16. "Arm-0 coverage" is the fraction of distinct problems for which the policy actually produced the top-ranked shared opening; the prefix is grafted into *every* problem regardless, so it carries no information about the problem at hand even where coverage is low, but low coverage means it is less typical of what the policy would have written. Real lift is $R(k) - R(0)$ on the same traces; content is the difference; 95% cluster-bootstrap intervals. The $k = 8$ real lift comes from a separate transplant run on the same traces with fresh completions, not from the draw behind $k = 4$ and 16; on the distilled cell it lies below both neighbouring points, the $k = 4$ value above its interval's upper edge; the draws are independent, so this is run-to-run variation larger than the problem-level intervals capture on that cell, and we read no dip into it. Coverage is that of the top-ranked opening (arm 0); the generic arm pools the top three, as in Table 14.

cell	k	arm-0 coverage	real lift	generic (format)	content
Q0.5B/G	4	24%	+0.044 [+0.024,+0.065]	+0.030 [+0.012,+0.048]	+0.014 [−0.005,+0.033]
Q0.5B/G	8	2%	+0.059 [+0.038,+0.079]	−0.132 [−0.157,−0.106]	+0.191 [+0.161,+0.221]
Q0.5B/M	4	10%	+0.035 [+0.017,+0.054]	+0.033 [+0.020,+0.048]	+0.001 [−0.014,+0.018]
Q1.5B/G	4	38%	+0.035 [+0.013,+0.055]	+0.008 [−0.012,+0.027]	+0.027 [+0.008,+0.044]
Q1.5B/G	8	30%	+0.048 [+0.024,+0.069]	+0.005 [−0.015,+0.024]	+0.043 [+0.025,+0.063]
Q1.5B/G	16	2%	+0.066 [+0.044,+0.089]	−0.030 [−0.053,−0.008]	+0.096 [+0.072,+0.118]
Q1.5B/M	4	14%	+0.080 [+0.055,+0.107]	+0.050 [+0.033,+0.065]	+0.030 [+0.010,+0.051]
QM7B/G	4	58%	+0.264 [+0.227,+0.300]	+0.264 [+0.232,+0.296]	+0.000 [−0.023,+0.022]
QM7B/G	8	39%	+0.260 [+0.227,+0.295]	+0.266 [+0.230,+0.302]	−0.006 [−0.033,+0.021]
QM7B/G	16	7%	+0.264 [+0.231,+0.299]	+0.245 [+0.206,+0.282]	+0.019 [−0.012,+0.048]
QM7B/M	4	26%	+0.152 [+0.114,+0.191]	+0.128 [+0.087,+0.171]	+0.023 [−0.008,+0.054]
QM7B-Oat/G	4	44%	+0.278 [+0.240,+0.316]	+0.229 [+0.195,+0.265]	+0.049 [+0.025,+0.073]
QM7B-Oat/G	8	34%	+0.297 [+0.259,+0.335]	+0.252 [+0.218,+0.287]	+0.045 [+0.022,+0.070]
QM7B-Oat/G	16	6%	+0.300 [+0.263,+0.336]	+0.255 [+0.220,+0.290]	+0.045 [+0.024,+0.067]
QM7B-SFT/G	4	79%	+0.152 [+0.105,+0.202]	+0.070 [+0.035,+0.107]	+0.082 [+0.045,+0.116]
QM7B-SFT/G	8	67%	+0.072 [+0.020,+0.124]	−0.027 [−0.070,+0.017]	+0.099 [+0.056,+0.143]
QM7B-SFT/G	16	16%	+0.198 [+0.150,+0.245]	+0.080 [+0.038,+0.121]	+0.118 [+0.082,+0.155]
L8B/G	4	1%	+0.047 [+0.025,+0.070]	−0.197 [−0.221,−0.172]	+0.244 [+0.212,+0.275]
L8B/G	16	0%	+0.109 [+0.083,+0.135]	−0.253 [−0.282,−0.223]	+0.362 [+0.324,+0.401]
Q7B/G	4	58%	+0.029 [−0.000,+0.060]	−0.002 [−0.027,+0.023]	+0.031 [+0.006,+0.056]
Q7B/M	4	28%	+0.078 [+0.046,+0.114]	+0.035 [+0.011,+0.059]	+0.043 [+0.013,+0.072]
Q7B-ORZ/G	4	35%	+0.010 [−0.013,+0.032]	−0.031 [−0.051,−0.012]	+0.041 [+0.019,+0.062]
Q7B-ORZ/M	4	42%	+0.075 [+0.042,+0.110]	−0.010 [−0.037,+0.015]	+0.085 [+0.054,+0.118]

Table 32: Out-of-sample validation of the worth criterion. Positions are ranked by a statistic computed on one set of rollouts and the value term is measured on an independent set drawn at the same positions with fresh seeds, so no part of the ranking enters the outcome. The right-hand columns calibrate the dispersion criterion against the shared-rate null of Table 25, with *both* sets of outcomes redrawn at each position's pooled rate. Position counts can be lower than in the source runs because only positions with a complete fresh-seed draw for every action enter. † marks the two runs (Q1.5B/G and QM7B/G at m=32 over the window) whose high-resolution rollouts were drawn under the same seeds as the first set and may share draws with it; they are excluded from the pooled numbers in the text. Intervals are 95% cluster bootstraps over problems.

		Spearman of the in-sample statistic with the out-of-sample value term				$\hat{\sigma}_b$ under the null			signed $\hat{\sigma}^2$		partial	
cell	positions	$\hat{\sigma}_b$	ε_t	H_b	$\lvert A_t \rvert$	null	excess	p	obs./null	p	obs./null	p
Q0.5B/G (whole trace, m=32)	365	+0.37 [+0.27,+0.47]	−0.06 [−0.17,+0.05]	−0.05 [−0.16,+0.05]	+0.48 [+0.37,+0.57]	+0.28 [+0.20,+0.35]	+0.10	0.003	+0.15/−0.16	<0.003	−0.07/−0.04	0.650
Q0.5B/G (m=32)	261	+0.40 [+0.28,+0.50]	−0.06 [−0.18,+0.05]	−0.05 [−0.16,+0.06]	+0.52 [+0.38,+0.63]	+0.28 [+0.19,+0.37]	+0.12	0.007	+0.10/−0.16	<0.003	−0.04/−0.04	0.507
Q0.5B/G (m=64)	172	+0.36 [+0.20,+0.50]	−0.11 [−0.25,+0.05]	−0.10 [−0.25,+0.05]	+0.47 [+0.25,+0.63]	+0.26 [+0.14,+0.36]	+0.10	0.037	+0.04/−0.14	0.013	+0.03/−0.04	0.193
Q0.5B/M (m=32)	177	+0.52 [+0.41,+0.60]	−0.01 [−0.18,+0.16]	−0.03 [−0.20,+0.14]	+0.64 [+0.47,+0.76]	+0.31 [+0.20,+0.42]	+0.20	<0.003	+0.18/−0.20	<0.003	+0.17/−0.03	0.010
Q1.5B/G (whole trace, m=32)	262	+0.31 [+0.19,+0.42]	−0.08 [−0.20,+0.04]	−0.07 [−0.19,+0.06]	+0.34 [+0.18,+0.48]	+0.20 [+0.10,+0.29]	+0.11	0.017	+0.10/−0.13	<0.003	−0.03/−0.03	0.480
Q1.5B/G (m=32)†	151	+0.34 [+0.18,+0.48]	−0.04 [−0.18,+0.11]	−0.05 [−0.19,+0.09]	+0.47 [+0.28,+0.62]	+0.21 [+0.07,+0.33]	+0.13	0.007	+0.11/−0.13	<0.003	−0.00/−0.02	0.437
Q1.5B/G (m=64)	174	+0.32 [+0.16,+0.46]	−0.09 [−0.24,+0.07]	−0.10 [−0.24,+0.06]	+0.39 [+0.19,+0.54]	+0.19 [+0.06,+0.32]	+0.14	0.023	+0.11/−0.12	<0.003	+0.08/−0.02	0.063
Q1.5B/M (m=32)	246	+0.44 [+0.31,+0.55]	−0.09 [−0.22,+0.04]	−0.09 [−0.21,+0.03]	+0.43 [+0.23,+0.57]	+0.27 [+0.16,+0.36]	+0.17	<0.003	+0.16/−0.15	<0.003	+0.01/−0.04	0.213
QM7B-Oat/G (whole trace, m=32)	182	+0.18 [+0.04,+0.32]	−0.18 [−0.32,−0.03]	−0.18 [−0.31,−0.04]	+0.08 [−0.15,+0.28]	+0.07 [−0.05,+0.20]	+0.11	0.053	+0.09/−0.05	0.027	−0.04/−0.00	0.710
QM7B/G (whole trace, m=32)	183	+0.44 [+0.26,+0.58]	−0.02 [−0.16,+0.12]	−0.01 [−0.16,+0.13]	+0.41 [+0.25,+0.54]	+0.15 [+0.03,+0.28]	+0.28	<0.003	+0.21/−0.10	<0.003	+0.21/−0.01	<0.003
QM7B/G (m=32)†	160	+0.27 [+0.11,+0.42]	−0.25 [−0.39,−0.10]	−0.26 [−0.41,−0.10]	−0.05 [−0.22,+0.12]	+0.05 [−0.10,+0.20]	+0.22	<0.003	+0.22/−0.05	<0.003	+0.06/−0.01	0.160
QM7B/M (m=32)	160	+0.16 [−0.00,+0.32]	−0.04 [−0.19,+0.10]	−0.04 [−0.19,+0.11]	+0.26 [+0.02,+0.49]	+0.13 [−0.00,+0.26]	+0.04	0.280	+0.04/−0.09	0.020	−0.03/−0.00	0.643
QM7B-SFT/G (m=32)	108	+0.21 [+0.02,+0.39]	+0.03 [−0.15,+0.20]	+0.02 [−0.16,+0.20]	+0.06 [−0.17,+0.29]	+0.04 [−0.14,+0.22]	+0.17	0.030	+0.13/−0.05	0.030	+0.16/−0.00	0.037
L8B/G (whole trace, m=32)	227	+0.26 [+0.11,+0.40]	−0.04 [−0.19,+0.12]	−0.03 [−0.18,+0.12]	+0.29 [+0.13,+0.43]	+0.19 [+0.08,+0.31]	+0.07	0.117	+0.05/−0.10	0.013	+0.01/−0.01	0.343
L8B/G (m=32)	233	+0.20 [+0.07,+0.33]	+0.07 [−0.08,+0.22]	+0.06 [−0.08,+0.20]	+0.42 [+0.31,+0.53]	+0.17 [+0.06,+0.28]	+0.02	0.360	+0.01/−0.11	0.027	−0.04/−0.03	0.603
L8B/G (m=64)	183	+0.27 [+0.15,+0.38]	+0.04 [−0.13,+0.21]	+0.00 [−0.15,+0.16]	+0.37 [+0.20,+0.51]	+0.17 [+0.05,+0.30]	+0.09	0.063	+0.07/−0.10	0.010	+0.06/−0.02	0.143

Table 33: Split-half reliabilities (Spearman–Brown corrected to the full m) of the full dispersion estimate, of $|A_t|$ and of the value term $|Q_b - V_b^{\text{alt}}|$, the raw Spearman correlations with $|A_t|$ of Table 22 (alternatives-only dispersion), and the value term's correlation with rate, for every ledger run. The rate and entropy columns of $|A_t|$ are dominated by the room factor (Corollary 1); the value-term correlations are the informative ones.

