A BENCHMARK SCORE IS NOT A MODEL PROPERTY: FAILURE COMPOSITION CHANGES ACROSS OPERATING REGIMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models are typically evaluated through aggregate metrics such as accuracy, which count how often a model fails without recording how it fails. A rising benchmark score can therefore conceal whether an intervention actually improves reliability or merely changes the type of errors a model makes. We study this question with a framework that decomposes failures into various mutually exclusive mechanisms, such as a wrong action (a decision error) or a misreported resulting state (a state error), and tracks how their composition changes across operating regimes and interventions.

Using an exact-simulator benchmark for multi-step reasoning, with no model-based judging, we analyze failure composition across 13 models from 9 families, including open-weight and proprietary systems. We further evaluate five interventions across reasoning, decoding, and representation changes, measuring the aggregate performance and failures movement.

Our analysis reveals two consistent patterns. First, failure composition changes systematically across models and capability regimes. Models with similar aggregate scores can fail through different mechanisms, and across eleven models the dominant failure type shifts with the model's reliable operating horizon (Spearman $\rho = 0.79$). Second, an intervention can eliminate its targeted failure while other failure modes absorb the errors, sometimes without improving overall reliability. In Qwen3-4B, position-based representations reduce state errors from 0.506 to 0.006 while introducing a new formatting failure, and self-consistency turns half of Qwen3-1.7B's would-be decision errors into state errors.

These findings show that aggregate metrics alone provide an incomplete picture of reliability. We argue that evaluations should measure how failures change under scaling and intervention, as well as how many remain.

1 INTRODUCTION

Suppose an intervention drives its target failure to zero. It is natural to conclude that the system became more reliable by the same amount. Test-time methods are usually judged this way: self-consistency by aggregate accuracy (Wang et al., 2023), and grammar-constrained decoding by whether the constraint holds (Park et al., 2024). In our experiments that conclusion fails. A targeted failure went to zero while total error stayed where it was, and a failure type absent from all 500 baseline answers appeared. An intervention can change the distribution of failure across mechanisms instead of reducing it, and an audit scoring only the targeted metric records a clean success in exactly that case.

One fact sets the problem up. At the largest size where Qwen3-1.7B holds a usable regime it fails mostly by emitting a non-optimal move (0.170 against 0.015 state-realization error); at the largest size where Qwen3-4B holds one, it mostly emits an optimal move with an incorrect resulting state (0.150 against 0.030; Section 5). Same benchmark, same scoring, opposite dominant mechanism. Measuring such differences needs exact grading and a fixed interface. We therefore reuse LEAD's environment, per-step formulation, grader and hard-step screening unchanged (Pushkin & Abbe, 2026); none of them is our contribution (Section 2).

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.

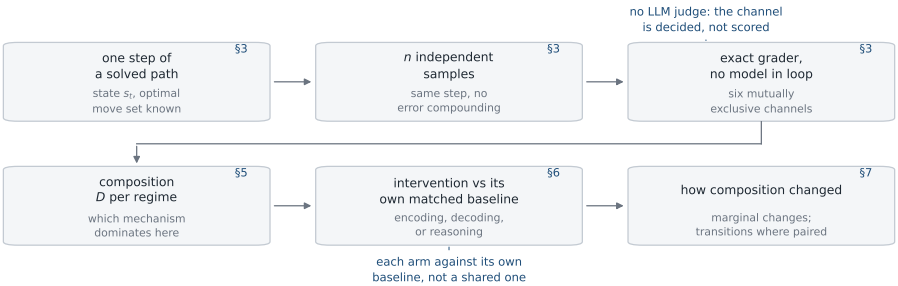

Figure 1: **How a per-step corpus becomes a statement about an intervention.** Schematic; no value here is measured. Marked: the grader is an exact simulator, and each intervention has its own baseline.

