NOT WHICH PATH, BUT WHAT HAPPENS BEFORE YOU WOULD STOP: WHAT REASONING POST-TRAINING CHANGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training for reasoning makes language models much better at mathematics, yet the abilities it exploits already exist in the base model and its lasting changes touch few tokens, so what those changes do is unclear. We find that they change what happens before the model commits to an answer rather than which path it takes, and that a few words carry the change. Our instrument is a transplant in which one model writes and, wherever the two disagree about the next token, at 3 to 6% of positions, the other model's word is substituted. Given to the base model, the reasoning model's words close 83% of the accuracy gap, taken from the reasoning model they remove 68% of its advantage, and the same dose of foreign words at random positions removes 12%. The difference is not made at the steps that decide a run's outcome. Forcing each option there shows that both models mostly agree and neither chooses reliably better. Where post-training lengthens the run, the words act before the stop rather than at it. In the primary pair they recover the effect when the base model is left to end its run on its own, while the word at the stop alone, or a fixed "Wait", recovers nothing. The words carry most of the gap in three model pairs, one trained by reinforcement learning alone, and the effect is largest where the base model solves a problem occasionally and small where it has mastered it. Within the base model's reach, post-training thus changes how existing capabilities are deployed rather than what they are. That gives interpretability a concrete target, locates the gain that sampling alone does not recover, and says which problems, measured against the base model itself, post-training reaches.

1 INTRODUCTION

Post-training for reasoning, by reinforcement learning with verifiable rewards (RLVR) or by distilling from a model trained that way, turns a pretrained model into a reasoning model, and the accuracy it buys is large. Watch the two models on the same problem and the difference is easy to see and hard to name. The base model writes a short derivation, boxes a plausible answer and stops. The reasoning model writes ten times as much, doubts, checks, tries another route, and is right far more often. What did post-training put into it?

Three well-supported findings make that question harder than it looks. The ability was already there: sampled enough times a base model reaches most of the answers its reasoning counterpart reaches (Yue et al., 2025), sharpening its distribution at decode time closes much of the gap without any reinforcement learning (Karan & Du, 2026), and probes find the relevant behaviours in base-model activations before post-training, with training supplying a sense of when to use them (Venhoff et al., 2026). The visible product is cheap to remove: suppressing the "wait, let me check" restarts that RLVR so conspicuously installs leaves accuracy nearly unchanged (Wang et al., 2025a; Zulaika et al., 2026). The durable edits are few. RLVR's lasting changes concentrate on a small fraction of uncertain positions, one to three percent in Akgül et al. (2026), and restricting updates to the highest-entropy fifth of tokens loses nothing (Wang et al., 2025b). Letting a reasoning model choose single tokens for a weaker one recovers much of the difference (Shen et al., 2026a). Behind that last finding stands a wider expectation that a solution is decided at a few forking steps and that training

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.

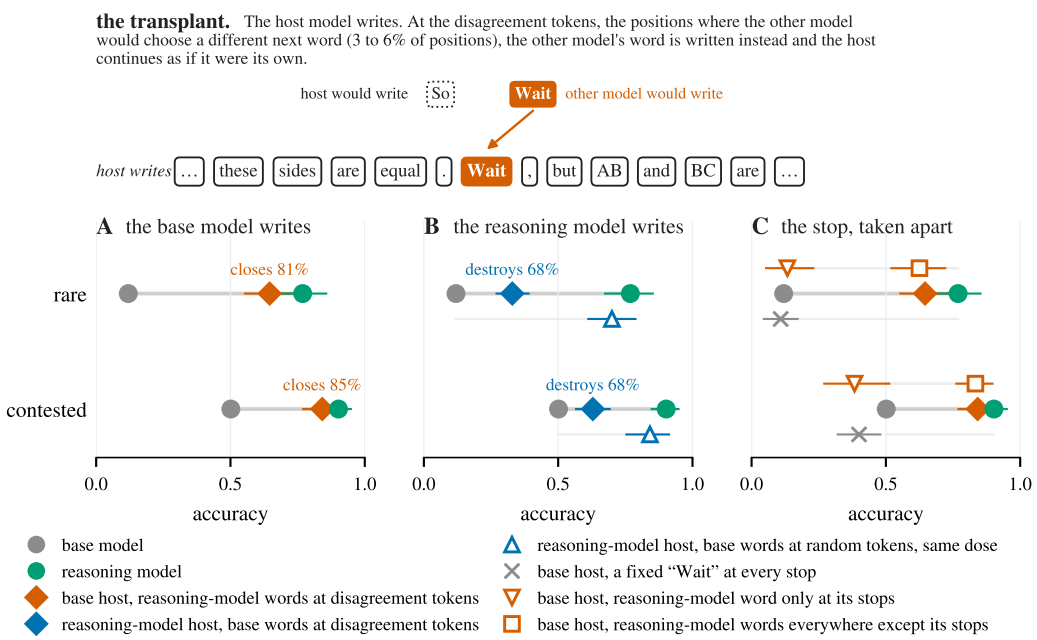

Figure 1: **The transplant and its main results, Qwen3-8B pair.** Top: one substitution in a transplanted run. Below, each row is a band of problems, rare (the base model solves up to a third) or contested (a third to two thirds), and runs from the base model's accuracy to the reasoning model's, with every condition and its interval. **A**: the base model writes with the reasoning model's words. **B**: the reverse, with the dose-matched random control. **C**: panel A taken apart at the stop.

acts there (Bigelow et al., 2025; Lin et al., 2025; Cui et al., 2025; Lightman et al., 2024), and its sharpest form reads reinforcement learning as sparse policy selection, committing to the right branch at a handful of uncertain positions (Akgül et al., 2026). A census of edits says *where* the edits are, not what they accomplish.

This paper asks what those few tokens buy. Our instrument is a transplant between a reasoning model and its base (Figure 1, Section 2). One model, the host, writes, and where the two most disagree about the next word the other model, the donor, supplies its word and the host continues as though it had written it. Run in each direction, and at random positions as a control, it asks whether the reasoning model's words are sufficient, whether they are necessary, and whether their position matters. We find that the words change what happens before the model commits to an answer, not which path it takes. Four findings carry that claim.

First, the words carry most of the gap in both directions (Section 3). Given to the base model they close 83% of the accuracy gap, and taken from the reasoning model they destroy 68% of its advantage. The same number of substitutions at random positions moves 4% of the gap, and the same dose of foreign words at random positions 12%.

Second, on the primary pair the positions that carry the gap are not the steps where a run's fate is decided (Section 4). At sampled positions we force each option the base model would consider, complete the run and grade it, which finds the positions where a run's fate is open. There the two models mostly agree and neither picks the winning option reliably better. The positions where the two models disagree lie largely elsewhere.

Third, what the words carry is what the run does before it commits (Section 5). The base model writes a plausible answer and stops, and the reasoning model does not. Replacing the words with the base model's own collapses its runs from a median of 8,264 tokens to 1,537, against 8,157 at random positions. Adding them takes the base model from 816 tokens to 8,528, the reasoning model's own length. Telling the base model not to stop is not enough, and neither is the word at the stop, while the same words everywhere except at the stop close as much of the gap as the full transplant. What the words carry is the work before the stop, after which the base model mostly

stops by itself. When the reasoning model's end-of-thinking token is never supplied, the words still close .69 and .70 of the gap on the rare and contested bands, against .78 and .83.

Fourth, the effect is graded by the base model's own solve rate (Section 3). It is largest where the base model solves a problem occasionally and small where it is already reliable, and on problems it never solves the words lift it from .00 to .05 against .15 for the reasoning model.

2 METHODOLOGY

In this section we present the techniques and setup used for performing the experiments of the paper. The main instrument of the paper is the transplant. One model, the host, writes a solution token by token. At every position both models' next-token distributions are computed on the host's prefix, and where they diverge beyond a threshold the donor's sampled token replaces the host's, after which the host continues as though it had written it. Nothing else changes: the same prompt, problems, decoding parameters, budgets and grader in every condition.

Positions. Let p_t^{R} and p_t^{B} be the reasoning and base models' next-token distributions on the host's prefix at position t. The donor's token replaces the host's wherever the symmetrised divergence

$$D_t = \tfrac{1}{2}\,\mathrm{KL}\big(p_t^{\mathrm{R}} \,\|\, p_t^{\mathrm{B}}\big) + \tfrac{1}{2}\,\mathrm{KL}\big(p_t^{\mathrm{B}} \,\|\, p_t^{\mathrm{R}}\big) \tag{1}$$

exceeds a threshold τ. We call the rule the *gate* and such a position a *disagreement position*, or gate position. The threshold is the 0.94 quantile of D_t on a stream of 60 reasoning-model solutions recorded before the core set existed, 31 of which later entered it. It is one scalar, set on divergence alone with no outcome in view and never re-tuned. The *substitution rate* of a condition, its dose, is the fraction of generated tokens the donor supplied, and for the gate it is 3 to 6% in every Qwen condition and source but the stop-free arm of Section 3, where it is 2 to 3%. A word is *off-support* when it lies outside the host's sampling nucleus (temperature 0.7, top-p 0.95). The two models end with different tokens, so the gate fires at essentially every stop the host would take, and the run ends there only if the donor's token is its own end token. The stop set holds both models' end tokens, so a donor end token supplied at the gate ends the run (Section 3). The intervention is that of Shen et al. (2026a), reimplemented.