cell	m	n	rel($\hat{\sigma}_b$)	rel($\lvert A\rvert$)	rel(value)	$r_s(\lvert A\rvert, \hat{\sigma}_b^{\text{alt}})$	$r_s(\lvert A\rvert, \varepsilon)$	$r_s(\lvert A\rvert, H_b)$	$r_s(\lvert Q_b - V_b^{\text{alt}}\rvert, \varepsilon)$
Q0.5B/G (window)	8	700	0.58	0.74	0.75	+0.33	0.43	0.43	−0.04
Q0.5B/M (window)	8	500	0.63	0.86	0.81	+0.42	0.44	0.42	+0.04
Q1.5B/G (window)	8	700	0.56	0.79	0.74	+0.25	0.55	0.54	−0.07
Q1.5B/M (window)	8	500	0.50	0.77	0.68	+0.31	0.45	0.46	−0.04
QM7B/G (window)	8	600	0.47	0.75	0.51	−0.02	0.75	0.74	−0.10
QM7B/M (window)	8	400	0.48	0.81	0.60	+0.05	0.73	0.74	+0.00
QM7B-Oat/G (window)	8	600	0.52	0.69	0.36	−0.06	0.80	0.79	−0.07
QM7B-SFT/G (window)	8	300	0.26	0.61	0.11	−0.10	0.69	0.65	−0.08
L8B/G (window)	8	500	0.45	0.78	0.71	+0.24	0.61	0.60	+0.06
Q0.5B/G (whole trace)	8	1250	0.63	0.81	0.79	+0.37	0.38	0.38	−0.07
Q0.5B/M (whole trace)	8	1000	0.59	0.80	0.81	+0.42	0.35	0.34	−0.01
Q1.5B/G (whole trace)	8	1250	0.51	0.79	0.76	+0.25	0.49	0.48	−0.11
Q1.5B/M (whole trace)	8	1000	0.61	0.82	0.81	+0.32	0.37	0.38	−0.08
QM7B/G (whole trace)	8	1000	0.56	0.79	0.68	+0.15	0.62	0.62	−0.12
QM7B/M (whole trace)	8	500	0.68	0.78	0.65	+0.08	0.61	0.62	−0.08
QM7B-Oat/G (whole trace)	8	750	0.56	0.73	0.45	−0.09	0.74	0.74	−0.15
QM7B-SFT/G (whole trace)	8	500	0.49	0.67	0.40	−0.00	0.66	0.59	−0.10
L8B/G (whole trace)	8	1000	0.44	0.73	0.58	+0.21	0.63	0.60	+0.01
Q7B/G (whole trace)	8	500	0.48	0.84	0.52	−0.07	0.76	0.76	−0.13
Q7B/M (whole trace)	8	400	0.56	0.69	0.60	+0.14	0.61	0.62	−0.10
Q7B-ORZ/G (whole trace)	8	500	0.77	0.72	0.79	+0.36	0.18	0.21	−0.19
Q7B-ORZ/M (whole trace)	8	400	0.59	0.72	0.75	+0.30	0.27	0.27	−0.24
Q0.5B/G (whole trace, m=32)	32	500	0.72	0.88	0.86	+0.35	0.58	0.58	−0.06
Q1.5B/G (whole trace, m=32)	32	300	0.66	0.87	0.80	+0.12	0.69	0.68	−0.06
QM7B/G (whole trace, m=32)	32	200	0.75	0.90	0.79	+0.02	0.85	0.84	−0.03
QM7B-Oat/G (whole trace, m=32)	32	200	0.59	0.93	0.52	−0.05	0.79	0.80	−0.17
L8B/G (whole trace, m=32)	32	300	0.54	0.86	0.68	+0.16	0.69	0.67	−0.06
Q0.5B/G (window, m=32)	32	300	0.67	0.89	0.82	+0.25	0.60	0.60	−0.06
Q0.5B/M (window, m=32)	32	250	0.62	0.90	0.90	+0.40	0.51	0.50	−0.03
Q1.5B/G (window, m=32)	32	160	0.57	0.84	0.72	+0.17	0.70	0.70	−0.07
Q1.5B/M (window, m=32)	32	250	0.66	0.90	0.82	+0.22	0.60	0.61	−0.10
QM7B/G (window, m=32)	32	160	0.62	0.90	0.56	−0.10	0.93	0.92	−0.25
QM7B/M (window, m=32)	32	160	0.55	0.92	0.63	−0.01	0.77	0.77	−0.04
QM7B-SFT/G (window, m=32)	32	120	0.36	0.77	0.19	−0.03	0.77	0.69	+0.03
L8B/G (window, m=32)	32	250	0.62	0.89	0.68	+0.16	0.72	0.69	+0.05
Q0.5B/G (window, m=64)	64	200	0.67	0.94	0.85	+0.28	0.64	0.64	−0.08
Q1.5B/G (window, m=64)	64	200	0.69	0.89	0.76	+0.11	0.73	0.72	−0.08
L8B/G (window, m=64)	64	200	0.70	0.91	0.74	+0.29	0.72	0.67	+0.05
QM7B/G (window, 4-shot base)	8	300	0.50	0.70	0.50	−0.08	0.78	0.78	−0.13

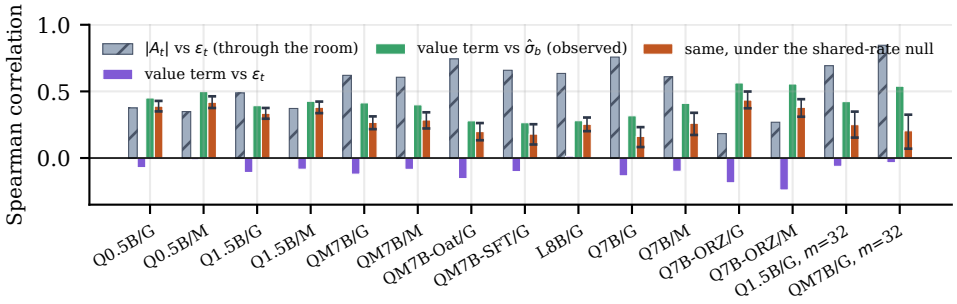


Figure 4: Predictors of worth, on the whole-trace runs of Table 22 (the last two groups are two of the five cells re-drawn at m=32 over whole traces). The hatched bar is the correlation of $|A_t|$ with rate, which runs through the room factor; the other bars are correlations with the value term. The last bar in each group is the dispersion correlation under the shared-rate null (mean and 95% band), which the observed bar must clear—and does on 12 of the 13 cells, most narrowly on Q1.5B/M (0.426 against a band top of 0.423); it does not on L8B/G (0.280 against a band top of 0.304). Clearing it says the association is not estimator noise; whether it is specific to the trained token is Table 26's question. The divergence statistics need both policies' distributions at the same positions and so are measured on the window runs; they are in Table 23 rather than here.

		GSM8K		MATH-500	
policy	cap	truncated	mean len	truncated	mean len
Qwen2.5-0.5B	1024	3.5%	185	10.2%	335
→RLVR	1024	0.9%	250	10.4%	477
Qwen2.5-1.5B	1024	3.2%	181	11.1%	316
→RLVR	1024	0.4%	297	13.8%	548
Qwen2.5-Math-7B	2048	21.9%	787	19.5%	920
→RLVR	2048	22.2%	722	9.0%	596
→RLVR (Oat)	2048	0.1%	393	–	–
→SFT-distill	2048	24.0%	1442	–	–
Llama-3.1-8B	2048	2.1%	187	–	–
→RLVR	2048	0.4%	126	–	–
Qwen2.5-7B	2048	11.5%	581	8.3%	614
→RLVR	2048	0.4%	321	1.5%	615
Qwen2.5-7B (ORZ format)	2048	0.2%	208	1.1%	368
→RLVR (ORZ)	2048	0.0%	445	13.2%	991

Table 34: Truncation rates and mean sampled lengths for every policy at the cap its cell uses, over the same 16-sample draws that produce the accuracies of Table 6. "Truncated" is the fraction of samples that reached the cap without emitting a stop token.

	truncated along the curve			recovered by k=4	share of
cell	$R(0)$	$R(4)$	$R(128)$	(points)	$R(4) - R(0)$
Q0.5B/G	3.5%	1.9%	2.1%	+1.5	35%
Q0.5B/M	10.2%	11.4%	12.2%	−1.2	0%
Q1.5B/G	3.2%	1.7%	1.1%	+1.4	41%
Q1.5B/M	11.1%	9.2%	12.3%	+1.9	23%
QM7B/G	21.8%	9.9%	10.1%	+11.9	45%
QM7B/M	19.5%	10.9%	11.4%	+8.6	57%
QM7B-Oat/G	21.8%	10.0%	12.1%	+11.8	43%
QM7B-SFT/G	21.8%	13.5%	8.7%	+8.3	55%
L8B/G	2.1%	1.9%	2.5%	+0.2	4%
Q7B/G	11.5%	10.1%	10.2%	+1.5	51%
Q7B/M	8.3%	7.0%	5.5%	+1.3	17%
Q7B-ORZ/G	0.2%	0.3%	0.2%	−0.0	0%
Q7B-ORZ/M	1.1%	1.8%	2.1%	−0.7	0%

Table 35: Truncation along the transplant curve: the share of the base's completions that hit the token cap after a handoff at k, the drop in that share by k=4, and the largest share of the four-token lift (the real lift of Table 14) it could explain if every recovered completion were graded correct. The $R(0)$ share is a mean over problems of the share of samples the generator stopped at the cap, so it can differ in the last digit from Table 34's, which counts samples of cap length.

Table 36: Temperature of the weights. Every whole-trace $m{=}8$ ledger statistic in the paper uses the base's temperature-1 probabilities as weights, while rollouts sample at temperature 0.8; here the weights are tempered to 0.8 and renormalised over the measured actions, with the rate re-read from the tempered distribution, and the same estimators re-run ($T{=}1$ columns repeat the paper's values). Λ and the wasted-rate share from the plug-in ceiling; search margin cross-fitted; rate above 1 nat is the share of total rate spent at $\varepsilon_t > 1$; Γ the allocation gap; the next pair is the raw Spearman of dispersion with the value term; the last column is the dense conservation ledger's $\sum_{t\le 4} \hat{A}_t$ (Table 8) with untempered and tempered weights, † marking a cell whose conservation verdict changes.

	Λ		wasted rate		search margin (cross-fit)		rate above 1 nat		Γ		$\rho(\hat{\sigma}_b, \text{value})$		$\sum_{t\le 4} \hat{A}_t$
cell	$T{=}1$	$T{=}0.8$	$T{=}1$	$T{=}0.8$	$T{=}1$	$T{=}0.8$	$T{=}1$	$T{=}0.8$	$T{=}1$	$T{=}0.8$	$T{=}1$	$T{=}0.8$	$T{=}1 \to 0.8$
Q0.5B/G	20%	21%	71%	71%	−0.0012	−0.0014	59%	66%	2.5	2.6	+0.45	+0.45	+0.022→+0.019
Q0.5B/M	19%	20%	79%	79%	−0.0004	−0.0004	57%	64%	3.1	3.3	+0.50	+0.50	+0.058→+0.055
Q1.5B/G	18%	18%	66%	65%	+0.0001	+0.0000	61%	66%	2.5	2.7	+0.39	+0.39	+0.050→+0.049
Q1.5B/M	27%	27%	74%	75%	−0.0015	−0.0014	55%	62%	2.8	3.0	+0.43	+0.42	+0.052→+0.050
QM7B/G	16%	16%	72%	73%	−0.0000	+0.0002	63%	70%	2.8	3.2	+0.41	+0.39	+0.226→+0.219
QM7B/M	20%	21%	71%	71%	+0.0021	+0.0017	54%	62%	3.2	3.5	+0.40	+0.39	+0.133→+0.127
QM7B-Oat/G	47%	52%	74%	76%	+0.0017	+0.0016	36%	43%	2.9	3.1	+0.28	+0.28	+0.264→+0.259
QM7B-SFT/G	1%	1%	62%	65%	−0.0062	−0.0061	84%	87%	2.2	2.4	+0.26	+0.26	+0.094→+0.087
L8B/G	37%	39%	62%	61%	−0.0007	−0.0005	64%	66%	2.1	2.2	+0.28	+0.27	+0.053→+0.048
Q7B/G	18%	19%	71%	75%	+0.0014	+0.0010	54%	62%	2.6	2.9	+0.32	+0.30	+0.028→+0.025
Q7B/M	20%	21%	66%	66%	+0.0016	+0.0015	27%	33%	2.3	2.4	+0.41	+0.43	+0.049→+0.042
Q7B-ORZ/G	8%	8%	84%	84%	+0.0007	+0.0008	76%	83%	3.5	3.6	+0.56	+0.56	−0.009→−0.012
Q7B-ORZ/M	24%	24%	89%	89%	+0.0001	+0.0001	63%	73%	4.4	5.0	+0.56	+0.55	+0.079→+0.077

Table 37: Action-set coverage on the whole-trace ledgers. The base's top four tokens are always scored; the realised token is scored in addition when it lies outside them, so a position has four or five measured actions. "Weighted" applies the inverse-inclusion weights of the population statistics. "Excluded tail mass" is the base probability outside the measured set, averaged over the five-action and four-action positions separately.

		realised token outside the base's top four			excluded tail mass	
cell	positions	count	share	weighted	when outside	when inside
Q0.5B/G	1250	8	0.6%	0.4%	0.285	0.023
Q0.5B/M	1000	7	0.7%	0.4%	0.179	0.016
Q1.5B/G	1250	10	0.8%	0.4%	0.106	0.014
Q1.5B/M	1000	4	0.4%	0.2%	0.097	0.013
QM7B/G	1000	5	0.5%	0.3%	0.206	0.009
QM7B/M	500	0	0.0%	0.0%	–	0.010
QM7B-Oat/G	750	0	0.0%	0.0%	–	0.003
QM7B-SFT/G	500	64	12.8%	7.1%	0.284	0.075
L8B/G	1000	28	2.8%	1.6%	0.415	0.082
Q7B/G	500	3	0.6%	0.3%	0.144	0.009
Q7B/M	400	0	0.0%	0.0%	–	0.006
Q7B-ORZ/G	500	11	2.2%	1.2%	0.152	0.009
Q7B-ORZ/M	400	1	0.2%	0.1%	0.042	0.010

			policy's own opening			base's own opening (placebo)			
cell	problems	$R(0)$	prefix	cover	lift	prefix	cover	lift	policy − base
Q1.5B/G	250	0.481	`First, we need`	38%	+0.008 [−0.012,+0.028]	`Step 1:`	57%	+0.042 [+0.020,+0.061]	−0.034 [−0.052,−0.016]
QM7B/G	200	0.434	`First, let's`	58%	+0.264 [+0.233,+0.294]	`To solve this problem`	59%	+0.267 [+0.231,+0.304]	−0.003 [−0.022,+0.018]
L8B/G	200	0.371	`Number of pounds of`	1%	−0.197 [−0.222,−0.172]	`Step 1:`	32%	−0.021 [−0.040,−0.001]	−0.176 [−0.200,−0.152]
Q7B/G	150	0.607	`To find out how`	58%	−0.002 [−0.028,+0.022]	`Step 1:`	97%	+0.009 [−0.016,+0.035]	−0.012 [−0.034,+0.011]
Q7B-ORZ/G	150	0.789	`To determine how many`	35%	−0.031 [−0.052,−0.011]	`First, let's`	87%	+0.011 [−0.007,+0.030]	−0.042 [−0.066,−0.020]

Table 38: Placebo for the format arm, on the five cells where it was run. The generic-prefix control is run twice on the same problems: with the *policy's* most frequent 4-token openings (the paper's arm, the top three pooled, the top one named) and with the *base's* (likewise; the two pools share one string on the primary cell, `To solve this problem`, the policy's third). Intervals are re-drawn, so they differ from Table 14's in the third decimal. "Cover" is the share of problems for which that model produces that opening in at least one of its own 16 samples. Lifts are over $R(0)$, with cluster-bootstrap intervals over problems.