C1. Interventions can change failure composition, and the pattern differs by intervention. Three methods aimed at different mechanisms are scored against matched per-intervention baselines on one corpus with one exact grader (Figure 3). Invariant-constrained decoding drives invariant-violating outputs to zero by construction, yet total state-realization error does not detectably move, $0.392 \rightarrow 0.368$, difference -0.024 [-0.192, +0.128]: the violating sub-mode falls 0.224 and the sub-mode the constraint cannot reach rises 0.200. The arms are separate generations over five step clusters, so this is a marginal *reallocation signature*, not a record of individual errors changing channel. Position labels take state-realization error from 0.506 to 0.006 while a non-compliance channel appears where there had been none. **For self-consistency the transfer is identified sample by sample**, because it selects among its own baseline generations, at k times their sampling: for Qwen3-1.7B at $N{=}10$, 0.507 of would-be decision errors become state-realization errors and 0.121 become correct, and at $N{=}12$ a quarter (0.249) do (Section 7.2). Reallocation is not uniform futility: the share of removed failure mass that reappears in another channel runs from 0.07 where labelling buys real reliability to above 1 in the hardest regimes, and it tracks regime difficulty in a second environment as well (Section 7.1).

C2. Failure is a mixture of channels, and the mixture differs across operating regimes. An exact grader separates six mutually exclusive outcomes; pooling truncation with substantive error would have made two thirds of one cell's error a `max_tokens` artefact (Section 4). At nearly equal error, Qwen3-1.7B at $N{=}4$ (0.240) and Qwen3-8B at $N{=}10$ (0.265) have state shares of content errors of 0.081 and 0.860. Within Qwen3 the capability horizon is monotone in scale, but composition separates only across the largest gap and is flat from 4B to 8B, so this is not a scaling law. Across the eleven models of nine families that the inclusion rule admits, the dominant failure shifts from decision to state error as the horizon grows ($\rho{=}0.79$, exploratory; Section 5.4).

C3. Within the regimes we evaluated, composition predicts where an intervention will be *sensitive*; it does not identify the right intervention or how much reliability it buys. We committed a prediction rule (dominant channel $\rightarrow$ intervention family) before the arm that tests it existed. In the five regimes where it named a *reasoning* intervention, decision error fell in four with intervals excluding zero, and those regimes gained roughly five times the accuracy of the regimes it passed over ($+0.122$ against $+0.023$): one rule and one intervention family, so evidence of actionable signal rather than a selection procedure. The dominant channel does not vary across steps where an intervention is measurable, so we do not propose a controller and report that we looked for one and found no target.

Scope: the apparatus is audited (Section 4); the reallocation results cover two exact-simulator tasks and two model families; the composition results cover nine families on one task.

2 BACKGROUND AND RELATED WORK

LEAD (Pushkin & Abbe, 2026) supplies the task, the prompt text and the framing we argue with; its per-step formulation is what makes any of this measurable. Appendix I discusses each work below in full.

Nearest work: failure diagnosis. Barke et al. (2026) localise the critical failing step of an agent trajectory with an LLM-based judge. Wang et al. (2026) study how agents degrade with horizon across more than 3,100 trajectories, attribute failures with a human-validated LLM-as-judge pipeline, and do not evaluate interventions. Zhu et al. (2025) annotate failure trajectories and show that targeted feedback lets agents recover, measured by task success. We claim no taxonomy and no diagnosis method. Our evidence differs in kind: per-step outcomes graded by an exact simulator with no model in the loop, and each intervention scored on the full distribution of mutually exclusive channels rather than on its target or on task success alone. Zhu et al. (2026) extend diagnosis toward automated recovery.

The interface as a locus of failure is also established (Zhang et al., 2026; Xu et al., 2026; Rozanov & Rei, 2024; Chae et al., 2024); the inference-time gate of Zhang et al. (2026) is structurally the channel-aware controller our data argues against. Measurement-validity auditing is shared practice (Messing, 2026; Tu et al., 2026; Thomas et al., 2026), and constrained decoding (Park et al., 2024; Li et al., 2026) supplies the method whose effect we characterise.

Our contribution is controlled exact-grader evidence that an intervention must be evaluated across the full channel distribution, not on its target alone.