Necessity and sufficiency. On a set of problems let B and R be the base and reasoning models' accuracies, T that of the base model writing with the reasoning model's words at the gate, and G that of the reasoning model writing with the base model's words there. The words are *sufficient* to the extent that they close the gap when given, $(T-B)/(R-B)$, and *necessary* to the extent that they destroy the edge when taken, $(R-G)/(R-B)$, both fractions of the same gap, so that 1 means the words carry all of it and 0 none.

Forks and basins. Consider a base-model prefix $x_{<t}$ and its two most likely next tokens a_t and b_t. Let $\hat{s}_t(o)$ be the success rate of n forced completions of option o from that prefix, and let ν_t be the noise floor, the mean absolute difference between two disjoint halves of the completions of the same option. Position t is a *fork* if

$$\big|\hat{s}_t(a_t) - \hat{s}_t(b_t)\big| > \nu_t, \tag{2}$$

that is, if choosing one option over the other changes the chance of solving the problem by more than resampling noise, and the size of that change is the fork's swing. Forks matter because a better chooser would show its advantage there, the first thing the paper tests, and they turn out to be largely distinct from the gate positions (Section 4). The *basin* of an answer y at t is the set of continuations of the prefix that end in y, and its mass $q_t(y)$ is the probability, under the model's own decoding, of ending there. A run has settled into that basin once $q_t(y)$ is near one, and it *commits* at the first position where it writes the boxed answer it will keep (Section 5).

Models. The transplant needs a pair that differ only by post-training. Qwen3-8B and Qwen3-8B-Base are the primary pair. Throughout, the member of a pair after reasoning post-training is the *reasoning model* and its partner the *base model*. The 8B is distilled from the flagship Qwen3 models, which are the ones that go through the reinforcement-learning pipeline (Qwen Team, 2025). Olmo-3-7B-Think and Olmo-3-1025-7B are a second family with different pretraining data and a recipe

Table 1: **The problem set by source.** Bands are measured against the base model and cut across the benchmark labels, each cell giving the count and its share of the source's problems. Cap rate is the share of base-model reference samples that hit the budget, and the last column their median length.

source	problems	budget	cap rate	unreached	rare	contested	reliable	mastered	median base run (tokens)
GSM8K	30	3,072	0.001	2 (7%)	3 (10%)	1 (3%)	13 (43%)	11 (37%)	241
MATH-500	151	12,288	0.002	18 (12%)	14 (9%)	33 (22%)	40 (26%)	46 (30%)	465
AMC23	29	16,384	0.005	3 (10%)	8 (28%)	11 (38%)	5 (17%)	2 (7%)	828
AIME pool	90	24,576	0.019	37 (41%)	35 (39%)	15 (17%)	2 (2%)	1 (1%)	916
all	300		0.008	60 (20%)	60 (20%)	60 (20%)	60 (20%)	60 (20%)	650

of supervised fine-tuning, preference optimisation and RLVR (Team Olmo, 2025), and SimpleRL-Zoo-7B with Qwen2.5-7B a third, trained by RLVR alone (Zeng et al., 2025). Within a pair every run uses one prompt for both models, so a transplant never changes the prefix: the reasoning model's chat template for the Qwen pairs, and a four-shot prompt for Olmo, whose base model needs worked examples before it answers in a form the grader can read (its reasoning model's reference alone uses its own chat template).

Problems. Difficulty is measured against the base model, because the effect depends on its competence. All experiments run on one set of 300 problems drawn from GSM8K, MATH-500, AMC23 and an AIME pool (Table 1). We measure difficulty by each problem's base-model solve rate over 8 samples, extended to 32 where undecided and to 64 for problems never solved. Next, we place each of them in one of five bands of 60 problems: *unreached* (never solved), *rare* (up to one third), *contested* (one to two thirds), *reliable* (two thirds to almost always) and *mastered* (always). A problem enters a band only if two estimators agree, one counting truncated runs as wrong and one ignoring them so budget-dependent problems are excluded. Every tunable setting, the census of Section 4 included, was fixed on a separate set of 36 problems drawn before the core set and never used in any reported number, except the gate threshold τ, set earlier on the MATH-500 stream. The bands stay fixed for every condition, and a split-half check on them moves the headline by 0.01.

Budgets and grading. Token budgets are set per source (Table 1), in two tiers, so that truncation is rare for terminating runs. The base model reaches the ceiling on 0.8% of its reference samples, while the reasoning model reaches it on 19.8% of rare-band and 9.2% of contested-band runs, so its column and the transplant's are scoped to the budget to that extent. One grader, extracting the last boxed expression, is used for every condition, and where an intervention changes termination we report two conventions (Section 3). Every interval in the paper is a 95% bootstrap interval over problems. Appendices A–D hold the account, the problem set, the protocol and stop analyses with every interval, and the census and sampling details.

3 TRANSPLANTING THE DISAGREEMENT WORDS

This section runs the transplant in both directions on the Qwen pair (Table 2), against two controls that put the same words at random positions.

The two directions tell one story. On the rare and contested bands, where both conditions ran (the matched bands), corrupting the reasoning model's disagreement positions destroys 0.68 of its edge over the base model, and supplying those positions to the base model builds 0.83 of it.

We run a random control to rule out that any substitution would do the same. By construction the gate fires where the models diverge, so most of the words it substitutes are ones the host would not have written itself, *off-support*, while words substituted at random positions mostly are not. The control therefore substitutes only off-support words, at random positions, at the gate's own off-support rate. It destroys 0.12 of the edge against the gate's 0.68, and in the forward direction it closes .12 of the gap against .81. A looser random control at a higher nominal rate but far less off-support dose does even less (Table 2), so position, not dose, carries the effect. Running the reasoning model through the transplant code with nothing substituted costs about .04 on rare problems (Appendix C).

Table 2 also shows that the effect depends on the base model's solve rate. On the rare and contested bands the words carry the gap in both directions. On the reliable and mastered bands there is little or

Table 2: **The transplant in both directions, Qwen3-8B pair.** Bands of 60 problems, and k is the number of rollouts per problem. Left: the base model writes with reasoning-model words at the gate ($k = 4$, one rollout per problem on the unreached band). The second column counts only runs that write a boxed answer, a fifth of the rare band being silent, and the reasoning model scored that way reaches .932 and .982. Right: the reasoning model writes with base-model words at the gate ($k = 8$), at random positions at a comparable rate ($k = 4$), or at the gate's off-support dose ($k = 2$). Fractions as defined in Section 2, undefined where the gap is near zero. Intervals are given for the fractions, the paper's claims, and for every accuracy in Appendix C.

		base model writes		reasoning model writes				fraction of the gap	
band	base	+reasoning words	runs that answer	+base words	random	random, dose-matched	reasoning	sufficiency	necessity
unreached	.000	.050	.250	.069	.062	—	.146	.34 [.00, .80]	.53 [.21, .74]
rare	.120	.646	.803	.329	.729	.700	.769	.81 [.71, .92]	.68 [.59, .76]
contested	.501	.842	.944	.629	.904	.842	.902	.85 [.72, .95]	.68 [.54, .83]
reliable	.895	—	—	.869	.942	—	.935	—	—
mastered	1.000	—	—	.946	.979	—	.979	—	—

Table 3: **The transplant on two further pairs**, on the same 120 matched problems. Olmo-3-1025-7B and Olmo-3-7B-Think, forward direction, at one word in five ($k = 4$) and at Qwen's 3.4% ($k = 2$). Qwen2.5-7B and SimpleRL-Zoo-7B, both directions and the random-position control ($k = 2$). Fractions as in Table 2, omitted where the band's gap is too narrow to divide, then the median run lengths of the two reference models.

				base model writes			reasoning model writes		median tokens	
pair	band	base	reasoning	+reasoning words	closes	random	+base words	removes	base	reasoning
Olmo-3	rare	.017	.738	.571	.77 [.68, .86]	—	—	—	231	11,017
one word in five	contested	.062	.892	.588	.63 [.53, .73]	—	—	—	198	6,985
Olmo-3	rare	.017	.738	.442	.59 [.47, .71]	—	—	—	231	11,017
at Qwen's rate	contested	.062	.892	.350	.35 [.24, .46]	—	—	—	198	6,985
SimpleRL-Zoo	rare	.054	.142	.142	—	.092	.067	—	926	1,108
	contested	.208	.596	.608	1.03 [.80, 1.32]	.242	.233	.94 [.69, 1.22]	801	974

no gap to carry and gate words cost a few points, the floor every substitution pays. On the unreached band gate words and random words damage the reasoning model alike, and given to the base model they lift it only from .00 to .05.

The accuracy numbers conceal a change in the shape of the runs. With base-model words at its disagreement positions the reasoning model's runs on the rare and contested bands collapse to a fifth of their length (Figure 4). It stops checking itself and the runs that still answer are wrong half the time, while the same words at random positions leave its checking intact and its accuracy near its own. The collapse runs through the base model's stop. The gate fires wherever the base model would end the run, and 94% of these runs end on its end token, against 62% under dose-matched random words, where it arrives later and accuracy holds at .77. The loss of accuracy does not need the stop. In a stop-free arm, which never substitutes the base model's end token or `</think>`, the base model's words still remove .92 of the edge. The runs are then as long as the reasoning model's own at the median, 46% run to the ceiling against 13% with nothing substituted, and 98% still give an answer (Appendix C). Given the reasoning model's words, the base model runs to the reasoning model's own length and starts checking itself (Section 5).

What the gate selects is the vocabulary of deliberation, in plain words rather than special tokens. Words such as *wait*, *but*, *let me check* and *alternatively* make up 18% of gate tokens against 3% elsewhere, and the words logged in the forward transplant match this audit. Features active just before a *wait* are known to modulate the reasoning that follows (Troitskii et al., 2025).