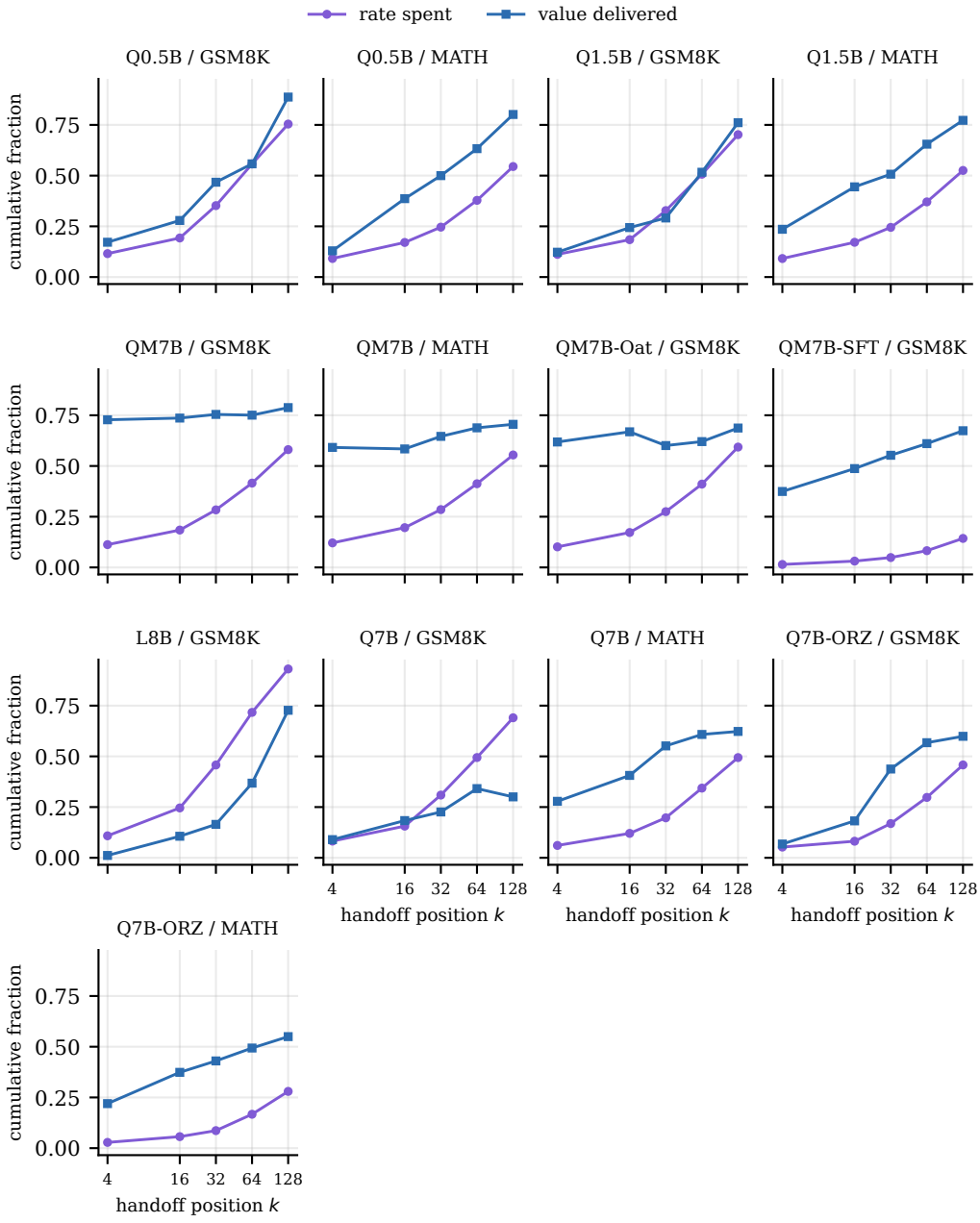


Figure 5: Cumulative share of the rate spent and of the gain delivered ($\rho(k)$) as a function of position, per cell; point estimates.

F EXTENSION RUNS

What was run. Ten extensions use the pipeline, decoding settings and estimators of the main cells unchanged (except the longer cap and handoff grid of (iv), the baselines that (vi), (viii), (ix) and (x) change by design, and the first-answer grade that (ix) and (x) also record) and are analysed with the same scripts on their own records, so no interval reported elsewhere changes (Table 39, Figure 6). (i) *Recipe crossing*: the SimpleRL-Zoo recipe on `Qwen2.5-Math-7B`, scored against the primary pair's base, template and value function (the first 168 GSM8K problems). (ii) A second math-specialised base, `DeepSeek-Math-7B`, with its SimpleRL-Zoo descendant, and (iii) a small math base, `Qwen2.5-Math-1.5B`, with its Oat-Zero descendant, each under the template its trained model prefers. (iv) A start that already writes the answer format: `DeepSeek-R1-Distill-Qwen-1.5B` with its RL descendant `DeepScaleR-1.5B-Preview` (Luo et al., 2025), under their shared chat template with a 4096-token cap. (v) A contamination control: GSM-Symbolic (Mirzadeh et al., 2025), whose problems re-instantiate GSM8K templates with new names and numbers, so a memorised answer to the original is wrong; 100 templates, two instances each, for the primary pair and both reward controls. (vi) The reverse handoff on the primary cell. (vii) Template pilots on 50 held-out GSM8K problems for the bases and the primary base's descendants, and the divergence pass on the `Qwen2.5-7B` cells (Appendix G). (viii) The primary base's descendants (and the 1.5B Qwen-Math pair) against the base under the Question/Answer template. (ix) The primary cell with the base sampling at other temperatures. (x) Three RLVR descendants of the primary base against the base with no template. Labels extend Section 4's: QM7B-Zoo, DSM7B (`DeepSeek-Math-7B`), QM1.5B-Oat, R1D1.5B-DSR; /S marks GSM-Symbolic and -rand, -gt the reward controls. The DeepSeek-Math and R1-distilled checkpoints ship byte-level BPE tokenizers under a Llama tokenizer class; they are loaded as the fast tokenizer their `tokenizer.json` defines, and every served model passes an exact encode–decode round trip first.

Recipe crossing. The SimpleRL-Zoo recipe, whose descendants of `Qwen2.5-7B`, the 0.5B and 1.5B Qwen bases and Llama-3.1-8B do not front-load, does front-load on `Qwen2.5-Math-7B`: $\rho(4) = 0.64$ $[0.56, 0.72]$ (gap 0.391), and $\rho(128) = 0.64$ (Figure 6(a)). Its early lift is format, as on the primary cell: its most common opening is the base's own (`To solve this problem`), and the generic prefix lifts the base by +0.266 against a real four-token lift of +0.238, content -0.028 $[-0.056, +0.001]$; four worked examples close 53% of its gap. Its dense check passes (shortfall -0.007 against the Stage-1 lift; shortfall is lift minus summed advantages, so negative is the over-count side); against the direct lift on the same positions the two-sided error, +0.051 [+0.006,+0.095], resolves an under-count, the direction truncation predicts, the two lift estimates differing by their noise. Over whole traces its ledger resembles its siblings' (Table 40): $\Lambda = 36\%$ lies between theirs, its cross-fitted search margin, -0.0009 [-0.0030,+0.0012], covers zero, and 63% of the rate sits at zero estimated dispersion, somewhat less than on either sibling. With the Oat-Zero and SimpleRL-Zero descendants and both reward controls, five RL-trained descendants of this base (four recipes, one of them on a random reward) front-load, and the same recipe under the same chat template on `Qwen2.5-7B` does not, nor on `DeepSeek-Math-7B`, whose base partly echoes that template (below): in the checkpoints tested, front-loading goes with the base under its pair's template, not with the recipe (whose training subsets are chosen per base). It depends on the template too: against the base under the Question/Answer form the primary cell front-loads less (Table 44).

The reverse handoff. Swapping the roles (the base writes k tokens, the primary policy finishes) gives the other conserved decomposition: $R_{\mathrm{rev}}(0)$ is the policy's accuracy, $R_{\mathrm{rev}}(\infty)$ the base's traces', and the increments are advantages of the base's tokens under the *policy's* value function. It front-loads too, less: the base's first four tokens cost $\rho_{\mathrm{rev}}(4) = 0.48$ [0.41, 0.55] of the gap and the next 124 a further 0.31 ($\rho_{\mathrm{rev}}(128) = 0.79$), against 0.06 forward (Table 41, Figure 6(b); 350 traces reverse, 385 forward). Each share is of its own traces' gap (0.422 reverse, 0.361 forward), and the reverse length guard keeps the base's longer, harder traces. The two decompositions credit positions differently, as they must, but both put a large share of the gain at the first four tokens; no generic arm was run in this direction, so that the base's opening costs the policy through its format is a reading, not a measurement.

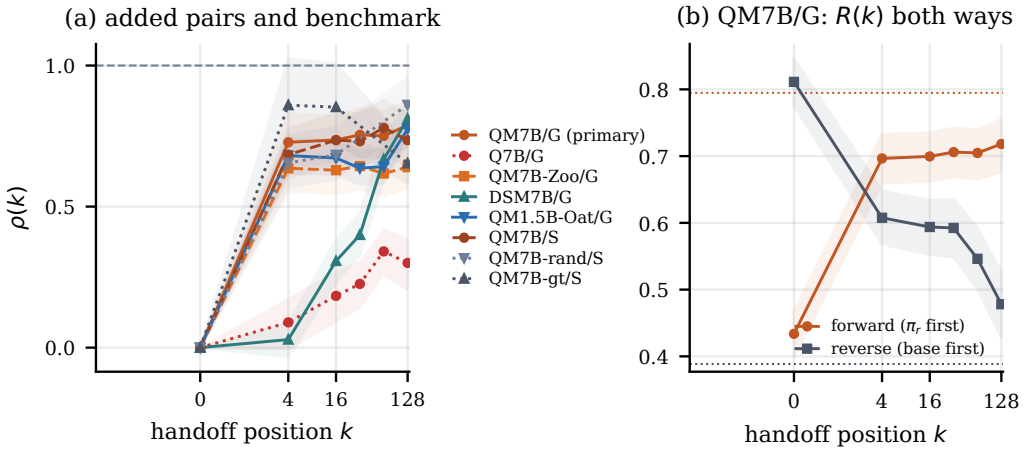


Figure 6: Extension runs. (a) $\rho(k)$ for the added pairs and for the primary pair and both reward controls on GSM-Symbolic (/S), with QM7B/G and Q7B/G for reference; labels as in Table 39. (b) The primary cell's transplant curve $R(k)$ in both directions; dotted lines mark $R(\infty)$. Both panels: 95% cluster-bootstrap bands.

Table 39: Extension cells (below the rule) against main cells (above): accuracies, $\rho(k)$ at $k = 4, 16$ and the largest grid point (the subscript), and the four-token format/content split (real lift, generic-prefix lift, content = real − generic with its interval, format share, printed only where the real lift's interval lies above zero at the printed precision and the generic lift is not negative; –: not run or not printed), the placebo difference (policy's opening minus the base's own), and the share of the gap four worked examples close.