3 SETUP

Task and grading. We use the Checkers-jumping puzzle from LEAD (Pushkin & Abbe, 2026) in its per-step form, with that paper's Appendix H prompt text byte-identical and hash-pinned. A board of $2N + 1$ cells holds N red checkers on the left, N blue on the right, one empty cell. The model is shown one intermediate position on the canonical optimal path and must return the next optimal move and the resulting board as `{'move': ..., 'new_state': ...}`. The model does not solve the whole puzzle. We sample each step of a length-L path separately, so we see behavior along the whole path and an early error cannot carry into later steps. A response is correct when its move is in the optimal set and its board equals what that move produces; `optimal_moves` was verified equal to the breadth-first shortest-path set for every board with $N \leq 7$. There is no rubric and no model in the loop. Section 7.1 adds Tower of Hanoi, graded the same way.

Six channels, and the split is load-bearing. We score six mutually exclusive outcomes: `correct`; `wrong_move`, move-selection or *decision* error (the move was not optimal); `wrong_state`, *state-realization* or *state* error (the move was optimal and the resulting state is wrong); `budget_exhausted` (cut off); `malformed` (the model bound the requested keys and our reader could not read them); and `non_compliant` (it never bound them at all). The last three are usually pooled as *invalid*, and pooling them cost this project two conclusions. The `malformed`/`non_compliant` boundary is decided by one question about the completion (did it ever bind the keys to literal values) and never by whether our parser succeeded, so it does not move when the parser improves.

Why `wrong_state` is called *state-realization* error. The channel is defined as *the move was optimal and the board is wrong*, which fits a copying failure, a misapplied transition, or interference from emitting both in one object, so we keep the weaker, definitional name (Appendix B).

Encodings, models, intervals and zeros. Each variant of the state representation is derived from the verbatim prompt by editing four enumerated sites and nothing else: *position-labelled* writes each cell as `'<index>:<symbol>'`, and two further variants appear only in Appendix F. Qwen3-1.7B, 4B and 8B are served locally under vLLM at bfloat16 on two GPUs, thinking mode on unless stated. Hosted models are described in Section 5.4. Every rate carries an exact Clopper–Pearson interval; differences of proportions carry Newcombe's hybrid-score interval, which is *a score-based approximation and not exact*, and only where the two arms share their units, a permutation test

standing in where they do not (Appendix C reports a case where the approximation would have produced a false finding). Intervals are within-run: the measured run-to-run component is $1.38\times$ sampling, [0.65, 2.20], an interval containing 1, so no inflation is applied. Where a cell produced no errors we write the count, the sample size and the exact upper bound, every time. **An observed zero is not a zero rate.**

4 THE INSTRUMENT

We claim no contribution for the apparatus, but three facts about it bind every result (Appendix C). A token ceiling can be two thirds of a headline error rate: at 6144 tokens, 0.186 of a total per-step error of 0.270 [0.240, 0.302] is budget exhaustion, so every arm below runs at 13312 and every rate is quoted with its ceiling and truncation. Truncation is *informative* censoring, so a rate measured under material truncation is a lower bound. Two parser defects manufactured 38 state-realization errors that were never made, so all 28,155 samples were re-graded behind a consistency gate, and the reader's own residual error is measured, at 0.00031 [0.00013, 0.00061] per sample. Selection and measurement use independent seeds, so the winner's curse is observed rather than warned about: $+0.119$ [+0.037, +0.197], against a split-half forecast of $+0.037$ to $+0.075$.

5 FAILURE COMPOSITION REORGANIZES ACROSS OPERATING REGIMES

The scale dimension has three points (Qwen3-1.7B, 4B and 8B) measured under one protocol on one corpus: $N \in \{4, 6, 8, 10, 12\}$, the *same* twenty steps per N for all three models drawn once from a fixed seed, n=10 per step, 200 samples per cell, thinking on, ceiling 13312, bf16, one serving configuration. Sharing the step set is what makes the models comparable at fixed difficulty. The thinking-on/off dissociation that fixes this regime is in Appendix D.