Two further model pairs. Table 3 repeats the transplant on two further pairs. On the Olmo pair the substitution rate is one word in five, four times Qwen's, and at that rate the reasoning model's words close .70 of the matched gap, with the transplanted host ending its own runs 83% of the time.

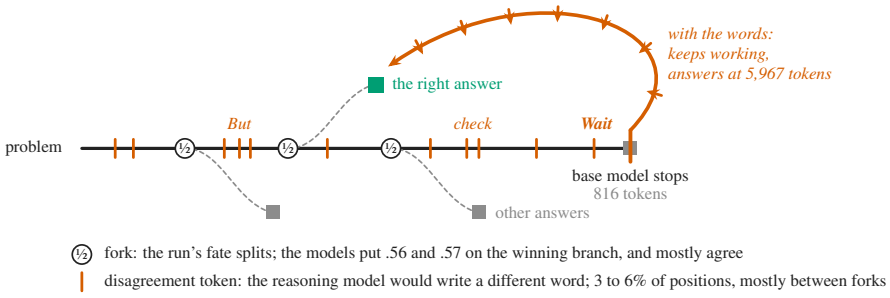

Figure 2: **The two kinds of position.** A run leaves the problem, passes the forks where its fate splits, and settles in an answer; the disagreement tokens lie mostly between the forks and at the stop.

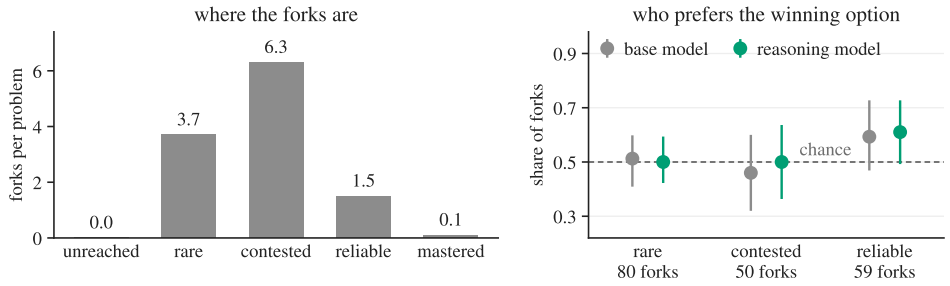

Figure 3: **The census, Qwen pair.** Left: forks per problem after the noise-floor correction. Right: how often each model puts more probability on the winning option, with 95% intervals.

At Qwen's rate the same words close .46, so the replication is a dose-response rather than a single point. Because the bands are Qwen's, Olmo's base model solves these problems at only .017 and .062. SimpleRL-Zoo is the pair trained by RLVR alone, with no distillation, a case neither of the first two can answer. There the words carry the contested gap in both directions, and the same words at random positions carry almost none of it. Its rare gap is too narrow to divide, and its gains are not confined to its training set (.28 on MATH-500, .32 on AMC23, .19 on AIME). What does not replicate in the third pair is the run length. Its two models write runs of almost the same length, so the words carry the gap there without moving the stop, which Section 5 takes up.

4 Do the models choose differently at the forks?

If the words move the outcome, one hypothesis is that they win the decisions, i.e. a solution is settled at a few forking or critical tokens (Bigelow et al., 2025; Lin et al., 2025). Furthermore, post-training acts there, which is how the sparse-position accounts read their own results (Cui et al., 2025; Shen et al., 2026a; Akgül et al., 2026). We tested it directly with a *census* of the forks. At sampled positions along a base-model trace we force each of the two most likely options, complete it to an answer and grade it, six completions per option, and call the position a fork when Equation 2 holds. The resampling method is that of Bogdan et al. (2025), to which we add the two-model comparison and the noise floor. The census forces only the base model's own top two options, on base-model prefixes, so an advantage the reasoning model expresses through a token the base model would never consider, or only on prefixes it would itself write, is outside what it can see.

Figure 3 gives the result on the Qwen pair, the only one the census ran on, and Table 6 in the appendix the numbers. Forks exist, and they concentrate where the effect does. They are nearly absent at both ends of the range, where everything works or nothing does, and cluster on the rare and contested bands, four to six per problem. The stakes at a fork are high, with a mean swing of .46 between the better and the worse option. Yet neither model chooses better. Consider the 85 candidate forks, among the 585 scored under both models, where the choice moves the success rate

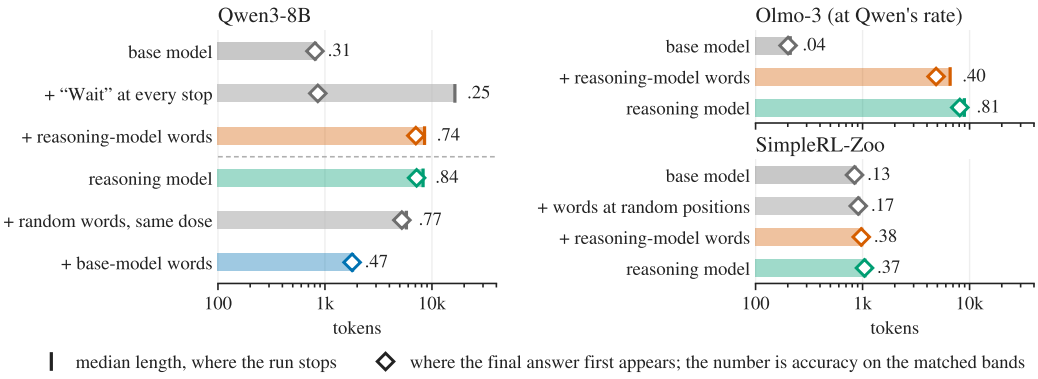

Figure 4: **Where the run stops and where it commits, on the matched bands of each pair.** Each bar runs to the condition's median length; the diamond marks where the run's final answer first appears; the number is accuracy. For the reference models the diamond is read from one saved run per problem of the same model. In the Qwen panel the two rows below the dashed line are the reasoning model's runs with base-model words at the gate (from the second seed, whose text was saved) and with random words at the same dose.

by more than half. Each model puts a little over half its probability on the winning option, the two differ by a point, and they pick the same side three times in four. Neither is distinguishable from chance on its own, so we read the result as a bound. Even the largest per-fork advantage the interval allows, taken at every fork, would move accuracy by at most a fifth of the rare-band gap and half of the contested one. The gap is not made at the forks. Section 6 returns to what the forks hold.

Words at disagreement positions carry the gap, yet at the forks neither model is better. Both can hold only if the two sets of positions differ, and they do. At a fork the reasoning model keeps almost all of its probability inside the base model's own two options, a median 94%. At the positions the gate selects, the word it supplies falls outside the base model's sampling nucleus 61% of the time, though usually within its top five (Appendix C). The two sets barely overlap: nine in ten forks are not gate positions, the transplant's gain on a problem does not track how many forks it has, and forks thin out along a run while disagreement does not. The gate selects a kind of moment, not a region (Figure 2).

5 Run length, stopping and commitment

If the words are not better choices at the forks, what do they change? The most visible difference is where the two models stop. The base model stops early, with a plausible but wrong answer, and not because the budget cuts it off. Cutting every base-model reference run at four thousand tokens costs it nothing in any band, while the reasoning model keeps converting tokens into accuracy up to the ceiling (Figure 5). The transplant shows that the stop travels with the words (Figure 4). With the reasoning model's words the base host runs to the reasoning model's own length, and 81% of its runs end at a stop those words supply, and only one in 480 at a stop of its own. With the base model's words the reasoning model stops at a fifth of its length, on the base model's own stop (Section 3). Where the run ends is set by the words, in both directions.

Since the gate fires at every stop the host would take, the transplant could work simply by keeping the base model going. Three conditions, all on the base host, take that apart (Figure 1C). First, a fixed "Wait" at every stop, budget forcing (Muennighoff et al., 2025) with no donor anywhere, sends nearly every run to the ceiling and leaves accuracy at or below the base model's own. Second, the reasoning model's word at the stops only also keeps the runs going and buys nothing. Third, the reasoning model's words at every disagreement position except the stops, with the base model left to end its own run, close .78 of the rare gap and .83 of the contested one, as much as the full transplant, and most of those runs end at a stop the base model chooses itself. The spared positions are those where the host means to end its run. In 69% of these runs the reasoning model's end-of-thinking token still closes the deliberation, and those runs carry the effect (Appendix C). With that token refused throughout, which avoids selecting runs by length, the words still close .69 of the rare gap

and .70 of the contested one. Neither more tokens nor the right word at the end carries the effect. The words carry work done along the way, and once that work is done the base model mostly stops by itself.

The third pair says how far this account travels. SimpleRL-Zoo and its base write runs of the same length (Table 3), so there is no stop to move. The transplanted run ends at the host's own stop nine times in ten, and base-model words leave the reasoning model's length where it was. The words still carry the gap there, and what they carry is visible in the commitment. Given the reasoning model's words the base host writes one answer and keeps it in every run, whereas the same words at random positions leave it writing two and changing its mind in a fifth of the runs. Base-model words give the reasoning host that habit in turn, all at unchanged length. In runs of a thousand tokens the position of the commitment cannot be read, so only the claim that the host commits once can be tested there, and it holds. The claim about stopping concerns the two long-chain pairs.