					$\rho(k)$			four-token lift					4-shot
cell	n	$J(\pi_b)$	$J(\pi_r)$	gap	k=4	16	last	real	generic	content	format (%)	placebo diff.	elicited (%)
QM7B/G	200	0.434	0.813	+0.379	0.73 [0.64,0.83]	0.74	0.79_{128}	+0.264	+0.264	+0.000 [−0.023,+0.022]	100	−0.003 [−0.022,+0.018]	56
QM7B-Oat/G	200	0.434	0.893	+0.459	0.62 [0.55,0.68]	0.67	0.69_{128}	+0.278	+0.229	+0.049 [+0.025,+0.073]	83	–	46
Q7B/G	150	0.607	0.894	+0.287	0.09 [−0.01,0.17]	0.18	0.30_{128}	+0.029	−0.002	+0.031 [+0.006,+0.056]	–	−0.012 [−0.034,+0.011]	−1
Q1.5B/G	250	0.481	0.764	+0.283	0.12 [0.04,0.21]	0.24	0.76_{128}	+0.035	+0.008	+0.027 [+0.008,+0.044]	23	−0.034 [−0.052,−0.016]	27
Q0.5B/G	250	0.274	0.505	+0.231	0.17 [0.07,0.27]	0.28	0.89_{128}	+0.044	+0.030	+0.014 [−0.005,+0.033]	69	–	31
L8B/G	200	0.371	0.813	+0.443	0.01 [−0.07,0.08]	0.11	0.73_{128}	+0.047	−0.197	+0.244 [+0.212,+0.275]	–	−0.176 [−0.200,−0.152]	15
QM7B-Zoo/G	168	0.434	0.825	+0.391	0.64 [0.56,0.72]	0.63	0.64_{128}	+0.238	+0.266	−0.028 [−0.056,+0.001]	112	+0.003 [−0.010,+0.016]	53
DSM7B/G	200	0.234	0.701	+0.467	0.03 [−0.03,0.08]	0.31	0.81_{128}	+0.025	+0.041	−0.016 [−0.037,+0.005]	–	+0.107 [+0.085,+0.128]	69
QM1.5B-Oat/G	250	0.376	0.829	+0.454	0.68 [0.62,0.74]	0.67	0.78_{128}	+0.311	+0.252	+0.059 [+0.039,+0.081]	81	+0.146 [+0.128,+0.164]	55
R1D1.5B-DSR/G	200	0.842	0.874	+0.032	−0.05 [−0.74,0.23]	0.17	0.20_{64}	−0.002	+0.009	−0.010 [−0.023,+0.001]	–	+0.007 [−0.003,+0.017]	–
QM7B/S	200	0.420	0.768	+0.348	0.68 [0.60,0.78]	0.74	0.74_{128}	+0.232	+0.250	−0.018 [−0.041,+0.005]	108	+0.070 [+0.045,+0.095]	60
QM7B-rand/S	200	0.420	0.731	+0.312	0.65 [0.57,0.75]	0.68	0.86_{128}	+0.217	+0.220	−0.002 [−0.027,+0.022]	101	–	–
QM7B-gt/S	200	0.420	0.743	+0.323	0.86 [0.74,1.02]	0.85	0.65_{128}	+0.262	+0.209	+0.053 [+0.031,+0.074]	80	–	–

Two more bases. On `DeepSeek-Math-7B`, math-specialised like Qwen-Math but pretrained separately, the SimpleRL-Zoo descendant does not front-load: $\rho(4) = 0.03$ $[-0.03, 0.08]$, climbing through 0.31 at $k = 16$ to 0.81 at $k = 128$, as on the small Qwen and Llama pairs; its four-token lift, +0.025 $[+0.000, +0.048]$, is small, and four worked examples close 69% of a 0.467 gap, more than on any front-loading cell but the random-reward control (71%), so elicitation by prompting does not by itself imply front-loading (though here the worked examples also cure the template echo described next). Under the chat template its descendant prefers, the base often echoes the template itself (`<|im_` opens 82% of its samples), so its own most frequent opening is no placebo there (it costs −0.066). On `Qwen2.5-Math-1.5B` the Oat-Zero descendant front-loads like the 7B pairs, $\rho(4) = 0.68$ $[0.62, 0.74]$, and its early lift is mostly format: the generic prefix gives +0.252 of a +0.311 lift (81%), leaving a resolved content of +0.059 $[+0.039, +0.081]$—the pattern of Oat-Zero on the 7B base (83%, +0.049). Its placebo separates, as on GSM-Symbolic: the base's own opening lifts it by +0.106 against +0.252 for the policy's (difference +0.146 [+0.128,+0.164]). Its base, under the chat template, truncates 39% of its samples at the 1024-token cap, and the handoff cuts that to 13% by $k = 4$, so up to 83% of that lift could be the base being taught to stop (at most 45% on the primary cell), itself a matter of format. The dense checks pass on both (shortfalls −0.018 and −0.008). On DeepSeek-Math the whole-trace sum of the sampled advantages overshoots its target by +0.86 [+0.13,+1.66], the one cell where that check resolves a discrepancy under temperature-1 weights (the primary cell's is resolved only under the sampler's; Appendix D), so its whole-trace quantities (Table 40) are reported, not interpreted; on the 1.5B pair the sum covers its target (difference +0.30 [−0.32,+0.95], an interval wider than the gap, so the check has no power there) and the ledger reads

Table 40: Whole-trace ledgers of the extension cells (below the rule) and of the primary base's two main RLVR cells: Λ (the gap over the mean per-trace ceiling total, plug-in), the search margin's naive and cross-fitted bounds, the share of the rate at zero estimated dispersion (m=8), the whole-trace sum of sampled advantages minus its target (Appendix D), and the dense check (shortfall: $R(4) - R(0)$ minus $\sum_{t\le 4} \hat{A}_t$, negative on the over-count side, and whether its interval meets the truncation allowance). †: the whole-trace sum of advantages misses its target (Appendix F); that row's whole-trace quantities are not interpreted.

cell	positions	Λ (%)	margin, naive	margin, cross-fit	wasted rate (%)	whole-trace sum − target	dense shortfall	in band
QM7B/G	1000	16	−0.0035 [−0.0051,−0.0021]	−0.0000 [−0.0017,+0.0014]	72	+0.99 [−0.01,+2.19]	+0.046 [+0.006,+0.086]	yes
QM7B-Oat/G	750	47	−0.0019 [−0.0033,−0.0007]	+0.0017 [+0.0001,+0.0032]	74	+0.18 [−0.36,+0.72]	+0.015 [−0.022,+0.052]	yes
QM7B-Zoo/G	500	36	−0.0040 [−0.0062,−0.0021]	−0.0009 [−0.0030,+0.0012]	63	−0.43 [−1.17,+0.39]	−0.007 [−0.066,+0.050]	yes
DSM7B/G†	500	19	−0.0077 [−0.0114,−0.0043]	−0.0014 [−0.0051,+0.0020]	54	+0.86 [+0.13,+1.66]	−0.018 [−0.070,+0.032]	yes
QM1.5B-Oat/G	625	41	−0.0018 [−0.0031,−0.0005]	+0.0008 [−0.0008,+0.0022]	74	+0.30 [−0.32,+0.95]	−0.008 [−0.064,+0.047]	yes

Table 41: The primary cell's handoff in both directions: forward (the paper's ledger, under the base's values) and reverse (under the policy's values), on the length-guarded traces of each direction.

	$R(k)$; reverse row $R_{\mathrm{rev}}(k)$							$\rho(k)$; reverse row $\rho_{\mathrm{rev}}(k)$				
handoff	k=0	4	16	32	64	128	∞	4	16	32	64	128
forward: π_r writes k, π_b finishes	0.434	0.697	0.700	0.706	0.705	0.718	0.795	0.73	0.74	0.75	0.75	0.79
reverse: π_b writes k, π_r finishes	0.811	0.608	0.594	0.592	0.546	0.478	0.389	0.48	0.51	0.52	0.63	0.79

like its 7B siblings' ($\Lambda = 41\%$, cross-fitted margin +0.0008 [−0.0008,+0.0022], 74% of the rate at zero estimated dispersion). Front-loading thus follows the Qwen-Math family across two sizes (four recipes at 7B, a random-reward control among them, and one at 1.5B, where up to 83% of the lift could be stopping but the curve still front-loads against the Question/Answer base, Table 44); it is not observed on `DeepSeek-Math-7B`, but under a template that base partly echoes rather than follows, so that contrast is weaker than the recipe crossing.

A start that already writes the format. `DeepSeek-R1-Distill-Qwen-1.5B`, distilled from long reasoning traces, already writes in the register its RL descendant writes: the two models' most frequent openings are the same string, `Okay, so I` (88% of the base's problems, 90% of the policy's). RL adds 0.032 on GSM8K (0.842 to 0.874), too little for ρ to be read ($\rho(4) = -0.05$ $[-0.74, 0.23]$), but the four-token lift itself is bounded near zero: -0.002 $[-0.015, +0.011]$, with the generic prefix at +0.009 and the base's own opening at +0.002. Where the start already writes the format the first four tokens buy nothing, consistent with the format reading, though with so small a gap little else was possible: the pair is a sanity check, not a measurement of where RL's gain lives, and it is tabulated but not plotted. The two pools share two of their three strings (the policy's "`Okay, so I`", "`Okay, let's`", "`Okay, let me`"; the base's "`Okay, so I`", "`Alright, so I`", "`Okay, let me`"), so its generic and placebo arms nearly coincide. Its curve's largest point under the 128-token guard is $k = 64$ (the run's grid is 4, 16, 64, 256, 1024), and at the 4096-token cap the base truncates 14% of its samples and the policy 8%. Its dense check passes (shortfall +0.007, two-sided error +0.007 [−0.023,+0.038]).

A contamination control. GSM-Symbolic re-instantiates GSM8K templates with new names and numbers (Mirzadeh et al., 2025); a model that had memorised the original problem's answer would give a wrong one, so an opening that "unlocks" memorised answers cannot lift accuracy here (memorised solution procedures for a template still could). We take templates 0–99 and instances 0 and 1 of each (`apple/GSM-Symbolic`, main configuration); intervals resample instances; the two instances of a template are not independent, but resampling templates instead moves no interval edge quoted here by more than 0.01 (`code/gsmsym_template_boot.py`). The base scores 0.420 and the SimpleRL-Zero policy 0.768 (0.348 gap, against 0.379 on GSM8K), the curve front-loads as on GSM8K ($\rho(4) = 0.68$ $[0.60, 0.78]$, $\rho(128) = 0.74$), and its two most frequent openings are the same strings as on GSM8K ("`First, let's`", "`First, we need`"; its third is "`To determine how many`"). Grafted into every problem the three lift the base by +0.250 against a real four-token lift of +0.232: content -0.018 $[-0.041, +0.005]$, format share 108%; four worked examples close 60% of the gap (Table 39). The placebo separates here, unlike on GSM8K: the base's own most frequent openings (`To solve this problem`, `Let's break down`, `To solve the problem`, none of them in the policy's pool here) lift the base by +0.180 against +0.250 for the policy's, a difference of +0.070 [+0.045,+0.095], so on these problems the policy's opening is a

better format than the base's own—still a problem-agnostic one. The early gain on the primary cell is format on problems where memorised answers cannot pay. The reward controls' format/content split replicates: on GSM-Symbolic the random-reward policy's four-token lift is all format (generic +0.220 against +0.217, content -0.002 $[-0.027, +0.022]$), and the correct-reward policy's exceeds it by resolved content, +0.053 $[+0.031, +0.074]$, as on GSM8K. Their timing does not: here the correct-reward policy has the higher point estimate ($\rho(4) = 0.86$ against 0.65, overlapping intervals) and its curve then falls ($\rho(128) = 0.65$), so the GSM8K reading that the correct reward adds later-trace value is not replicated here (the ordering is not resolved either way; Table 39).

Held-out template pilots. The template pilot of Table 1 ran 12 problems on the trained models. Table 42 repeats it on 50 GSM8K problems never scored elsewhere (indices 250–299; 8 samples each) for the four bases scored under the chat template and for SimpleRL-Zero and the two reward controls. Every base scores higher under the Question/Answer form (by 0.07–0.24, point estimates on 50 problems) and each of the three piloted descendants under the chat form (by 0.15–0.24), so on the chat-template pairs the shared template, chosen on the trained model, is the base's worse one (the Question/Answer pairs use the form Table 1 chose for their trained models, the form this pilot finds better for every base). The measured gaps there contain a template component, which Table 4 and the next paragraph measure on the primary base; the few-shot arm, a different baseline, runs on 11 main cells.

Table 42: Template pilot on 50 held-out GSM8K problems (indices 250–299), 8 samples each at temperature 0.8: accuracy and share of samples reaching the token cap under both prompt forms. Last column: the form the model is scored under in its pair (the Open-Reasoner-Zero pair scores `Qwen2.5-7B` under that model's own prompt instead).