5.1 THE CAPABILITY HORIZON IS MONOTONE IN SCALE OVER THE TESTED THRESHOLDS

Let $N^*(q)$ be the largest board on which a model's cell-mean per-step error stays at or below q, its *reliable operating horizon*, and let H_q be the interpolated crossing from a monotone logistic fit, with a step-level bootstrap interval. We chose the 0.30 bar, so we also report the crossing for every $q \in \{0.20, \ldots, 0.40\}$ and a threshold-free area.

The crossings are 5.02 [4.14, 5.77], 8.86 [8.24, 9.56] and 10.55 [9.93, 11.34] for the 1.7B, 4B and 8B (fits and threshold-free areas in Appendix D). Error rises with board size for all three models, and no point was smoothed or dropped. For every tested $q \in [0.20, 0.40]$ the order is $1.7\text{B} < 4\text{B} < 8\text{B}$ with disjoint bootstrap intervals. The one exception is q=0.40, where the 4B and 8B intervals *touch* at 10.59. The 8B's N=12 error (0.490) exceeds 0.30, so no horizon is open-ended.

5.2 COMPOSITION DOES NOT INHERIT THAT ORDERING

Let $D(m, N) = P(\text{state-realization} \mid \text{state-realization or move error})$ be the state-versus-move share of *content* errors. Budget exhaustion, malformed output and non-compliance are excluded from D by construction and reported separately, so the exclusion is auditable rather than silent (Figure 5). Where a cell has no content errors, D is undefined and not imputed. Because every model runs the same steps, contrasts are paired at the step level and every interval resamples the twenty step clusters rather than the two hundred samples. Per-cell values and the full paired-contrast table are in Appendix D.

The largest capability gap reorganizes composition everywhere it is measurable. 8B $-$ 1.7B is positive and excludes zero at all four rungs where both models produce content errors: $+0.661$, $+0.594$, $+0.303$, $+0.257$. The intermediate gap reorganizes it at the harder rungs: 4B $-$ 1.7B excludes zero at N=8, 10, 12 but not at N=4 or N=6.

The 4B $\rightarrow$ 8B step is flat. It excludes zero at exactly one rung, N=6, and that is the least stable cell in the table: the 8B has *four content errors in total* there, giving D=1.000 with an exact interval of [0.398, 1.000]. At N=8, 10 and 12 the contrast is $+0.108$, $+0.015$ and -0.045, all including zero. So the strict three-way monotone chain is not established at any single rung.

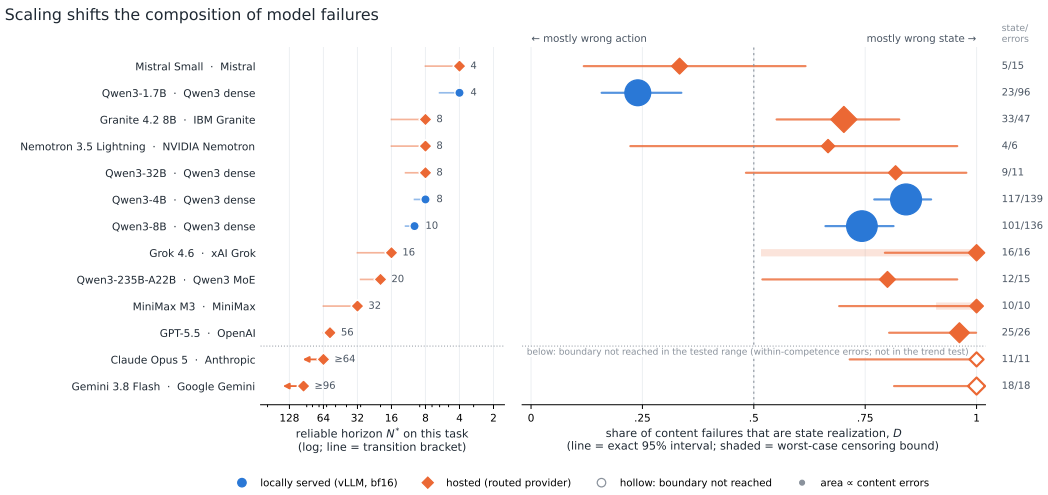

Figure 2: **Scaling shifts the composition of model failures.** One row per model, sorted by reliable horizon N^* (left, log; line = transition bracket). Right: state share D of content failures, with its exact 95% interval and, where answers were truncated, the worst-case censoring bound (shaded). Circles locally served, diamonds hosted; hollow rows never reached a boundary and are not in the trend test. N^* is a horizon on this task, not a capability score.