The same result shows from the answer's side. Scalena et al. (2026) find a commitment boundary inside reasoning traces, after which the final answer no longer changes, and show that the chain after it is epiphenomenal. Our saved runs show what moves that boundary (Figure 4). Read as the point where the final boxed answer first appears, the base model commits at the very end of its short run and the reasoning model 87% of the way into its long one. Given the reasoning model's words, the base host commits 83% of the way in, once, and its first written answer is almost always its final one. The words do not make the host revise an early answer. They withhold it until it holds, and end the run once it does. The words move the boundary, and what follows is the decoration that Zulaika et al. (2026) and Scalena et al. (2026) remove at negligible cost.

6 What sampling and selection recover

A cheaper alternative to the reasoning model's words is to sample the base model more, and what sampling recovers says what the words are worth (Figure 7 in Appendix D). On contested problems a majority vote over eight base runs reaches the reasoning model's accuracy at the same token cost, because the right answer is the plurality of the base model's attempts there. On rare problems it is not, because the right answer is a minority among some twenty distinct answers per problem. Moreover, the vote saturates at a third of the reasoning model's accuracy and even a perfect selector over sixteen draws only matches the reasoning model's single run. Sampling fails exactly where the transplanted words are worth most.

Selection is where post-training shows again. Over the same base runs, picking by either model's likelihood does no better than the vote and the reasoning model's own likelihood is the worst rule of all. Asked instead to read each run and say whether it is correct, the reasoning model picks nearly twice what the vote does on rare problems and recovers three quarters of what a perfect selector would. This comparison tests whether post-training installed a recogniser or only a sharpened distribution, and it is the former, in line with the generation-verification asymmetry others report (Song et al., 2025; Chen et al., 2026). The forks hold value too and neither model collects it. A one-step lookahead over the two options at the census's positions doubles the base model's accuracy on rare problems, while a linear probe on the prefix cannot read the winning side (Appendix D). That value is real, and the sparse-token literature works on it, but it is not what post-training moved.

7 Related work

Base models know. Base models reach most of their reasoning counterparts' answers given enough samples (Yue et al., 2025), post-training amplifies behaviours pretraining already produces (Zhao et al., 2025), decode-time sharpening closes much of the gap without reinforcement learning (Karan & Du, 2026; Guo et al., 2026), and RLVR improves the reasoning behind the answers more than the answers themselves (Wen et al., 2026). The behaviours post-training seems to install are probeable in base activations beforehand, training supplying the sense of when to use them (Venhoff et al., 2026; Ward et al., 2025a), and gains appear only where the base model already succeeds sometimes (Zhang et al., 2026), as expected if RL ascends a transform of the success probability (Davis & Recht, 2025). We give that *when* a measured form and a carrier.

Edits are sparse. Concurrent work places RLVR's durable changes on few high-entropy positions and recovers much of the gain by correcting single tokens (Akgül et al., 2026; Wang et al., 2025b; Cui et al., 2025; Shen et al., 2026a; Huang et al., 2026a), and on the parameter side the update lives in a small subnetwork (Mukherjee et al., 2025), a rank-one adapter or one bias vector per layer recovering most of the gain (Ward et al., 2025b; Sinii et al., 2025). Earlier work located the forking and critical tokens (Bigelow et al., 2025; Lin et al., 2025), and Nguyen et al. (2026) trace the coverage reasoning models lose to forks in their training data. Shen et al. (2026a) let the stronger model take over at high-disagreement positions, and Akgül et al. (2026) substitute the RL model's greedy choice at every disagreement and read the result as sparse policy selection. The intervention is theirs, run here under sampling; what is new is asking it in both directions against a control that moves the same words to random positions at a matched dose, testing at the forks whether either model chooses better, and taking the effect apart at the stop. We confirm the sparsity and qualify the address: the gate finds uncertain positions, mostly not the most uncertain ones.

The visible might be decorative. Suppressing the installed restarts barely moves accuracy (Wang et al., 2025a; Zulaika et al., 2026), and past a commitment boundary the rest of a chain is epiphenomenal (Scalena et al., 2026; Boppana et al., 2026). Our transplant joins the two: the few words that matter act before the boundary, and the text after it can be decorative.

Test-time compute. Budget forcing extends a run with a cue (Muennighoff et al., 2025). Reasoning models solve a problem in rounds, the first usually right (Chen et al., 2025), their intermediate answers become a stopping signal (Fu et al., 2025), and their hidden states know when one is right (Zhang et al., 2025). With thinking suppressed they recover their accuracy from parallel samples at matched cost (Ma et al., 2025).

8 Conclusions and limitations

The picture that emerges is narrower than "post-training teaches reasoning" and we believe is also more actionable. On problems within the base model's reach, the failure is rarely a wrong turn at a hard step, where both models are near chance and mostly agree. The behaviour that hurts the base model is stopping too early with a plausible answer, and forbidding the stop does not cure it. What post-training adds on these problems is what the model does before it would stop, carried by a few words at identifiable positions, which is why it can be transplanted.

Difficulty measured against the base model itself separates these readings and much of the disagreement about what post-training adds looks like a mixture of bands, which Noukhovitch et al. (2026) see from the training side as gains in proportion to initial competence, Huang et al. (2026b) as an implicit curriculum at the edge of competence, and Shen et al. (2026b) as amplification on easy problems and surfacing on hard ones. We consider one reading fits all of this, the words carry a sequential search that never needs the winner in advance, and post-training taught the model not to choose but to keep going until an answer holds. We present that as a reading, not a result (Appendix A).

Several limits follow. Our models are 8B-scale, our problems are mathematics, and our claims are scoped to problems within the base model's reach, which prolonged or curriculum RL can extend (Liu et al., 2025; Cai et al., 2026). A sharper risk is family-specific. Much of the evidence that base models already contain these behaviours is Qwen-derived, other base families do not all show them (Gandhi et al., 2025), and spurious rewards give large gains on Qwen (Shao et al., 2026). Neither long-chain reasoning model is pure RLVR (Qwen3-8B is distilled, Olmo-3-Think adds supervised and preference stages), so our claims concern the installed behaviour. The fork census and the stop decomposition ran on the Qwen pair only, and the third pair, whose models run the same length, says nothing about stopping. A second seed reproduces both headline results. The gate supplies stop tokens along with words, and withholding them leaves both results standing (Sections 3 and 5). The bands are Qwen's, reused for the other pairs, and 3 to 6% of tokens is several hundred words in a long run. The unreached band was run once per problem, so its cells are lower bounds.

REPRODUCIBILITY STATEMENT

The problem set with its band assignments, every rollout of every condition (each row records its host and donor models, and for the two earliest conditions the analysis scripts establish both from the run configuration and a length fingerprint), the reference samples, and the analysis scripts that print each table (with the sanity assertions that failed at some point during the campaign left in place) are released with the paper at `https://anonymous.4open.science/r/not-which-path-842D`. Appendix B gives the construction of the problem set, Appendix C the grader, hardware and seeds, the gate and the stop analyses. The stop analyses of Appendix C are printed by `analyze_revision_checks.py`, and the two masking arms were launched by `run_e47_masking.sh`.

ETHICS STATEMENT

This work studies the mechanics of mathematical reasoning in open-weight models on public benchmarks. It involves no human subjects, no personal data and no deployed system, and we see no direct ethical risk beyond those of the underlying models.

AI USE STATEMENT

In this work we used generative AI tools for the following tasks that require disclosure: to implement the methods (the experiment, analysis and figure code), to propose experiments and controls and give feedback on them, from which the authors chose, and to help develop the account of Appendix A, formulate its propositions and draft their proofs. We did not use them to generate the research question, to propose or refine hypotheses, to interpret results, or to produce or alter any reported measurement, and the remaining tasks that require disclosure (synthetic data, translation, dataset cleaning, qualitative data analysis) do not apply to this work. We also used them to launch and monitor the experimental campaign and to draft and edit the text and the figures. We have reviewed all AI-assisted work. Every reported number was checked by the authors against the raw outputs through the scripts named in the reproducibility statement, AI-written code was read and tested by the authors before any reported run, the statements and proofs of Appendix A were checked by the authors line by line. The claims, scope statements and limitations are the authors' own, and we take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

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.

Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking paths in neural text generation. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=8RCmNLeeXx`.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which LLM reasoning steps matter? In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL `https://openreview.net/forum?id=VnSlfeRCaU`.

Siddharth Boppana, Annabel Ma, Max Loeffler, Raphaël Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. Reasoning theater: Disentangling model beliefs from chain-of-thought. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL `https://openreview.net/forum?id=1lyagkzogH`.

Pengxiang Cai, Tianchen Fang, Xiaohan Li, Qingyuan Zeng, Guocong Li, and Jintai Chen. Curriculum reinforcement learning can incentivize reasoning capacity in LLMs beyond the base model. *arXiv preprint arXiv:2606.22317*, 2026.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? On the overthinking of long reasoning models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9487–9499, 2025.

Yuxin Chen, Yu Wang, Yi Zhang, Ziang Ye, Zhengzhou Cai, Yaorui Shi, Qi Gu, Hui Su, Xunliang Cai, Xiang Wang, An Zhang, and Tat-Seng Chua. Learning to self-verify makes language models better reasoners. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://openreview.net/forum?id=rSDj9EM1HV.

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.

Damek Davis and Benjamin Recht. What is the objective of reasoning with reinforcement learning? *arXiv preprint arXiv:2510.13651*, 2025.

Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Yonghao Zhuang, Yian Ma, Aurick Qiao, Tajana Rosing, Ion Stoica, and Hao Zhang. Efficiently scaling LLM reasoning programs with Certaindex. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=nn51ewu5k2.