	Question/Answer		chat, boxed		
model	accuracy	truncated (%)	accuracy	truncated (%)	pair's template
Qwen2.5-Math-7B	0.618	6	0.527	18	chat, boxed
SimpleRL-Zero	0.590	14	0.828	16	chat, boxed
GRPO, random reward	0.665	7	0.810	6	chat, boxed
GRPO, correct reward	0.738	3	0.890	0	chat, boxed
DeepSeek-Math-7B	0.312	24	0.245	1	chat, boxed
Qwen2.5-Math-1.5B	0.632	2	0.388	38	chat, boxed
Qwen2.5-7B	0.740	0	0.627	13	chat, boxed

The base under the Question/Answer template. The pilots above find the Question/Answer form better for every base, so the chat template the Qwen-Math pairs share is the base's worse one, and part of what the early tokens repair could be that template. We reran the primary cell's transplant curve and its generic arm with the base reading the Question/Answer template (same problems, same trained traces, same three grafted openings; Tables 4 and 43). On the 99 of the first 100 GSM8K problems whose traces pass the 128-token guard (Table 4) the base starts +0.164 [+0.115, +0.214] higher, at 0.588, and the policy's first four tokens take it to the same $R(4)$ as under the chat template; the four-token lift that remains is +0.100 [+0.054, +0.145], $\rho(4) = 0.43$ [0.26, 0.60] against 0.67 under the chat template on these problems, so the curve still front-loads, less. The generic openings no longer carry it (Table 43, all 100 problems, traces longer than 32 tokens as in Table 14): they lift the base by +0.025 $[-0.007, +0.058]$ against a real +0.099 [+0.054, +0.144], content +0.074 [+0.034, +0.113] (on the traces whose opening is not one of the grafted strings, +0.095 [+0.051, +0.141], 92 problems). The format credit of Section 4.2 is therefore credit relative to the pair's template: a better template supplies it, and against the Question/Answer baseline the early lift is smaller and resolved as content. The policy's openings are partly chosen per problem: its most frequent opening by samples, `To determine how many`, is not among the grafted strings, which are the openings covering the most problems, because it varies with the question (`how much`, `the total`). Under the distilled policy's chat template, where the base starts higher still (0.773), neither arm lifts it (real +0.009 $[-0.015, +0.034]$, generic -0.002 $[-0.027, +0.021]$; on the 99 guarded problems of Table 4, +0.002 $[-0.023, +0.028]$).

The other descendants against the Question/Answer baseline. The same rerun for the primary base's other descendants (Table 44; first 100 GSM8K problems, the traces of the main cells, each policy's own three most frequent openings grafted) shows the same two effects on every RLVR one

Table 43: Four-token format/content split on the primary cell with the base reading three templates: the pair's chat template (the paper's arm, all problems), the Question/Answer template and the distilled policy's chat template (first 100 problems). Real lift $R(4) - R(0)$ with the policy's own four tokens; generic lift with the three pooled grafted openings under the same template; content = real − generic; 95% cluster-bootstrap intervals.

base's template	problems	$R(0)$	real lift	generic lift	content
chat, boxed (the pair's)	200	0.434	+0.264 [+0.227,+0.300]	+0.264 [+0.232,+0.296]	+0.000 [−0.023,+0.022]
Question/Answer	100	0.591	+0.099 [+0.054,+0.144]	+0.025 [−0.007,+0.058]	+0.074 [+0.034,+0.113]
the distilled policy's chat	100	0.773	+0.009 [−0.015,+0.034]	−0.002 [−0.027,+0.021]	+0.011 [−0.015,+0.041]

of them. The template supplies about half of the early lift (46–63%): each RLVR policy's four-token lift falls from 0.23–0.29 against the chat-template base to 0.09–0.15 against the Question/Answer one, and the curves still front-load, less ($\rho(4) = 0.39$–0.57 where the gap is large enough to normalise by). Most of what a *random* reward adds to this base is what the Question/Answer template also gives it: that policy's gap falls from 0.262 to 0.096, while the correct-reward policy keeps 0.259 of 0.426 and its early lift a resolved content, +0.060 [+0.027, +0.093]. What remains of the early lift is still mostly generic on the other 4 RLVR descendants (59–80%), though its content is resolved on the random-reward and correct-reward policies (+0.032, +0.060); the primary cell, at 25%, is the exception. The distilled descendant's early lift nearly vanishes against the Question/Answer base (+0.152 to +0.027, $\rho(4)$ from 0.37 to 0.11), and with it the early content it carried against the chat-template base (now +0.013 [−0.022, +0.049]), so the RLVR–distillation ordering of Section 4.2 is itself relative to the template. The 1.5B Qwen-Math pair behaves like the primary cell (bottom row; all 250 problems, `code/formal/run_r9k.sh`): the Question/Answer base starts +0.275 higher, the four-token lift falls from +0.311 to +0.107 (by 66%), the curve still front-loads ($\rho(4)$ from 0.68 to 0.59 [0.45, 0.73]), and what remains is resolved content, +0.061 [+0.041, +0.081].

Table 44: The primary base's descendants (and, below the rule, the 1.5B Qwen-Math pair on all 250 problems) against the base under its pair's chat template and under the Question/Answer template (first 100 GSM8K problems, the same trained traces). $R(0)$ (over problems), $R(\infty)$ (over traces), the gap and $\rho(4)$ on the traces longer than 128 tokens (n: problems with one); four-token lifts on all the problems' traces longer than 32 tokens, with each policy's three grafted openings under the same template; 95% cluster-bootstrap intervals over problems. The primary cell's Question/Answer entries are those of Tables 4 and 43; its chat-template entries are on the same 100 problems. ‡: gap below 0.15, ρ not interpreted.

		$R(0)$			gap		$\rho(4)$		four-token lift			
policy	n	chat	Q/A	$R(\infty)$	chat	Q/A	chat	Q/A	real, chat	real, Q/A	generic, Q/A	content, Q/A
QM7B/G (SimpleRL-Zero)	99	0.424	0.588	0.822	0.398	0.234	0.67	0.43 [0.26,0.60]	+0.268	+0.099	+0.025	+0.074 [+0.034,+0.113]
QM7B-Oat/G	100	0.424	0.591	0.895	0.471	0.304	0.58	0.49 [0.35,0.62]	+0.275	+0.148	+0.118	+0.030 [−0.008,+0.067]
QM7B-Zoo/G	100	0.424	0.591	0.827	0.403	0.236	0.59	0.39 [0.24,0.56]	+0.230	+0.090	+0.059	+0.031 [−0.005,+0.070]
QM7B-rand/G	100	0.424	0.591	0.686	0.262	0.096	0.96	1.32‡	+0.255	+0.128	+0.096	+0.032 [+0.001,+0.067]
QM7B-gt/G	100	0.424	0.591	0.850	0.426	0.259	0.68	0.57 [0.42,0.75]	+0.289	+0.149	+0.088	+0.060 [+0.027,+0.093]
QM7B-SFT/G	100	0.424	0.591	0.830	0.406	0.239	0.37	0.11 [−0.04,0.24]	+0.152	+0.027	+0.014	+0.013 [−0.022,+0.049]
QM1.5B-Oat/G	250	0.376	0.650	0.832	0.456	0.182	0.68	0.59 [0.45,0.73]	+0.311	+0.107	+0.046	+0.061 [+0.041,+0.081]

The base with no template. Liu et al. (2025b) find the Qwen2.5-Math bases strongest with no template at all, and so is ours: reading the question text alone (`code/formal/run_r9m.sh`, `run_r9n.sh`; first 100 GSM8K problems, the same trained traces; `code/template_raw.py`), `Qwen2.5-Math-7B` scores 0.703, against 0.588 under the Question/Answer template and 0.424 under the chat template, truncating 4% of its samples. Against that base the SimpleRL-Zero policy's gap on these problems is 0.119 (not 0.398), and its first four tokens add nothing: the four-token lift is −0.005 and $\rho(4) = -0.02$ [−0.46, 0.28], and the curve reaches only 0.43 [0.05, 0.78] by $k = 128$, so what the policy still adds over this base arrives later in the trace. Grafting the policy's three most frequent openings moves this base by −0.041, and the content left over, +0.036 [−0.004, +0.081], is not resolved on last answers and only barely on first answers (+0.047 [+0.003, +0.094]). The Oat-Zero and SimpleRL-Zoo descendants agree: against the no-template base their four-token lifts are +0.028 and −0.001, $\rho(4)$ is 0.14 [−0.02, 0.33] and −0.03 [−0.48, 0.24], and no content is resolved (+0.022 [−0.016, +0.060] and +0.001 [−0.036, +0.040]). Front-loading on this base is therefore a property of the pair's chat template: against the base's strongest template-free form there is no early credit to decompose (Table 45).

Table 45: RLVR descendants of `Qwen2.5-Math-7B` against the base under the pair's chat template, the Question/Answer template (both as in Table 44) and no template (the question text alone), first 100 GSM8K problems, the same trained traces. $R(0)$ over problems; gap and $\rho(4)$ on the traces longer than 128 tokens; four-token lifts on the traces longer than 32 tokens, with each policy's three most frequent openings grafted for the generic arm; 95% cluster-bootstrap intervals over problems.

	base's template	$R(0)$	gap	$\rho(4)$	real	generic	content
QM7B/G (SimpleRL-Zero)	chat	0.424	0.398	0.67 [0.56,0.79]	+0.268	+0.275	−0.007 [−0.046,+0.029]
	Q/A	0.588	0.234	0.43 [0.26,0.60]	+0.099	+0.025	+0.074 [+0.034,+0.113]
	none	0.703	0.119	−0.02 [−0.46,0.28]	−0.005	−0.041	+0.036 [−0.004,+0.081]
	none, first answers	0.755			+0.004	−0.043	+0.047 [+0.003,+0.094]
QM7B-Oat/G	chat	0.424	0.471	0.58 [0.50,0.66]	+0.275	+0.223	+0.051 [+0.017,+0.085]
	Q/A	0.591	0.304	0.49 [0.35,0.62]	+0.148	+0.118	+0.030 [−0.008,+0.067]
	none	0.705	0.190	0.14 [−0.02,0.33]	+0.028	+0.006	+0.022 [−0.016,+0.060]
	none, first answers	0.755			+0.027	−0.008	+0.034 [−0.004,+0.072]
QM7B-Zoo/G	chat	0.424	0.403	0.59 [0.49,0.70]	+0.230	+0.268	−0.038 [−0.075,−0.001]
	Q/A	0.591	0.236	0.39 [0.24,0.56]	+0.090	+0.059	+0.031 [−0.005,+0.070]
	none	0.705	0.122	−0.03 [−0.48,0.24]	−0.001	−0.002	+0.001 [−0.036,+0.040]
	none, first answers	0.755			−0.004	+0.014	−0.018 [−0.053,+0.018]

The pricing layer against the Question/Answer baseline. The primary cell's whole-trace ledger was re-run with the base reading the Question/Answer template (`code/formal/run_r9i.sh`: the first 100 GSM8K problems, the same trained traces, positions re-drawn by the rates the base assigns under that template, fresh rollouts at m=8; 500 positions) and read against the main ledger restricted to the same problems. The baseline changes the scale of what is priced, not the pattern (each pair below: chat template, then Question/Answer): the gap falls from 0.390 to 0.223 and Λ with it (13% to 9%), while the share of the rate at zero estimated dispersion stays at 70% [59, 81] against 72% [60, 84], rate tracks $|A_t|$ at 0.71 and 0.70 and the value term at −0.11 and −0.13, dispersion tracks the value term at 0.39 and 0.39 (excess over the shared-rate null +0.15 and +0.14, both beyond every null draw), and the cross-fitted search margin covers zero on both (+0.0005 [−0.0017, +0.0025] and +0.0004 [−0.0028, +0.0032]). The whole-trace sums cover their targets on both (+1.33 [−0.16, +3.18] and −0.86 [−3.47, +1.36]). The room–worth reading and the rate accounting are properties of the base's values along the trained traces, not of the chat template's handicap.

The base at other temperatures. A policy includes its decoding, so the base's sampling temperature is part of the baseline too, and a base sampled at $T = 0.8$ could lose to the trained policy by drifting into bad openings that a sharpened policy avoids. We re-measured the primary cell with the base sampling at $T = 0.3, 0.6, 0.8, 1.0$ (nucleus 0.95; `code/formal/run_r9j.sh`: the first 100 GSM8K problems, the same trained traces, Stage 0 and the transplant and generic arms re-run at each temperature; `code/temperature_sweep.py`). Table 46 gives the results. A cooler base is stronger and the four-token lift it leaves smaller, a hotter one the reverse: $R(0)$ runs from 0.569 at $T = 0.3$ to 0.334 at $T = 1.0$ and the real lift from 0.218 to 0.327. At every temperature the grafted openings carry 95–106% of the lift and no content interval excludes zero, and the curve still front-loads ($\rho(4) = 0.66$–0.87, highest for the coolest base). Graded on first answers the lift is larger at every temperature (by +0.025 to +0.060), with content resolved only at $T = 0.8$. The base's temperature, like its template, sets how large the early lift is; it does not change what carries it.

Table 46: The primary cell against the base sampling at other temperatures (first 100 GSM8K problems, the same trained traces). $R(0)$, $R(\infty)$ and $\rho(k)$ on the traces longer than 128 tokens (n problems); four-token lifts on the traces longer than 32 tokens, graded on the last answer (as everywhere) and on the first; format share = generic/real where the real lift is resolved. *The paper's baseline, $T = 0.8$: its curve from the main records, its four-token arms from a rerun that records both grades (run to run: Appendix D). 95% cluster-bootstrap intervals.