Under the decision rule fixed before the 8B ladder was run, this is the middle outcome of three: the hypothesis holds *in direction* but not monotonically, so we state that failure composition reorganizes across operating regimes, not a scaling law (Appendix D).

5.3 The dominant channel inverts between the two smaller models

Reading each model at its own N^*, rather than at a shared N, shows opposite dominant failures. At its $N^*{=}4$, the 1.7B has move-selection error 0.170 and state-realization error 0.015. At its $N^*{=}8$, the 4B has 0.030 and 0.150. The smaller model mostly picks a non-optimal move. The larger model mostly picks an optimal move and then writes the wrong resulting board. Aggregate error hides this difference. The 1.7B at $N{=}4$ and the 8B at $N{=}10$ have similar cell error (0.240 and 0.265) but $D{=}0.081$ and 0.860 (a post-hoc sensitivity analysis, Appendix D). Section 6 shows that the two failures call for different interventions. Outside the ladder, Qwen3-235B-A22B-Thinking solves transplanted steps but does not give the answer in the requested format. We report this as a phenomenon, not a rate (Appendix D).

5.4 Beyond the ladder: thirteen models, nine families

We ran the same task, pinned prompt and exact grader on hosted models from eight more families (Figure 2). They include open-weight models and proprietary models that are only available through vendor APIs (GPT-5.5, Claude Opus 5, Gemini 3.8 Flash). Appendix L lists every model. One rule, applied to every model regardless of the direction of its result, decides what is plotted. We read composition where a model starts to fail, and keep the reading only if it rests on at least five content errors and its worst-case censoring bound stays on one side of 0.5. Two models fail only where they also run out of tokens, so we report them but do not plot them. Two others never cross the bar in the tested range, and we show them separately (Appendix L). The two models with the shortest horizon fail mostly by choosing the wrong move. Every model with a longer horizon fails mostly by writing the wrong state. Qwen3-1.7B and Mistral Small ($N^*{=}4$) have $D{=}0.24$ and 0.33. The other nine, from Granite 4.2 8B to GPT-5.5, lie between 0.67 and 1.00. Across these eleven, $\rho = 0.79$ ($p = 0.005$), and the correlation survives each sensitivity analysis we ran; since the rule was fixed after the data, we treat it as exploratory.

6 THE INTERVENTION AT EACH MODEL'S OWN HORIZON

We evaluate five interventions: three representation changes (position labels, the empty-cell glyph within them, and an action-only output interface, Appendix J), one decoding change (invariant-constrained decoding), and one reasoning change (self-consistency, Section 7).

6.1 AT THE 4B, AT ITS OWN HORIZON, ON ITS OWN STEPS, ON A THIRD SEED

Nothing here is transplanted: Stage 0 located $N{=}8$ as the 4B's horizon, Stage 1 sampled all 80 steps with a second seed and found 27 hard, and Stage 2 measured the top 20 with a third seed fixed before the screen ran, 500 samples per arm. No sample ever selected a step and then measured it.

Raw scores `correct` 0.388 [0.345, 0.432] with state-realization error $\mathbf{0.506}$ [0.461, 0.551]; the position-labelled arm scores 0.928 [0.902, 0.949] with state-realization $\mathbf{0.006}$ [0.001, 0.017]. All six channels, both arms and all 20 steps are in Appendix K. Two readings were registered together before the data existed, with the reconciliation rule fixed in advance. Pooled, the difference is $+\mathbf{0.500}$ [+0.455, +0.544] by Newcombe; with STEP as the unit the mean per-step difference is $+\mathbf{0.500}$, paired bootstrap [+0.412, +0.588], exact sign test 20 positive, 0 negative, $p = 1.91 \times 10^{-6}$. The two agree to three decimals, the registered agreement case in which the effect survives clustering and does not rest on a few steps (Figure 7).