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 *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=QGJ9ttXLTy.

Hong Guo, Nianhui Guo, Christoph Meinel, and Haojin Yang. Sample where you struggle: Sharpening base model reasoning via entropy-guided power sampling. *arXiv preprint arXiv:2606.09926*, 2026.

Kexin Huang, Haoming Meng, Junkang Wu, Jinda Lu, Chiyu Ma, Ziqian Chen, Xue Wang, Bolin Ding, Jiancan Wu, Xiang Wang, Xiangnan He, Guoyin Wang, and Jingren Zhou. Beyond magnitude: Leveraging direction of RLVR updates for LLM reasoning. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=r6Pw3RiMYL.

Yu Huang, Zixin Wen, Yuejie Chi, Yuting Wei, Aarti Singh, Yingbin Liang, and Yuxin Chen. On the emergence of implicit curriculum in RLVR learning dynamics. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026b. URL https://openreview.net/forum?id=KxYCE98u1d.

Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Vsgq2ldr4K.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, 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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37906–37918, 2025.

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*, volume 38, 2025. URL https://openreview.net/forum?id=YPsJha5HXQ.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*, 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, 2025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tür, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=0NdS4xCngO.

Ngoc-Hieu Nguyen, Parshin Shojaee, Phuc Minh Nguyen, Nan Zhang, Chandan K. Reddy, Khoa D. Doan, and Rui Zhang. Why do reasoning models lose coverage? The role of data and forks in the road. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=ivza8JIW9s.

Michael Noukhovitch, Hamish Ivison, Nathan Lambert, and Aaron Courville. Learning to solve hard problems in RL for LLMs by never giving up. *arXiv preprint arXiv:2609.13443*, 2026.

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

Daniel Scalena, Sara Candussio, Luca Bortolussi, Elisabetta Fersini, Malvina Nissim, and Gabriele Sarti. Beyond the commitment boundary: Probing epiphenomenal chain-of-thought in large reasoning models. *arXiv preprint arXiv:2606.13603*, 2026.

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 *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://openreview.net/forum?id=tqTNOpkP5j.

Changshuo Shen, Leheng Sheng, Yuxin Chen, An Zhang, and Xiang Wang. Reasoning can be restored by correcting a few decision tokens. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026a. URL https://arxiv.org/abs/2605.16874.

Jingyan Shen, Ang Li, Salman Rahman, Yifan Sun, Micah Goldblum, Matus Telgarsky, and Pavel Izmailov. Understanding reasoning from pretraining to post-training. *arXiv preprint arXiv:2607.16097*, 2026b.

Viacheslav Sinii, Alexey Gorbatovski, Artem Cherepanov, Boris Shaposhnikov, Nikita Balagansky, and Daniil Gavrilov. Steering LLM reasoning through bias-only adaptation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9202–9211, 2025. URL https://aclanthology.org/2025.emnlp-main.467/.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham M. Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mtJSMcF3ek.

Team Olmo. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Dmitrii Troitskii, Koyena Pal, Chris Wendler, and Callum Stuart McDougall. Internal states before wait modulate reasoning patterns. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18640–18649, 2025. URL https://aclanthology.org/2025.findings-emnlp.1012/.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Base models know how to reason, thinking models learn when. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://openreview.net/forum?id=2BniakOS4q.

Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. Wait, we don't need to "Wait"! Removing thinking tokens improves reasoning efficiency. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7459–7482, 2025a. URL https://aclanthology.org/2025.findings-emnlp.394/.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui 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*, volume 38, 2025b. URL https://openreview.net/forum?id=yfcpdY4gMP.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. Reasoning-finetuning repurposes latent representations in base models. In *ICML 2025 Workshop on Actionable Interpretability*, 2025a. URL https://arxiv.org/abs/2507.12638.

Jake Ward, Paul Riechers, and Adam Shai. Rank-1 LoRAs encode interpretable reasoning signals. *arXiv preprint arXiv:2511.06739*, 2025b.

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*, 2026. URL https://openreview.net/forum?id=jGbRWwIidy.

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*, volume 38, 2025. URL https://openreview.net/forum?id=4OsgYD7em5.

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 *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=vSMCBUgrQj.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=O6I0Av7683.

Charlie Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and RL on reasoning language models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://openreview.net/forum?id=TBaUfO9znF.

Rosie Zhao, Alexandru Meterez, Sham M. Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. Echo chamber: RL post-training amplifies behaviors learned in pretraining. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=dp4KWuSDzj.

Unai Zulaika, Miguel Fernandez-de-Retana, Rubén Sánchez-Corcuera, Aritz Bilbao-Jayo, and Aitor Almeida. Backtracking is decorative: A mechanistic account. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=IUX6hpNZBd.

A THE ACCOUNT AND TWO PROPOSITIONS

The Conclusions offer a reading of the results: post-training moves accuracy through a few positions where a word changes the run's fate, and everything else it visibly changes is style. This appendix states that reading as two propositions about any sequential sampler, not about transformers, and says which premises the measurements cover. The propositions hold for any process with the stated properties. The census measures how rare forks are and how much they swing, not how flat fate is away from them, and nothing here shows that our models have those properties for the reasons the account gives.

Setup. A problem fixes a grader. A policy π continues a prefix h (the problem and a partial solution) by sampling tokens until an end token, and the finished text is graded right or wrong. The *fate* of a prefix is its probability of ending right when π finishes the run,

$$\varphi^{\pi}(h) = \Pr_{\tau \sim \pi(\cdot|h)}\big[\tau \text{ ends right}\big],$$

so fate depends on the policy, and we drop the superscript where the policy is clear. The census of Section 4 measures the base model's fate φ^{B}. It forces a next token y, lets the base model finish, grades the result, and averages over completions to estimate $\varphi^{\mathrm{B}}(h \oplus y)$.

Assumption 1 (absorption)**.** A prefix h is ϵ-*absorbed* if every token the sampler would consider leaves fate unchanged up to ϵ, that is $|\varphi(h \oplus y) - \varphi(h \oplus y')| \le \epsilon$ for all y, y' in the union of the two models' sampling supports. It is a δ-*fork* if two such tokens differ in fate by at least δ. Every position is one or the other, with $\epsilon < \delta$.

Assumption 2 (sparse forks)**.** Along typical runs at most a fraction ρ of positions are δ-forks. The largest difference in fate at a fork is $\delta_{\max} \le 1$.

Equation 2 is the measured form of the fork definition, with the noise floor in place of δ. Table 6 gives ρ (3.7 and 6.3 forks per problem on the rare and contested bands, against runs of thousands of tokens) and the typical swing at a fork (.46). Nothing measures ϵ. The census resolves a difference in fate only down to its noise floor, near .3 with six completions per option when fate is near one half, while the conclusions below need ϵ small against $\rho\,\delta_{\max}$. Absorption is therefore assumed, not measured.

The census also differs from the account in three ways, and each can hide from it the forks the reasoning model's words use. It forces the base model's own two most likely options, whereas the account's forks range over every token either model would consider, and 61% of the words the transplant supplies lie outside the base model's nucleus (Section 4). It measures φ^{B}, while the fate that matters for the reasoning model is its own φ^{R}. A position where one option loses under base-model completions but the reasoning model recovers from both is a fork for the census and absorbed under φ^{R}, and there the reasoning model has no reason to prefer a side. It also scores prefixes the base model writes, while Lemma 1 weights the prefixes of the modified run. That is how, in Section 4, the reasoning model's words carry the gap while neither model chooses better at the forks the census finds.

Both propositions rest on one lemma, the performance-difference identity written for fates.

Lemma 1 (fate difference)**.** *Let π' be any modification of π whose sampling support stays within the union of the two models' supports, let TV_t be the total-variation change of the next-token distribution at step t of a run, let F_t be 1 if the prefix at step t is a fork and 0 if it is absorbed, and let S_t be the spread of φ^{π} over the tokens whose mass moved. With expectations over runs of π' of typical length T,*

$$\big|\varphi^{\pi'}(h_0) - \varphi^{\pi}(h_0)\big| \le \sum_{t<T} \mathbb{E}\big[\mathrm{TV}_t\, S_t\big] \le \delta_{\max} \sum_{t<T} \mathbb{E}\big[\mathrm{TV}_t\, F_t\big] + \epsilon \sum_{t<T} \mathbb{E}\big[\mathrm{TV}_t\,(1 - F_t)\big].$$

Proof. Replace π by π' one step at a time along the run and telescope. Following π' up to step t and π after it, exchanging the distribution at step t changes the value by $\sum_y (\pi' - \pi)(y \mid h_t)\, \varphi^{\pi}(h_t \oplus y)$ with h_t drawn from π', which is at most $\mathrm{TV}_t\, S_t$ because the fates enter only through their spread over the mass that moved. Absorption (Assumption 1) bounds S_t by ϵ off forks, and $S_t \le \delta_{\max}$ on forks by definition. Splitting the sum on F_t gives the second inequality. With $\overline{\mathrm{TV}}$ the mean of TV_t,

it reduces to $T\,\overline{\mathrm{TV}}\,(\rho\,\delta_{\max} + (1-\rho)\,\epsilon)$ only when the push is independent of fork status, which a targeted intervention need not be. □

When ϵ is small against $\rho\,\delta_{\max}$, an intervention that pushes evenly moves accuracy through one channel only, the $\delta_{\max}$ term at forks. The same total push is capped by the ϵ term when placed at absorbed positions and by the $\delta_{\max}$ term when placed on forks, so the ceilings on the effect of equal doses differ by a factor $\delta_{\max}/\epsilon$. That is the shape of the dose-matched control of Section 3, where the same words at the same dose destroy .12 of the edge at random positions and .68 at the gate's.