				$\rho(k)$		four-token lift, last answer				first answer			
base T	n	$R(0)$	$R(\infty)$	k=4	128	real	generic	content	format %	$R(0)$	real	generic	content
0.3	99	0.569	0.822	0.87 [0.69,1.12]	0.97	+0.218	+0.230	−0.012 [−0.054,+0.028]	106	0.582	+0.243	+0.235	+0.008 [−0.032,+0.046]
0.6	99	0.509	0.822	0.74 [0.59,0.90]	0.82	+0.235	+0.225	+0.009 [−0.029,+0.047]	96	0.540	+0.283	+0.262	+0.021 [−0.016,+0.056]
0.8*	99	0.424	0.822	0.67 [0.55,0.79]	0.74	+0.279	+0.265	+0.014 [−0.022,+0.052]	95	0.472	+0.339	+0.300	+0.039 [+0.004,+0.073]
1.0	99	0.334	0.822	0.66 [0.57,0.75]	0.74	+0.327	+0.309	+0.017 [−0.021,+0.054]	95	0.383	+0.387	+0.364	+0.022 [−0.012,+0.055]

G FURTHER CHECKS

The analyses in this appendix re-read the released records (plus the divergence pass of Appendix F (vii)); none changes a number reported elsewhere in the paper.

A two-sided dense check. The conservation check of Section 4.1 compares $\sum_{t\le 4}\hat{A}_t$ with $R(4)-R(0)$ against a one-sided allowance. A two-sided check needs no allowance. Transitions are deterministic, so the realised token's own rollouts at position $t-1$ are a direct estimate of $V^{\pi_b}(s_t)=Q^{\pi_b}(s_{t-1},y_{t-1})$ (at $t=1$, the base's accuracy on the problem), from rollouts disjoint from those behind the renormalised estimate $\hat{V}_{\mathrm{est}}(s_t)$ that every $\hat{A}_t$ uses. Their difference estimates the estimator's error at each position, of either sign, and its sum over $t\le 4$ is exactly the telescoped direct lift minus the summed advantages (the realised tokens' values cancel), so a positive sum is an under-count. On 15 of 15 cells the summed error's 95% interval covers zero (point estimates -0.016 to +0.046, positive (the direction truncation predicts) on 11), and no interval edge exceeds 0.11 in magnitude (Table 47; 80–250 problems per cell), a margin larger than the four-token lift on most cells. Pooled with equal weight over the 13 main cells (the reward controls, above the rule in the table, are left out), the summed error is +0.014 $[+0.002,+0.025]$ with each cell's bootstrap replicates averaged, and the same, [+0.002, +0.025], when each benchmark's problems are resampled jointly across the cells that share them, so that cells sharing a base and problems keep their correlation: a small under-count over the four positions, of the sign the truncated action set predicts, carried by the first two (per-position means +0.009, +0.010, -0.003 and -0.003 at $t=1$–4). Of the 4 extension cells with a dense ledger the check resolves an under-count on QM7B-Zoo/G and an over-count on none. The estimator therefore under-counts slightly over the first four positions and is not detected to over-count there (over whole traces no bias common to the cells is detected; Appendix D). The error measured here is that of the whole estimator, so it includes the temperature-1 weighting as well as the truncated action set, and its sign bounds their net effect.

Table 47: Two-sided dense check: the renormalised value estimate minus the direct one from the previous position's realised rollouts, per position and summed over $t\le 4$ (cluster bootstrap over problems), with the summed estimated advantages and the direct lift they should equal; extension cells below the rule.

		$\hat{V}_{\mathrm{est}}(s_t)-\hat{V}_{\mathrm{direct}}(s_t)$						
cell	n	t=1	2	3	4	summed [95%]	$\sum_{t\le 4}\hat{A}_t$	direct lift
Q0.5B/G	250	+0.006	+0.007	+0.007	+0.000	+0.020 [−0.009,+0.049]	+0.022	+0.043
Q0.5B/M	200	+0.010	−0.001	−0.004	−0.007	−0.001 [−0.028,+0.024]	+0.058	+0.056
Q1.5B/G	250	−0.011	+0.011	+0.009	−0.022	−0.014 [−0.049,+0.021]	+0.050	+0.037
Q1.5B/M	200	+0.009	+0.020	−0.007	−0.004	+0.018 [−0.018,+0.054]	+0.052	+0.070
QM7B/G	150	+0.029	−0.008	−0.000	+0.003	+0.024 [−0.015,+0.061]	+0.226	+0.250
QM7B/M	80	+0.014	+0.001	+0.030	−0.011	+0.034 [−0.015,+0.086]	+0.133	+0.167
QM7B-Oat/G	200	+0.029	−0.008	+0.009	+0.004	+0.034 [−0.002,+0.069]	+0.264	+0.297
QM7B-SFT/G	100	+0.024	+0.012	+0.015	−0.005	+0.046 [−0.011,+0.105]	+0.094	+0.140
L8B/G	200	−0.004	+0.013	+0.015	−0.016	+0.008 [−0.027,+0.045]	+0.053	+0.060
Q7B/G	100	−0.015	+0.019	−0.025	+0.014	−0.007 [−0.071,+0.055]	+0.028	+0.021
Q7B/M	80	+0.003	+0.022	−0.037	−0.002	−0.016 [−0.066,+0.030]	+0.049	+0.033
Q7B-ORZ/G	100	+0.021	+0.012	+0.002	−0.015	+0.020 [−0.025,+0.065]	−0.009	+0.011
Q7B-ORZ/M	80	+0.009	+0.032	−0.051	+0.023	+0.014 [−0.035,+0.060]	+0.079	+0.093
QM7B-rand/G	100	+0.022	+0.018	−0.026	+0.025	+0.040 [−0.010,+0.089]	+0.213	+0.253
QM7B-gt/G	100	+0.047	−0.000	−0.035	+0.013	+0.024 [−0.027,+0.075]	+0.275	+0.300
QM7B-Zoo/G	100	+0.029	+0.016	−0.027	+0.033	+0.051 [+0.006,+0.095]	+0.215	+0.266
DSM7B/G	100	−0.002	−0.014	−0.020	+0.003	−0.033 [−0.071,+0.008]	+0.053	+0.020
QM1.5B-Oat/G	100	+0.014	+0.007	−0.031	+0.048	+0.037 [−0.019,+0.093]	+0.275	+0.312
R1D1.5B-DSR/G	100	+0.001	−0.002	−0.009	+0.017	+0.007 [−0.023,+0.038]	+0.003	+0.010

First-answer grading. Every arm is graded on its last answer. A base that answers correctly and then keeps writing can overwrite its answer, and a handoff that teaches it to stop is then credited to the trained tokens. Re-grading every sample on its *first* answer changes the gap by at least 0.01 on 8 of 15 cells (QM7B/G, QM7B/M, QM7B-Oat/G, QM7B-SFT/G, Q7B/G, Q7B/M, QM7B-rand/G, QM7B-gt/G; Table 48): it falls most on Q7B/G, by 0.082 (28% of it), and by 0.018 (5%) on the primary cell; on QM7B-rand/G it rises (+0.028), the policy overwriting more than the base. On these bases part of the measured gap is a stopping failure of the base. On the primary cell the transplant arms were re-graded too (a rerun of $k=4$ and the generic arm on the first 100 problems

that records both grades; `code/formal/run_r9j.sh`, `code/temperature_sweep.py`): on first answers $R(0)$ rises from 0.424 to 0.472 and the four-token lift from +0.279 to +0.339, the generic arm taking +0.300 of it (content +0.039 [+0.004, +0.073]; Table 46). The base overwrites correct answers after the handoff as well as before it, so the early lift is not the handoff teaching it to stop overwriting; truncation, the other stopping channel, is bounded in Appendix D.

Table 48: Accuracy graded on the last answer (as everywhere in the paper) and on the first, the share of samples with more than one answer, and the resulting gaps.

		base			trained policy			gap	
cell	n	last answer	first answer	>1 answer (%)	last	first	>1 (%)	last	first
Q0.5B/G	250	0.274	0.274	0	0.505	0.505	0	+0.231	+0.231
Q0.5B/M	200	0.129	0.128	2	0.370	0.370	0	+0.241	+0.241
Q1.5B/G	250	0.481	0.481	0	0.764	0.764	1	+0.283	+0.283
Q1.5B/M	200	0.214	0.215	1	0.558	0.558	1	+0.343	+0.343
QM7B/G	200	0.434	0.486	36	0.813	0.847	22	+0.379	+0.361
QM7B/M	150	0.443	0.486	26	0.730	0.733	6	+0.287	+0.246
QM7B-Oat/G	200	0.434	0.486	36	0.893	0.893	0	+0.459	+0.407
QM7B-SFT/G	100	0.424	0.472	36	0.846	0.848	76	+0.422	+0.376
L8B/G	200	0.371	0.371	0	0.813	0.813	0	+0.443	+0.443
Q7B/G	150	0.607	0.695	17	0.894	0.900	1	+0.287	+0.205
Q7B/M	100	0.489	0.537	13	0.756	0.751	1	+0.267	+0.214
Q7B-ORZ/G	150	0.789	0.785	1	0.945	0.941	65	+0.156	+0.156
Q7B-ORZ/M	100	0.439	0.438	3	0.786	0.786	57	+0.347	+0.349
QM7B-rand/G	200	0.434	0.486	36	0.732	0.812	12	+0.298	+0.326
QM7B-gt/G	200	0.434	0.486	36	0.865	0.865	0	+0.431	+0.379

The search margin where the comparator has budget. The comparator of Section 4.4 draws $N = e^{\varepsilon_t}$ tokens, about one at most positions. Splitting each whole-trace run's positions into rate terciles, the mean N is at most 1.03 in the bottom tercile and 2.02 in the middle one, where the margin is near zero almost by construction (cross-fitted margin within ± 0.0028 on the RLVR runs, resolved above zero on 2 of them by at most 0.0010; ± 0.0149 on the distilled one). In the top tercile N averages 1.6–399 draws, but the mean is driven by rare high-rate tokens: the median is 1.2–3.3 draws on the RLVR runs (16 on the distilled one), and 15–87% of top-tercile positions allow two or more. There the naive estimate (biased down) is negative on all 15 runs and the cross-fitted one (biased up) is resolved above zero on 2 (QM7B-Oat/G, Q7B/M), below on 1 (Q1.5B/M) and covers zero on 12, so the bracket contains zero everywhere except on Q1.5B/M, where the margin is negative (Table 49). Where the comparator has some budget the trained token still buys nothing resolvable beyond it, but the budget is small, so this too is a statement of resolution. How large a margin could the runs have seen? The naive estimate is biased down by at most the bracket width, so an 80%-power one-sided detection on a whole-trace m=8 run needs a true mean margin of about that width plus 2.8 standard errors of the naive estimate: 0.0052–0.0163 per position on the 12 RLVR runs (0.0018–0.0041 of it noise alone), or 2.2–4.4 times the mean realised advantage on the 4 cells where that is resolved above zero. A trained token that out-bought best-of-e^{ε_t} search by less than a few times its own average credit would not have been seen.

The wasted-rate share within one draw. Keeping only the first m' outcomes of every action in each run with $m \geq 32$ holds the positions fixed, so the trend is the resolution effect alone. The share of rate spent where the bias-corrected dispersion estimate is not positive falls from m'=8 to the largest m' on 13 of 16 runs, from 38–81% to 46–65% at m'=32 and 41–50% at 64 on the 3 runs that go that far; it does not fall on QM7B/M (window, m=32), QM7B-Oat/G (whole trace, m=32), QM7B-SFT/G (window, m=32) (Table 50). What remains at the largest m' is mostly *clipping* (interior success rates whose unbiased dispersion estimate is not positive, 34–63% of the rate), and only 0–26% sits where every rollout of every measured action failed or succeeded (0–40% at m'=8). The $m \geq 32$ runs are separate draws on 24–100 problems, and a rate-weighted share is carried by few positions, so their m'=8 values need not match the m=8 runs': on the 182 positions QM7B-Oat/G's whole-trace m=32 draw shares with its m=8 run, one position carries 20% of the rate, and the two m=8 readings of the same positions give 84% and 33% (58% and 48% of positions, unweighted). The nested series isolates resolution; its values carry cluster-bootstrap intervals of ±7–21 points, and the m=8 whole-trace shares ±5–11 (`code/ledger_extensions.py`). The unambiguous part of the wasted rate is small at $m \geq 32$; the rest is rate spent where the base's value spread is below what the rollouts resolve.