6.2 THE INTERVENTION IS NOT FREE, AND THE TRADE IS STEREOTYPED

The position-labelled arm introduces a failure the raw arm does not have: `non_compliant` 0.040 [0.025, 0.061] against 0 of 500. All 20 are one defect, a dropped closing quote after the empty cell's label, and none is truncated; all 23 of the arm's non-correct, non-budget, non-move outcomes are failures to write that label, and not one is a wrong board. We did not repair it in the reader: restoring a dropped quote means deciding what the model meant, the class of rule that fabricated 38 errors here. We bound it instead, at worst 0.046 [0.029, 0.068] against raw's 0.506. **The labels therefore move part of the error into a different mechanism.** Most of this new failure comes from the notation, not from labelling itself. Changing one glyph cuts format failure to 0.008 [0.002, 0.020] at the same token cost, and state-realization error does not detectably change (Appendix F).

6.3 THE RESIDUAL MATRIX: THREE INTERVENTIONS, ONE CORPUS, ONE GRADER

We report the three interventions in one table. Each is scored against *its own* matched baseline. The glyph change modifies the position-labelled encoding, so scoring it against raw would credit it with the labelling gain. Every contrast is paired over the shared steps and uses $20{,}000$ step-cluster resamples. Table 1 in Appendix F gives the full matrix.

Each predeclared target moves. None of the $20{,}000$ resamples crosses zero, so each raw p sits at the resampling floor, $< 5 \times 10^{-5}$. Holm correction over the three targets gives $p < 1.5 \times 10^{-4}$. A paired cluster-level sign-flip test agrees for the two interventions with enough clusters to run one. The three interventions change the channels in different ways. **Position labels** remove almost all of the targeted error (residual ratio 0.01) and create a new channel: non-compliance rises from exactly 0.000 to 0.040 [+0.022, +0.060]. An audit of the target channel alone would have missed the second effect. **Constrained decoding barely changes its total target.** It drives invariant-violating outputs to zero ($0.224 \to 0.000$, guaranteed by construction), but total state-realization error does not detectably move: $0.392 \to 0.368$, a difference of -0.024 [-0.192, +0.128] and a residual ratio of 0.94. The displaced share rose by 0.200, and every remaining error has the right pieces in the wrong order. *An audit that scored only invariant violations would call this intervention a complete success.* We state this column as a mechanism rather than an inferential claim: five step clusters do not support one, and its interval spans 0.32.

Over the 80-step path at $N{=}8$, trajectory success reads as zero for every one of these methods, and the 50-point gap between them disappears (Appendix F). **Which side of the interface carries the effect.** Two registered single-side arms put it on the side the model must *write*: 0.940 [0.905, 0.970] of the reduction against 0.348 for the board it reads (Appendix F).