Proposition 1 (learning lands at the forks)**.** *Train π_θ with an outcome reward $R \in \{0,1\}$ and any state baseline b, and take fates under π_θ. At a prefix h that is ϵ-absorbed over the support π_θ samples from, the expected policy gradient of the step taken at h satisfies*

$$\big\|\mathbb{E}\big[(R-b)\,\nabla_\theta \log \pi_\theta(y \mid h)\big]\big\| \le \epsilon\,\mathbb{E}\big\|\nabla_\theta \log \pi_\theta(y \mid h)\big\|.$$

At a two-option fork with $\pi_\theta(y^+ \mid h) = p$, $\pi_\theta(y^- \mid h) = 1-p$ and fates differing by $\Delta = \varphi(h \oplus y^+) - \varphi(h \oplus y^-) \ge \delta$, the expected gradient equals $p(1-p)\,\Delta\,\big[\nabla_\theta \log \pi_\theta(y^+ \mid h) - \nabla_\theta \log \pi_\theta(y^- \mid h)\big]$.

Proof. Condition on h and take the baseline $b = \varphi(h)$. Any other baseline adds a mean-zero term. The reward's conditional mean after choosing y is $\varphi(h \oplus y)$, so the expectation over the next token is $\sum_y \pi(y \mid h)\big(\varphi(h \oplus y) - \varphi(h)\big)\nabla_\theta \log \pi_\theta(y \mid h)$. Since $\varphi(h)$ is the π_θ-weighted mean of the $\varphi(h \oplus y)$, absorption bounds every coefficient by ϵ in magnitude, which gives the bound. With two options of fates $\varphi^\pm = \varphi(h \oplus y^\pm)$ and $\Delta = \varphi^+ - \varphi^-$ the coefficients are $(1-p)\Delta$ and $-p\Delta$, which gives the identity. □

Corollary 1 (the style tax, informal)**.** *A sampled run with positive advantage multiplies every token's log-probability gradient by the same scalar (in the unclipped objective, since ratio clipping, whose bias Shao et al. (2026) show yields gains even under random rewards, makes it token-dependent). By Proposition 1 the systematic part of that update is $O(\epsilon)$ everywhere but at forks, and the remainder is zero-mean sampling noise of order one per position that finite groups do not cancel. Outcome-reward training therefore buys its accuracy at the ρ-fraction of forks. Everywhere else its updates are, to first order, the zero-mean noise of whichever runs happened to win, through which their style can drift into the policy without buying accuracy. We state this informally and rely on it nowhere.*

Corollary 1 is the training-side account the Conclusions refer to. It predicts that the durable changes of post-training concentrate on a small set of positions, which concurrent work reports (Akgül et al., 2026; Wang et al., 2025b; Shen et al., 2026a) and which the gate finds at 3 to 6% of tokens, and that the visible product elsewhere is removable at little cost, which Wang et al. (2025a) and Zulaika et al. (2026) report. It is also compatible with both findings of Section 4, the words carrying the gap and neither model choosing better at the census's forks. The gradient lands at the forks of the policy being trained, under its own fate and on its own prefixes, and these move away from the base model's forks as training changes both. Proposition 1 concerns outcome-reward training, so it applies to SimpleRL-Zoo and to the models Qwen3-8B is distilled from, not directly to the pair the census ran on.

Proposition 2 (behaviour-fitted parts are decorative)**.** *Consider an instrument that scores a candidate internal part by how well it predicts emitted behaviour, averaged over the positions a run visits, with a per-position contribution bounded by a constant. Two candidates that differ only in what they do at forks differ in score by $O(\rho)$, so the score separates a part that is causal for fate from one that is not only through an $O(\rho)$ term. If ablating a part found this way pushes the next-token distribution independently of fork status, Lemma 1 bounds its effect on fate by $T\,\overline{\mathrm{TV}}\,(\rho\,\delta_{\max} + \epsilon)$, however much it moves behaviour.*

Proof sketch. The score is an average over the position-visitation measure, on which forks carry mass at most ρ. Two candidates identical off forks differ by at most ρ times the bound on the per-position term, so the objective leaves their fork components undetermined. The fate bound is Lemma 1 applied to the ablation in its reduced form, which the independence assumption allows, with $\overline{\mathrm{TV}}$ the ablation's mean push per position. A statement with the visitation measure made explicit is straightforward and omitted. □

Proposition 2 predicts that parts selected by behavioural fit, a backtracking direction, a correctness probe, or features chosen on outputs, can be removed without accuracy loss, which is the pattern Zulaika et al. (2026) report. It does not predict our main result, which is about tokens rather than parts; that comes from Lemma 1 together with the measured location of the gate.

Scope. The propositions say where outcome training can put ability and why behaviour-first searches miss it. They do not say what computation happens at a fork, and they do not require that runs get longer. The third pair of Section 3, whose two models are the same length, is consistent with them, and the stopping account of Section 5 is an empirical description of the two long-chain pairs, not a consequence of anything here.

B THE PROBLEM SET

Every reported number runs on one set of 300 problems, so that results can be joined per problem instead of compared across benchmarks.

Difficulty pass. A pool of 503 candidate problems from GSM8K, MATH-500, AMC23 and an AIME pool was sampled eight times by the base model under the paper's prompt, then 32 more times wherever eight samples left the solve rate undecided. Each sample records whether it was right, whether it hit the token budget, and its length. Budgets are set per source (Table 1) and the pass runs in two tiers, a first at a quarter of the budget and a rerun at the full budget for every sample that hit the first, so that the final cap rate is .008 pooled and one problem in 503 changes band depending on how cut-off runs are counted.

Bands. Bands are cut on the measured base solve rate, with the definitions and the two-estimator rule of Section 2. The first cut gave 60, 38, 58, 60 and 60 problems in the five bands, and the code asserts that the core set and the calibration set do not intersect.

Sampling depth. Eight samples are too few for the two closed bands. Never solved in eight samples is weak evidence for a problem whose true rate is one in twenty, and always solved in eight is weak evidence of mastery. The unreached candidates were therefore resampled to 64 and the mastered candidates to 32. Of 115 problems with no correct answer in eight samples, 49 produced one by 64, at a median rate of .031, and the oracle coverage of that group rose from .027 at one sample to .150, .225, .326 and .426 at 8, 16, 32 and 64 with no plateau. Of 156 problems solved in all of eight samples, 89 failed at least once by 32. Both closed bands were then refilled to 60 with problems that held their band at the deeper count, giving the five bands of 60 in Table 1. The unreached band is therefore unreached within 64 samples at the stated budget, not out of reach. Because coverage keeps climbing with samples, as Yue et al. (2025) find, the comparison with sampling (Table 7) is stated at a fixed k.

Stability. A split-sample check re-estimates the base column from the even-numbered samples and re-derives each problem's band from the odd-numbered ones. The base column moves by at most .005 in any band, the necessity fraction on the matched bands moves from .679 to .673, and the half-sample bands agree with the frozen ones for 82%, 72% and 80% of rare, contested and reliable problems and for every problem in the two closed bands. Regression to the mean in the band cut is therefore small, and the paper's matched-band headline does not depend on it.

C PROTOCOL, GATE AND STOPS

Budgets. Token budgets are set per source (Table 1) so that truncation is rare for runs that end on their own. On the unreached band the reasoning model's cell and the transplant's are scoped to a 6,144-token budget by decision and are lower bounds. At that budget the reasoning model scores .146 and the base model with reasoning-model words .050 ([.00, .12], one rollout per problem). Figure 5 gives each reference model's accuracy within a budget, per band.

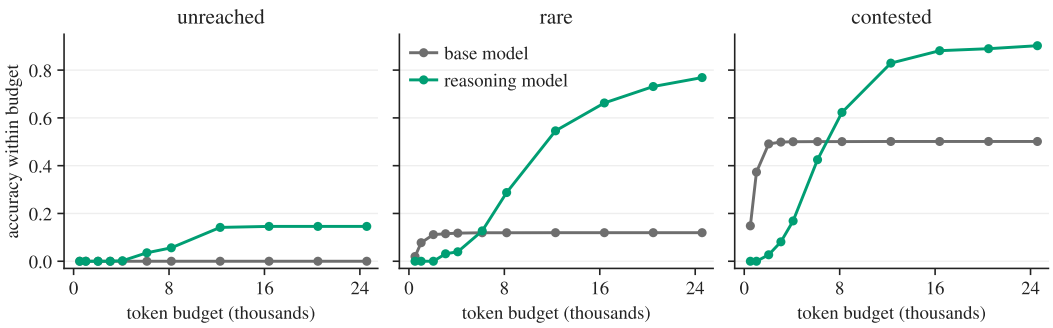

Figure 5: **Accuracy within a token budget, Qwen pair, per band.** Each reference run was sampled once at the full budget. The curve at budget B is the fraction of runs whose final answer is correct and already written within their first B tokens.

Hardware, seeds and the grader. Runs used two card types, Quadro RTX 8000 and RTX A6000, in half precision. The reference samples of 228 problems were drawn on the first and 72 on the second; within a source the two agree to $+.011$ on MATH-500 ($n = 283$) and $+.011$ on GSM8K ($n = 60$), and a permutation test stratified by source gives $p = .177$, so the card is not a hidden factor. Every condition uses generator seed 11 and the second seed of Section 8 uses 12. The grader extracts the last boxed expression, is the same function in every condition and both references, and is checked against eleven hand cases including nested braces; a run with no boxed expression counts as wrong. Every condition runs four to eight rollouts per problem, and every interval is a problem-level bootstrap over them. The second seed agrees with the first on the same 120 problems, with necessity .70 [.59, .82] against .68 [.60, .76] and sufficiency .79 [.69, .88] against .83 [.75, .90]. Figure 6 places every condition on every band, and Tables 4 and 5 give every cell and fraction with its interval.