Table 49: Search margin $\hat{A}_t - \hat{F}_t^{\mathrm{srch}}$ by rate tercile on the whole-trace $m{=}8$ runs (inverse-inclusion-weighted, cluster bootstrap over problems). $N \geq 2$: share of top-tercile positions whose comparator draws two or more tokens. $\bar{A}$: mean realised advantage in the tercile. Naive: plug-in, biased down; split (two-fold split-half) and cross-fit (leave-one-out) value each choice on held-out rollouts and are biased up (Appendix D); naive and split are point estimates. The cross-fitted $\hat{F}^{\mathrm{srch}}$ values each choice on a held-out rollout and can be negative, so the cross-fitted margin can exceed $\bar{A}$.

		mean $N = e^{\varepsilon_t}$ by rate tercile (median)			middle tercile	top tercile				
cell	n	low	middle	top	margin, cross-fit	$N \geq 2$	$\bar{A}$	margin, naive	split	margin, cross-fit
Q0.5B/G	1250	1.00	1.08	4.77 (2.0)	+0.0013 [−0.0001,+0.0027]	51%	−0.0047	−0.0392	−0.0063	−0.0086 [−0.0194,+0.0019]
Q0.5B/M	1000	1.00	1.03	7.67 (1.6)	+0.0003 [−0.0003,+0.0008]	38%	+0.0016	−0.0247	−0.0021	−0.0028 [−0.0141,+0.0075]
Q1.5B/G	1250	1.00	1.04	17.46 (1.8)	+0.0005 [−0.0001,+0.0013]	45%	+0.0047	−0.0301	+0.0011	−0.0005 [−0.0098,+0.0086]
Q1.5B/M	1000	1.00	1.02	4.18 (1.6)	+0.0004 [−0.0002,+0.0009]	36%	−0.0002	−0.0264	−0.0080	−0.0088 [−0.0181,−0.0000]
QM7B/G	1000	1.00	1.01	399.34 (1.4)	+0.0001 [−0.0001,+0.0002]	31%	+0.0080	−0.0189	−0.0002	−0.0002 [−0.0090,+0.0076]
QM7B/M	500	1.00	1.01	2.42 (1.5)	+0.0010 [+0.0001,+0.0022]	29%	+0.0104	−0.0128	+0.0125	+0.0096 [−0.0029,+0.0217]
QM7B-Oat/G	750	1.00	1.01	2.51 (1.3)	+0.0000 [−0.0003,+0.0003]	20%	+0.0084	−0.0104	+0.0066	+0.0089 [+0.0004,+0.0169]
QM7B-SFT/G	500	1.03	2.02	279.53 (16.3)	−0.0149 [−0.0309,+0.0006]	100%	+0.0117	−0.1005	−0.0045	−0.0136 [−0.0414,+0.0125]
L8B/G	1000	1.02	1.26	10.51 (3.3)	+0.0028 [−0.0003,+0.0058]	87%	+0.0098	−0.0650	−0.0049	−0.0095 [−0.0243,+0.0042]
Q7B/G	500	1.00	1.02	6.14 (1.5)	+0.0009 [+0.0002,+0.0017]	30%	+0.0044	−0.0318	−0.0034	+0.0061 [−0.0117,+0.0226]
Q7B/M	400	1.00	1.00	1.60 (1.2)	+0.0001 [−0.0000,+0.0003]	20%	+0.0091	−0.0079	+0.0088	+0.0084 [+0.0008,+0.0158]
Q7B-ORZ/G	500	1.00	1.03	19.42 (1.9)	+0.0006 [−0.0002,+0.0014]	48%	+0.0133	−0.0151	+0.0008	+0.0030 [−0.0056,+0.0120]
Q7B-ORZ/M	400	1.00	1.01	3.60 (1.6)	−0.0000 [−0.0005,+0.0005]	43%	−0.0023	−0.0163	+0.0006	+0.0006 [−0.0101,+0.0105]
QM7B-rand/G	500	1.00	1.00	1.82 (1.3)	−0.0000 [−0.0002,+0.0001]	23%	+0.0091	−0.0162	+0.0030	−0.0015 [−0.0121,+0.0100]
QM7B-gt/G	500	1.00	1.00	1.83 (1.3)	+0.0000 [−0.0001,+0.0002]	15%	+0.0014	−0.0161	−0.0008	+0.0020 [−0.0084,+0.0127]

Table 50: Share of the total rate spent at zero estimated dispersion, on nested subsamples of one draw; at the largest m', split into dead (every rollout failed), saturated (every rollout succeeded) and clipped (interior success rate, estimate ≤ 0).

		share of rate at zero dispersion (%), first m' rollouts					at the largest m': of which		
run	n	$m'{=}4$	8	16	32	64	dead	saturated	clipped
Q0.5B/G (whole trace, $m{=}32$)	500	73	75	68	60	–	14	2	43
Q0.5B/G (window, $m{=}32$)	300	74	63	62	60	–	12	4	45
Q0.5B/G (window, $m{=}64$)	200	64	61	56	55	43	7	3	34
Q0.5B/M (window, $m{=}32$)	250	80	79	69	65	–	23	3	38
Q1.5B/G (whole trace, $m{=}32$)	300	70	71	57	51	–	6	3	42
Q1.5B/G (window, $m{=}32$)	160	85	81	71	62	–	1	10	51
Q1.5B/G (window, $m{=}64$)	200	78	67	60	62	50	3	2	45
Q1.5B/M (window, $m{=}32$)	250	75	71	62	55	–	14	1	40
QM7B/G (whole trace, $m{=}32$)	200	69	60	52	48	–	0	0	48
QM7B/G (window, $m{=}32$)	160	78	64	60	63	–	0	0	63
QM7B/M (window, $m{=}32$)	160	66	50	57	53	–	4	0	49
QM7B-Oat/G (whole trace, $m{=}32$)	200	44	38	63	64	–	8	0	57
QM7B-SFT/G (window, $m{=}32$)	120	67	56	63	63	–	0	0	63
L8B/G (whole trace, $m{=}32$)	300	74	73	60	61	–	6	1	55
L8B/G (window, $m{=}32$)	250	62	62	56	55	–	7	1	47
L8B/G (window, $m{=}64$)	200	70	58	48	46	41	3	2	35

Specificity: a probability-matched null, and resampling. The two nulls of Table 26 differ in how probable the pseudo-realised token is under the base: the exchangeable draw is usually the trained token itself, the alternatives-only draw never is and is less probable. A third null draws uniformly among the alternatives whose base probability is within a factor of two of the trained token's, or takes the nearest in log-probability when none is (only 7–34% of positions have one within a factor of two at $m{=}8$, so this is mostly the nearest alternative, which can still be much less probable: the null only partly removes the probability confound). The p-values of Table 26 are also conditional: they compare the observed coefficient with null draws on the *same* rollouts, which is a randomisation test for the exchangeable null only. Here each null is fixed at 20 draws per position and the difference between the trained token's coefficient and the null tokens' mean coefficient is bootstrapped over problems (1000 replicates; Table 51). At $m{=}8$ the alternatives-only excess is positive on all 8 main cells of the two 7B bases (+0.04 to +0.16) and its interval excludes zero on 6 of them, the probability-matched excess (+0.04 to +0.17) on 3; on the 5 main cells of the other bases both are ≤ 0 (-0.07 to -0.01 matched), resolved on 1 (Q1.5B/M, where the trained token is resolvably *dis*favoured). Counts in this paragraph are of 95% intervals excluding zero; Table 54's uncorrected counts are one-sided tests and are larger, by up to four runs at $m{=}32$–64. The two reward controls split: the correct-reward policy shows the 7B pattern, the random-reward one does not (Table 51). At $m{=}32$–64, with fewer problems per run, the matched excess (-0.09 to +0.17) is resolved on 1 of 16 runs (the Llama window run at $m{=}64$), and the exchangeable null (the one comparison that is a valid randomisation test) on 3 by its randomisation p, 0 after Holm correction (3 and 1 by the bootstrap). Under Holm correction the matched null keeps 1 run at $m{=}8$ and 0 at $m{=}32$–64 (Table 54, where every family is listed). Against a less probable alternative the trained token is favoured on the 7B bases, resolvably on 6 of

8 cells once problems are resampled (1 of 15 runs after Holm, three of the six intervals starting at zero) but on only 3 when the alternative is matched in probability, and not on the other bases at m=8: specificity stays open, now for a reason one can quantify.

Table 51: Trained-token specificity under three nulls: the observed Spearman of $\hat{\sigma}_b^{\mathrm{alt}}$ with the value term, and its excess over the exchangeable-token, alternatives-only and probability-matched nulls (each null fixed at 20 draws per position), with a paired cluster bootstrap over problems (1000 replicates); bold intervals exclude zero. p_{MC}: the Monte-Carlo p of the matched null over 500 null coefficients, each from one draw per position, conditional on the sampled rollouts, shown only to contrast with the bootstrap (it is not a test of the population statement). Last column: share of positions with an alternative within a factor of two of the trained token's base probability.

				observed minus null, paired cluster bootstrap [95%]			matched	within
run	m	n	observed	exchangeable	alternatives only	probability-matched	p_{MC}	2× (%)
Q0.5B/G (whole trace)	8	1250	+0.45	−0.008 [−0.025,+0.009]	−0.034 [−0.090,+0.020]	−0.039 [−0.088,+0.014]	1.000	16
Q0.5B/M (whole trace)	8	1000	+0.50	+0.005 [−0.010,+0.020]	−0.004 [−0.062,+0.052]	−0.006 [−0.067,+0.053]	0.968	13
Q1.5B/G (whole trace)	8	1250	+0.39	−0.008 [−0.023,+0.007]	−0.040 [−0.092,+0.009]	−0.032 [−0.082,+0.020]	1.000	16
Q1.5B/M (whole trace)	8	1000	+0.43	−0.003 [−0.017,+0.010]	**−0.063 [−0.127,−0.004]**	**−0.070 [−0.131,−0.005]**	1.000	12
QM7B/G (whole trace)	8	1000	+0.41	+0.007 [−0.005,+0.019]	**+0.090 [+0.028,+0.154]**	**+0.080 [+0.018,+0.145]**	0.002	9
QM7B/M (whole trace)	8	500	+0.40	−0.001 [−0.014,+0.011]	+0.050 [−0.045,+0.156]	+0.054 [−0.042,+0.160]	0.002	10
QM7B-Oat/G (whole trace)	8	750	+0.28	+0.002 [−0.011,+0.015]	+0.037 [−0.046,+0.110]	+0.045 [−0.042,+0.126]	0.002	9
QM7B-SFT/G (whole trace)	8	500	+0.26	+0.028 [−0.035,+0.085]	**+0.106 [+0.004,+0.207]**	+0.084 [−0.021,+0.193]	0.002	34
L8B/G (whole trace)	8	1000	+0.28	+0.004 [−0.020,+0.026]	−0.009 [−0.065,+0.052]	−0.006 [−0.070,+0.064]	0.898	24
Q7B/G (whole trace)	8	500	+0.32	+0.006 [−0.013,+0.024]	**+0.100 [+0.001,+0.205]**	+0.091 [−0.018,+0.198]	0.002	10
Q7B/M (whole trace)	8	400	+0.41	+0.008 [−0.010,+0.024]	**+0.127 [+0.000,+0.251]**	+0.135 [−0.003,+0.268]	0.002	8
Q7B-ORZ/G (whole trace)	8	500	+0.56	+0.004 [−0.017,+0.027]	**+0.106 [+0.027,+0.190]**	**+0.106 [+0.015,+0.194]**	0.002	12
Q7B-ORZ/M (whole trace)	8	400	+0.56	+0.004 [−0.012,+0.022]	**+0.158 [+0.058,+0.268]**	**+0.167 [+0.060,+0.274]**	0.002	12
QM7B-rand/G (whole trace)	8	500	+0.40	+0.001 [−0.021,+0.020]	−0.016 [−0.100,+0.068]	−0.006 [−0.088,+0.081]	0.980	8
QM7B-gt/G (whole trace)	8	500	+0.34	+0.000 [−0.016,+0.016]	+0.096 [−0.001,+0.188]	**+0.109 [+0.012,+0.208]**	0.002	7
Q0.5B/G (whole trace, m=32)	32	500	+0.53	+0.013 [−0.019,+0.039]	+0.079 [−0.008,+0.171]	+0.072 [−0.015,+0.159]	0.002	16
Q1.5B/G (whole trace, m=32)	32	300	+0.42	−0.008 [−0.034,+0.015]	+0.015 [−0.090,+0.119]	+0.015 [−0.100,+0.134]	0.002	16
QM7B/G (whole trace, m=32)	32	200	+0.54	+0.009 [−0.021,+0.038]	+0.121 [−0.019,+0.257]	+0.098 [−0.061,+0.235]	0.002	10
QM7B-Oat/G (whole trace, m=32)	32	200	+0.40	**+0.023 [+0.007,+0.043]**	+0.131 [−0.012,+0.275]	+0.128 [−0.022,+0.300]	0.002	10
L8B/G (whole trace, m=32)	32	300	+0.40	+0.000 [−0.057,+0.053]	+0.088 [−0.037,+0.209]	+0.108 [−0.034,+0.241]	0.002	27
Q0.5B/G (window, m=32)	32	300	+0.46	**+0.028 [+0.000,+0.063]**	+0.040 [−0.067,+0.161]	+0.029 [−0.089,+0.149]	0.002	12
Q0.5B/M (window, m=32)	32	250	+0.58	−0.010 [−0.049,+0.025]	−0.002 [−0.114,+0.114]	+0.012 [−0.107,+0.137]	0.002	13
Q1.5B/G (window, m=32)	32	160	+0.42	**+0.050 [+0.017,+0.083]**	+0.194 [−0.031,+0.395]	+0.173 [−0.048,+0.368]	0.002	11
Q1.5B/M (window, m=32)	32	250	+0.50	−0.011 [−0.042,+0.023]	+0.030 [−0.082,+0.139]	+0.024 [−0.096,+0.138]	0.002	13
QM7B/G (window, m=32)	32	160	+0.27	−0.020 [−0.064,+0.020]	+0.015 [−0.169,+0.195]	+0.035 [−0.164,+0.232]	0.002	14
QM7B/M (window, m=32)	32	160	+0.28	−0.003 [−0.037,+0.039]	−0.070 [−0.204,+0.067]	−0.092 [−0.247,+0.069]	1.000	10
QM7B-SFT/G (window, m=32)	32	120	+0.00	−0.043 [−0.159,+0.056]	−0.060 [−0.290,+0.137]	−0.003 [−0.294,+0.266]	0.665	38
L8B/G (window, m=32)	32	250	+0.40	+0.006 [−0.039,+0.045]	+0.089 [−0.047,+0.221]	+0.130 [−0.019,+0.257]	0.002	26
Q0.5B/G (window, m=64)	64	200	+0.50	+0.029 [−0.010,+0.063]	+0.032 [−0.085,+0.154]	+0.002 [−0.110,+0.123]	0.531	12
Q1.5B/G (window, m=64)	64	200	+0.38	+0.000 [−0.051,+0.047]	+0.048 [−0.106,+0.215]	+0.050 [−0.095,+0.201]	0.002	16
L8B/G (window, m=64)	64	200	+0.46	+0.067 [−0.004,+0.133]	**+0.165 [+0.041,+0.288]**	**+0.151 [+0.016,+0.282]**	0.002	29