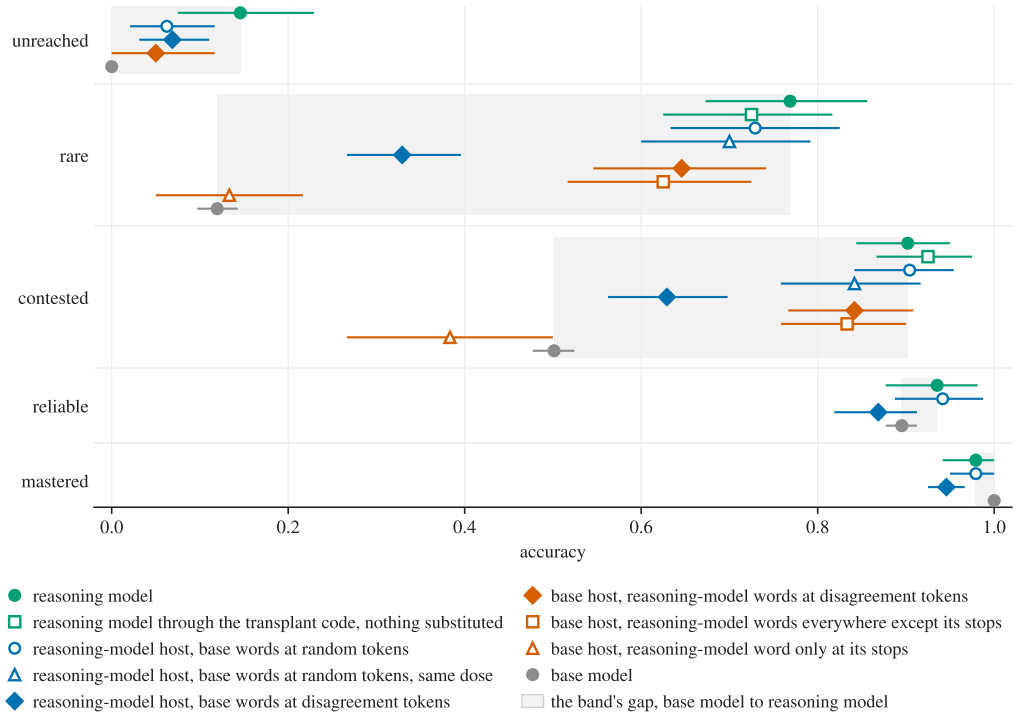

Figure 6: **Every condition on every band.** Each band's shaded span runs from the base model's accuracy to the reasoning model's, and each condition of Table 2 and each stop condition of Section 5 sits on its own line with its 95% problem-level bootstrap interval.

Table 4: **Every cell of Table 2 with its 95% problem-level bootstrap interval** (60 problems per band, the per-problem statistic being the mean over that problem's rollouts).

band	base	gate	random	reasoning	transplant
unreached	.000 [.000, .000]	.069 [.031, .113]	.062 [.021, .117]	.146 [.071, .229]	—
rare	.120 [.097, .144]	.329 [.267, .394]	.729 [.633, .821]	.769 [.673, .856]	.646 [.546, .742]
contested	.501 [.478, .525]	.629 [.560, .694]	.904 [.842, .954]	.902 [.844, .952]	.842 [.767, .912]
reliable	.895 [.877, .913]	.869 [.821, .912]	.942 [.883, .988]	.935 [.877, .983]	—
mastered	1.000 [1.000, 1.000]	.946 [.925, .965]	.979 [.950, 1.000]	.979 [.942, 1.000]	—

Table 5: **Sufficiency and necessity per band**, the fractions of Section 2, with 95% problem-level bootstrap intervals. Undefined where the band's gap is within .05 of zero.

band	gap $R - B$	sufficiency $(T - B)/(R - B)$	necessity $(R - G)/(R - B)$
unreached	$+.146$	.34 [.00, .80]	.53 [.21, .74]
rare	$+.649$	.81 [.71, .92]	.68 [.59, .76]
contested	$+.401$	.85 [.72, .95]	.68 [.54, .83]
reliable	$+.040$	—	—
mastered	$-.021$	—	—

The gate. The threshold predates the calibration set. It was set on a stream of 60 MATH-500 traces of the reasoning model, 31 of which later entered the core set, as the 0.94 quantile of divergence, with no outcome in view and never re-tuned. Its one structural consequence, that every stop is a gate position, does not depend on the threshold, since stops sit at 15 to 26 nats against a threshold of 2.2. The stream holds 89 of the reasoning model's end tokens and 56 of the base model's, and the gate fires at all of the former and 96% of the latter. By source the substitution rate is 3.1 to 4.3% with the reasoning model as host and 4.3 to 5.9% with the base model, slightly lower on harder problems, so a threshold set on MATH-500 transfers to the other sources. Pooled, the gated conditions fire at 3.4% of generated tokens with the reasoning model as host and 4.5% with the base model, and the nominal random control at 5.0%. On the calibration stream the 6% of positions with the highest host entropy and the 6% the gate selects overlap by 16% (Jaccard 0.09), though the two are rank-correlated at 0.80, and a typical gate position sits at the 76th entropy percentile.

Substituted words. On the stream the gate was calibrated on, the reasoning model's word at a gate position is the base model's first choice 18% of the time and in its top five 77% (median rank 3), against 88% and 99% at other positions. Its base-model probability has a median of .08 and is below .05 at two gate positions in five. Akgül et al. (2026) find the RL model's token in the base model's top five almost always (mean rank near 2) for short-chain models under greedy decoding. The long-chain pair's gate words sit further out, mostly words the base model ranks but would rarely sample, which is what the 61% off-support share measures. With the base model writing, its word at a gate position lies in the reasoning model's top five 91% of the time (median rank 2). Per run the gate makes 127 substitutions with the reasoning model as host and 446 with the base model, over mean lengths of 3,753 and 9,835 tokens, the medians of 1,537 and 8,528 being lower because lengths are skewed. With base words at the gate the reasoning model makes 2.8 self-checks per run against 17.4 under the nominal random condition, which leaves its runs at 8,157 tokens, and its answered runs are right .509 of the time against .920. The base host with reasoning-model words makes 49 self-checks per run against 18 for the reasoning model under the dose-matched random condition.

Controls. At the gate 67% of substitutions are off-support in the necessity direction and 61% in the sufficiency direction, against 7% at random positions, which per generated token is 2.3% off-support injection at the gate against 0.34% at random, 6.6 times less. The nominal-rate random control is therefore under-dosed. The dose-matched control tests a third of all positions and substitutes only where the base model's word is off-support, delivering 2.1% per token at random positions. It destroys 0.12 of the edge ([0.04, 0.21]) against the gate's 0.68, so the same dose of the same foreign words does five times the damage at the gate's positions (Section 3). In the forward direction, on the rare band at one rollout per problem, the nominal control closes .02 of the gap and the dose-matched

one .12, against .81 at the gate. The zero-substitution pass puts the reasoning model at .725 on rare and .925 on contested through the transplant code against .769 and .902 on its own.

Stops, necessity direction. The base model's words largely end the reasoning model's runs. On the necessity arm's second seed, whose substitutions were logged, 94% of the matched-band runs end on the injected base-model end token, at a median of 1,810 tokens against 8,351 for the reasoning model through the transplant code with nothing substituted. Under the dose-matched random control 62% of runs end on an injected base-model end token, later, at a median of 5,774 tokens, and accuracy holds at .771 against .825. No run in either arm has its thinking closed by an injected </think>. Keeping only the runs that do not end on the injected token leaves 13 problems, too few to read. The stop-free arm of Section 3 removes the stop instead. It never substitutes the base model's end token or </think>, and where the gate would supply either, the host keeps its own token, 4.2 times per run. With one run per matched-band problem, the base model's words then remove .93 [.78, 1.06] of the rare gap, .92 [.61, 1.24] of the contested one and .92 [.77, 1.07] of both, against .68 for the necessity arm on the same problems. The runs reach a median of 8,293 tokens against 8,351 with nothing substituted, and 46% of them run to the ceiling against 13%. Nearly all still give a boxed answer (98%), so the loss is in the answers, which are right .36 of the time against .51 in the necessity arm and .94 with nothing substituted. The gate substitutes 2.2% of their tokens (1.8 to 3.4% by source), 225 per run against 127 in the necessity arm and 164 in the dose-matched random control, which removes .12.