The literature's own statistics, and the Qwen2.5-7B cells. Table 23 has the two-policy divergences, but two statistics the token-level literature uses are not among them: the trained policy's own entropy $H(\pi_r)$, whose high-entropy tokens are singled out as forking points (Wang et al., 2025a), and the cross-entropy $\mathrm{CE}(\pi_r, \pi_b) = H(\pi_r) + \mathrm{KL}(\pi_r \| \pi_b)$, a disagreement statistic (Shen et al., 2026); we add both, and the policy's surprisal of the token it wrote, computed from the same top-20 lists. On the window positions of the 9 cells they correlate with $|\hat{A}_t|$ at $r_{\mathrm{s}} = 0.32$–0.80, at most 0.01 above the rate's coefficient in the same cell and up to 0.18 below it, and with the value term at $|r_{\mathrm{s}}| \leq 0.13$ (Table 52). The same statistics at the whole-trace ledger positions of the 4 `Qwen2.5-7B` cells (1800 positions; both policies' top-20 lists recorded there) give the same picture for worth: every one has $|r_{\mathrm{s}}| \leq 0.25$ with the value term, every one of those correlations negative, while the base's dispersion has r_{s} of +0.32 to +0.56 with it (Table 22). For room they split by pair, whose recipe and prompt both differ: on the two SimpleRL-Zoo cells every statistic tracks $|\hat{A}_t|$ ($r_{\mathrm{s}} = 0.58$–0.75), on the two Open-Reasoner-Zero cells, under that model's own prompt, only weakly ($r_{\mathrm{s}} = 0.14$–0.26), as the rate itself does there (0.18–0.27; Table 22). On the 10 RLVR window runs re-scored at m=32–64 the two-policy divergences reach $|r_{\mathrm{s}}| \leq 0.21$ against the value term and every statistic $|r_{\mathrm{s}}| \leq 0.26$, while dispersion has 0.27–0.58 (Table 53; `DIVX_HR=1 code/divergence_extended.py`). There, on the primary cell, every statistic's coefficient is negative and resolved, the inverse association of Section 4.3; with $|\hat{A}_t|$ the partial coefficient of KL given the rate reaches 0.24 (Table 23), so on some cells KL carries a little of $|A_t|$ beyond the room the rate fixes.

Reliability and weighting of the rank correlations. Two features of the measurement could hide a predictor of worth. The value term is itself an m=8 (or 32–64) estimate: its split-half reliability (Spearman–Brown) is 0.36–0.90 on the RLVR runs (Table 33), and dividing each rate or entropy coefficient by its square root raises the largest to 0.35 (QM7B/G, m=32), so correcting for the value term's noise leaves them weak (`code/worth_analysis.py`). Put through the same shared-

Table 52: Spearman correlations (over positions, unweighted) of the rate, three two-policy statistics (KL, JSD, the cross-entropy, which contains the KL) and two policy-side ones ($H(\pi_r)$, the policy's surprisal) with $|\hat{A}_t|$ and with the value term $|\hat{Q}(s_t, y_t) - \hat{V}_b^{\mathrm{alt}}|$, on the window runs and at the whole-trace ledger positions.

		against $\lvert\hat{A}_t\rvert$						against the value term					
run	n	rate	$\mathrm{KL}(\pi_r\Vert\pi_b)$	JSD	$H(\pi_r)$	$\mathrm{CE}(\pi_r,\pi_b)$	$-\log\pi_r(y_t)$	rate	$\mathrm{KL}(\pi_r\Vert\pi_b)$	JSD	$H(\pi_r)$	$\mathrm{CE}(\pi_r,\pi_b)$	$-\log\pi_r(y_t)$
Q0.5B/G (window)	700	+0.43	+0.40	+0.40	+0.32	+0.41	+0.32	−0.04	−0.04	−0.04	−0.12	−0.06	−0.12
Q0.5B/M (window)	500	+0.44	+0.40	+0.39	+0.37	+0.42	+0.37	+0.04	+0.02	+0.02	−0.00	+0.02	+0.00
Q1.5B/G (window)	700	+0.55	+0.52	+0.52	+0.50	+0.55	+0.50	−0.07	−0.08	−0.08	−0.07	−0.07	−0.07
Q1.5B/M (window)	500	+0.45	+0.46	+0.46	+0.39	+0.44	+0.38	−0.04	−0.02	−0.02	−0.06	−0.05	−0.06
QM7B/G (window)	600	+0.75	+0.60	+0.63	+0.69	+0.69	+0.70	−0.10	−0.09	−0.09	−0.12	−0.12	−0.13
QM7B/M (window)	400	+0.73	+0.65	+0.67	+0.60	+0.68	+0.59	+0.00	+0.05	+0.06	−0.03	−0.01	−0.05
QM7B-Oat/G (window)	600	+0.80	+0.80	+0.80	+0.66	+0.80	+0.62	−0.07	−0.06	−0.06	−0.03	−0.07	−0.05
QM7B-SFT/G (window)	300	+0.69	+0.68	+0.67	+0.59	+0.67	+0.57	−0.08	−0.04	−0.03	−0.07	−0.06	−0.08
L8B/G (window)	500	+0.61	+0.60	+0.60	+0.43	+0.61	+0.43	+0.06	+0.06	+0.06	+0.01	+0.06	+0.01
Q7B/G (whole trace)	500	+0.76	+0.74	+0.74	+0.72	+0.75	+0.72	−0.13	−0.12	−0.12	−0.14	−0.14	−0.14
Q7B/M (whole trace)	400	+0.61	+0.60	+0.60	+0.59	+0.60	+0.58	−0.10	−0.08	−0.08	−0.09	−0.10	−0.09
Q7B-ORZ/G (whole trace)	500	+0.18	+0.18	+0.18	+0.14	+0.18	+0.15	−0.19	−0.19	−0.19	−0.20	−0.19	−0.15
Q7B-ORZ/M (whole trace)	400	+0.27	+0.26	+0.26	+0.20	+0.26	+0.18	−0.24	−0.24	−0.24	−0.24	−0.25	−0.23

Table 53: Every statistic against the value term on the window runs re-scored at m=32 and 64: Spearman correlations over positions (unweighted), bold where the 95% cluster-bootstrap interval over problems (B=1000) excludes zero.

run (window)	m	n	rate	H_b	$\mathrm{KL}(\pi_r\Vert\pi_b)$	$\mathrm{KL}(\pi_b\Vert\pi_r)$	JSD	$\log\frac{\pi_r(y_t)}{\pi_b(y_t)}$	$H(\pi_r)$	CE	$-\log\pi_r(y_t)$	$\hat{\sigma}_b^{\mathrm{alt}}$
Q0.5B/G	32	300	−0.06	−0.05	−0.07	−0.02	−0.06	−0.04	**−0.17**	−0.09	**−0.17**	**+0.46**
Q0.5B/M	32	250	−0.03	−0.03	−0.03	−0.02	−0.02	+0.01	−0.03	−0.03	−0.02	**+0.58**
Q1.5B/G	32	160	−0.07	−0.07	−0.08	−0.04	−0.08	−0.01	**−0.15**	−0.08	**−0.14**	**+0.42**
Q1.5B/M	32	250	−0.10	−0.10	−0.07	−0.07	−0.07	−0.05	−0.10	−0.09	−0.10	**+0.50**
QM7B/G	32	160	**−0.25**	**−0.26**	**−0.20**	**−0.21**	**−0.21**	+0.13	**−0.24**	**−0.24**	**−0.24**	**+0.27**
QM7B/M	32	160	−0.04	−0.04	−0.01	+0.02	−0.01	+0.02	−0.04	−0.05	−0.05	**+0.28**
QM7B-SFT/G	32	120	+0.03	+0.03	+0.04	+0.03	+0.02	+0.03	+0.02	+0.04	−0.01	+0.00
L8B/G	32	250	+0.05	+0.04	+0.04	+0.09	+0.05	+0.07	−0.04	+0.05	−0.03	**+0.40**
Q0.5B/G	64	200	−0.08	−0.08	−0.08	−0.03	−0.07	−0.05	**−0.19**	−0.11	**−0.18**	**+0.50**
Q1.5B/G	64	200	−0.08	−0.08	−0.08	−0.07	−0.09	−0.05	−0.09	−0.06	−0.08	**+0.38**
L8B/G	64	200	+0.05	+0.01	+0.06	+0.10	+0.06	+0.07	−0.06	+0.05	−0.06	**+0.46**

rate null as dispersion (Table 25), the rate's coefficient with the value term has an excess over the null of -0.12 to -0.00 on the 13 main cells, resolved below zero (two-sided $p < 0.05$) on 12 and above it on 0, and the entropy's -0.12 to -0.02 (12 resolved below zero): the weak inverse association is not a scale artefact (`code/null_calibration_rate.py`). And the coefficients are computed over the rate-stratified sample; weighting every position by its inverse-inclusion weight (weighted mid-ranks, then a weighted Pearson correlation; `code/weighted_spearman.py`) moves $r_{\mathrm{s}}(|\hat{A}_t|, \varepsilon_t)$, $r_{\mathrm{s}}(\text{value term}, \varepsilon_t)$ and $r_{\mathrm{s}}(\text{value term}, \hat{\sigma}_b^{\mathrm{alt}})$ by at most 0.08 on the 22 RLVR runs at m=8 (window and whole trace, reward controls included; 0.13 on the distilled cell); weighted, the value term's largest coefficient with the rate is 0.25 in magnitude and its coefficient with dispersion 0.26–0.56.

Multiplicity. Unless marked "after Holm", every count of significant runs in the paper is one-sided and uncorrected, one test per run (the resampling counts of Appendix G are of 95% intervals). Table 54 gives the same counts under Holm's step-down correction within each family, with Monte-Carlo p-values taken conservatively as $(k+1)/(B+1)$. The shared-rate excess stays significant on 9 of 13 whole-trace runs and the resolution-series excess at m'=32 on 12 of 15; the scale-conditioned excess keeps 1 and 3 and the out-of-sample excess 3. With $B = 300$ the smallest attainable p, $1/301$, only just clears Holm's first threshold for 15 tests ($0.05/15$), so a null with fewer draws could not survive the correction whatever the data (the resolution series uses 400 null draws at every m', so its zero at m'=8, where the smallest p is $2/401$, is not structural). Correction is within families only; several families test overlapping hypotheses, so no paper-level error rate is claimed. Tables 25 and 32 print k/B (<0.003 for $k = 0$); the Holm counts use $(k+1)/(B+1)$.

Table 54: Significance counts with and without Holm correction (family-wise error 0.05 within each row). Null-draw and bootstrap p-values are $(k+1)/(B+1)$, with B the family's null draws or its 1000 bootstrap replicates.

family (one-sided tests, one per run)	tests	significant, uncorrected	Holm, FWER 0.05	smallest p
shared-rate null, excess (whole-trace, m=8)	13	12	9	0.0033
scale-conditioned (partial) excess, m=8	13	2	1	0.0033
scale-conditioned excess, m=32–64	15	6	3	0.0033
resolution series, excess at m'=8	15	8	0	0.0050
resolution series, excess at m'=16	15	9	7	0.0025
resolution series, excess at m'=32	15	13	12	0.0025
out-of-sample excess	14	9	3	0.0033
exchangeable-token null, m=8 (bootstrap)	15	0	0	0.1319
alternatives-only null, m=8 (bootstrap)	15	7	1	0.0020
probability-matched null, m=8 (bootstrap)	15	6	1	0.0020
exchangeable-token null, m=32–64 (bootstrap)	16	4	1	0.0010
alternatives-only null, m=32–64 (bootstrap)	16	5	0	0.0070
probability-matched null, m=32–64 (bootstrap)	16	3	0	0.0100