Stops, sufficiency direction. Budget forcing on its own, a bare "Wait" at every stop with no donor, fires 6.6 times per run, sends 97% of runs to the ceiling, and leaves the base model at or below its own accuracy under three gradings (-0.11 of the gap $[-0.21, -0.01]$ on the last boxed answer, -0.04 on the first, and in between on a majority over each run's boxed answers), while the transplant closes 0.83 on the same problems. The reasoning model's word only at the vetoed stops sends 58% of runs to the ceiling and scores .133 on rare and .383 on contested against the base model's .120 and .501. The words everywhere except the stops close .78 [.66, .89] of the rare gap and .83 [.69, .96] of the contested one against the full transplant's .81 and .85, and 82% of those runs end at a stop the base model chooses itself. In the gate-minus-stops arm 96% of substitutions come before the run's final answer first appears. The first tenth of that stretch holds 18% of them and each later tenth between 8 and 11%, with 4.9% in its last twentieth. Of 75,181 substitutions, 121 (0.16%) replace a token of a boxed answer, and the host never tries to close its own thinking at a gate position. But the reasoning model's </think>, which the base model never writes on its own, is supplied in 69% of these runs and in 84% of the full transplant's, and those runs carry the effect. In the gate-minus-stops arm they are right .903 of the time against .347 for the rest, 57% of which run to the ceiling without an answer (the answers they do give are right 93% of the time), and in the full transplant .876 against .039. The split selects on length, since runs whose donor has not closed its thinking by the ceiling are partly the long problems on which the reasoning model itself reaches the ceiling. The arm that refuses the donor's </think> throughout avoids that selection. With one run per problem against two, it closes .69 [.50, .89] of the rare gap and .70 [.44, .94] of the contested one, against .78 and .83 for the gate-minus-stops arm on the same problems, which is ahead by .10 [$-.03$, .24] over both bands. Its runs are longer, a median of 9,086 tokens against 7,190, and 32% reach the ceiling against 18%. Of its runs 78% give an answer against 80%, right .87 of the time against .91. The gate substitutes 3.9% of their tokens (3.3 to 5.7% by source), 417 per run against 421, and the donor's end token or </think> is refused 1.7 times per run.

Commitment. The base model with a bare "Wait" cue, whose runs never stop, commits at a median of 850 tokens, 5% of the way in, and then re-derives the same answer. The reasoning model on its own runs commits at 87% of its length (one saved run per matched problem, Figure 4), and read from the dose-matched random condition that leaves its runs nearly intact, once at a median of 4,018 tokens, 91%. The base model on its own runs commits at 99%, that is, it stops as soon as it has an answer. The base host with the reasoning model's words commits once at a median of 5,967 tokens, 83% of the way in, with its first written answer already its final one in 96% of runs. In the SimpleRL-Zoo pair the transplanted run ends at the host's own stop in 91% of cases and the donor's end token is never injected. With the reasoning model's words at the gate the base host writes 1.0 boxed answers per run with the first answer final in 100% of runs on both bands. The same words at random positions leave it writing 1.7 to 1.9 per run and changing its answer in 18 to 24% of runs,

and base-model words at the gate give the reasoning host 2.4 boxed answers per run with the first final in 72%. All conditions run at the same length, about a thousand tokens.

D FORKS, SAMPLING AND SELECTION

The census. Table 6 gives the census of Section 4 per band. With the same divergence stream and threshold the gate would fire on 9.7% of forks and 7.1% of the strongest ones, against 6.0% of ordinary positions, so most forks are not gate positions and most gate positions are not forks. Transplant gain shows no correlation with a problem's fork count (Spearman $+0.02$, $p = 0.81$, 119 problems). Forks fall by half from the first third of a base trace to the last while cross-model divergence is flat along it.

Table 6: **The census per band, Qwen pair.** Forks per problem after the noise-floor correction, the swing in success rate between the better and worse option, how often each model, scored on the same option set at the same prefix, puts more probability on the winning option, and the number of forks scored by both models (n).

band	forks per problem	swing	base	reasoning	n
unreached	0.0	—	—	—	0
rare	3.7	.45	.51	.50	80
contested	6.3	.46	.46	.50	50
reliable	1.5	.48	.59	.61	59
mastered	0.1	—	—	—	1

Lookahead and probe. A one-step lookahead that completes each of the top two options three times at the census's sampled positions and takes the best, scored on three held-out completions, lifts the base model from .12 to .24 on the rare band and from .50 to .57 on the contested band, and the optimistic bound with no hold-out is .44 and .67. The reasoning model sits at .77 and .90 on the same problems. The procedure loses six points on mastered problems, which is the size of its completion noise, so the achievable figures are if anything low. A pilot that predates the core set and the gate drew 327 AIME problems from 2013 to 2023 with no overlap with the evaluation pool, ran the base model eight times per problem (oracle coverage .440), labelled 205 forks by forced completion with balanced outcomes (mean fate .50), and trained logistic probes on the residual stream at two layers and their concatenation. The bar was a leave-one-problem-out AUC of .70, below which a probe-based selector would not be run. The probes reached .567, .597 and .596, and a pooled logistic probe .602, so the selector was not run. A probe trained on MATH forks and applied to the AIME forks scored .418 and .403, chance. The winning side of a fork is therefore not linearly readable from the prefix at this scale and with this much training data.

Majority vote. On the contested band a majority vote over eight of the base model's own runs reaches .84 for about 5,700 tokens, against .90 for the reasoning model at one run of about 6,800 tokens, and sixteen runs reach .92 (Table 7). The correct answer is the plurality of the base model's stored samples (32 to 64 per problem) on 97% of those problems, so on that band post-training compresses a vote into one run. On the rare band the vote reaches .19 with eight runs and .24 with sixteen, against .77, because the correct answer is the plurality on 23% of problems and the base model spreads over about twenty distinct answers per problem. On the rare band one run of the reasoning model is worth about sixteen base draws with a perfect selector (oracle .74 at k=16, resampled from the stored samples), and the transplant about eleven (Figure 7). Votes are resampled from the stored base samples per problem, and token costs are medians over all runs of the band.

With eight samples per problem the base model scores .497 per sample, .607 by majority vote and .700 by oracle, the quantities of Wang et al. (2023) and Snell et al. (2025), and the reasoning model .746, .797 and .807. The reasoning model's vote is within a point of its own oracle, so its remaining headroom is coverage, while the base model's nine-point gap between vote and oracle is selection it could still buy. Both readings are specific to $k = 8$. Coverage keeps climbing with k, and the reasoning model's oracle is censored by the 19.8% of its rare-band runs that reach the ceiling.

Table 7: **Majority vote over the base model's own runs** against the reasoning model at one run, per band, with token cost per problem. On the contested band the vote matches the reasoning model at the same cost, and on the rare band it saturates far below it.

band	base k=1	vote k=4	vote k=8	vote k=16	oracle k=8	tokens, vote k=8	reasoning k=1	tokens, reasoning
rare	.12	.15	.19	.24	.56	6,844	.77	10,514
contested	.50	.69	.84	.92	1.00	5,736	.90	6,812
reliable	.90	.99	1.00	1.00	1.00	2,938	.94	2,768

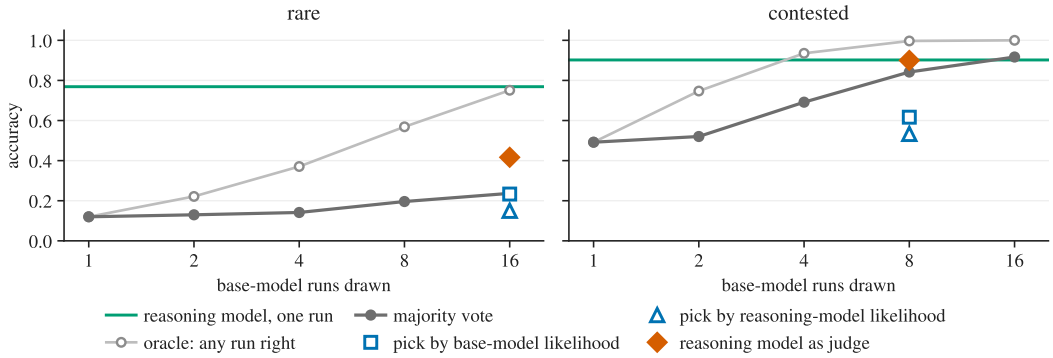

Figure 7: **What sampling and selection recover, Qwen pair.** Majority vote and oracle over the base model's own runs against the number drawn, and four selection rules over sixteen fresh runs per rare problem and eight per contested, against the reasoning model at one run (Table 8).

Selection. Sixteen fresh base runs per rare problem and eight per contested problem were scored under both models and judged by the reasoning model, thinking on, with a yes or no verdict. Selection rules pick among the same runs, and a decision rule reads the result. A judge more than .10 above the likelihood rules means post-training installed a recogniser, within .10 of them means amortised likelihood, and near the majority means no usable recogniser. Table 8 gives the result. The judge is .18 above the best likelihood rule and .20 above the majority on the rare band, and the reasoning model's own likelihood is a worse selector than the base model's, so the recogniser is not a sharpened base distribution. On the contested band the judge and the majority agree, .90 and .82, while the likelihood rules fall below both. The judge said yes to 75% of correct rare-band runs and 13% of wrong ones while returning no verdict on 59% of them, so its rare-band row is a lower bound. The oracle here is .55 at sixteen runs against .74 at sixteen draws of the stored reference samples (Table 7 gives .56 at eight), because these are fresh runs at an 8,192-token cap, and the two are not the same draw.

Table 8: **Selecting among the base model's own runs**, sixteen per rare problem and eight per contested, Qwen pair. Likelihood rules pick the highest-scoring run; the judge rule takes the plurality of runs the reasoning model marks correct; the oracle picks any correct run.

selector	rare	contested
base model, one run	.115	.475
majority vote	.217 [.12, .33]	.817 [.72, .92]
base-model likelihood	.233 [.13, .33]	.617 [.50, .73]
reasoning-model likelihood	.150 [.07, .23]	.533 [.40, .65]
reasoning model as judge	.417 [.30, .53]	.900 [.82, .97]
oracle	.550 [.43, .67]	1.000
reasoning model, one run	.769	.902