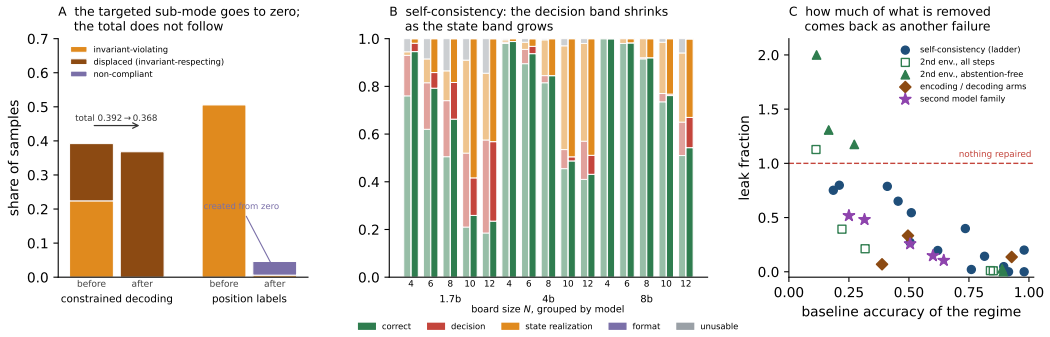

Figure 3: **The interventions we tested change failure composition, each in a different pattern.** A and C compare separately sampled arms (marginal changes); only B pairs outcomes within the same generations. **(A)** Constrained decoding eliminates the invariant-violating sub-mode by construction while the displaced one rises by almost as much; position labels collapse the state channel while a non-compliance channel appears. **(B)** Composition before (pale) and after (solid) self-consistency at k=5 on the ladder; the decision-to-state transition is identified sample by sample. **(C)** Leak fraction (Section 7.1) against baseline accuracy for every arm in the paper; above the dashed line total error does not fall. Values in `fig_relocation.csv`.

7 A PREDICTION MADE BEFORE THE ARM EXISTED

The order of operations is the evidence for this test, so we state it as a sequence. **(1)** Fifteen baseline cells (three models $\times$ five board sizes) were sampled and their composition measured; one, the 8B at N=4, has no failures, so no channel dominates it and it carries no prediction, leaving fourteen. **(2)** A four-branch map from dominant channel to intervention family was written and frozen in a timestamped registration file (Appendix A), nominating five of the fourteen as regimes where a decision-quality intervention should matter. The same file scores the map, post hoc, against the representation interventions whose outcomes were already known; for self-consistency the map is prospective, because no self-consistency outcome had been computed in any form. Self-consistency reuses the baseline generations, so they predate the map; its outcome does not. **(3)** Self-consistency (plurality move, then the complete answer of a voting sample, so both channels are free to change) was evaluated on all fourteen, the nominated and the passed-over alike, at k times the sampling of the single-sample baseline. **(4)** The comparison below was computed once. **The prediction separated the regimes**: nominated regimes gained $+0.122$ mean accuracy against $+0.023$ for the nine it passed over, and the decision channel fell in four of the five with intervals excluding zero.

It is not selective repair. The state channel rose almost wherever the decision channel fell, $+0.052$ against $+0.044$ in the two groups, far closer than the accuracy means. The clearest case is the regime the rule *declined*: 1.7B at N=10, where decision error falls 0.153 while state-realization error rises 0.194 and net correctness moves only $+0.049$.

7.1 THE SAME TRADE IN A SECOND ENVIRONMENT, AND A SECOND MODEL FAMILY

Tower of Hanoi is in the same corpus under the same grader with 25 samples per step, so the identical procedure runs on it with no new sampling, against the thinking-off Checkers cells at the same model and ceiling, matched on regime, so only the task differs. We summarize each cell by a *leak fraction*: the rise in the failure channels an intervention does not reduce, as a share of the fall in those it does. A value of zero means clean repair. A value of one means total error did not change, and a value above one means it rose. For separately sampled arms it is a ratio of marginal changes. The leak fraction follows regime difficulty, not the task (Figure 3C). On Hanoi, where the baseline is already 0.84–0.85 correct, it is 0.01. On Checkers it rises as baseline accuracy falls and reaches 1.13 at N=10, where accuracy changes by -0.006.

The pattern also holds outside Qwen. DeepSeek-R1-Distill-Llama-8B has a different base model and training recipe. We ran it through the same harness and grader, and it was the only one of four candidates to pass the comparability gate. At all five of its rungs, the decision channel falls and the

state channel rises. Its leak fraction also grows with difficulty: 0.10, 0.15, 0.26, 0.48, 0.52 as baseline accuracy falls from 0.645 to 0.250 (Figure 3C). The level differs: this model stays decision-dominant at every rung, while the Qwen 4B and 8B become state-dominant.

Part of the effect comes from the vote itself. A sample that states no move cannot win a vote, so voting removes unusable answers. On steps where every sample stated a move, the apparent benefit on Checkers disappears (leak 1.18–2.00, accuracy change ≤ 0), while the benefit on Hanoi remains. On Checkers, the harder task, the gain came from not selecting non-answers. That gain is real in deployment, but it does not change how the model decides.

7.2 Where the failures go, for the one arm that can say

Two sets of marginal rates cannot show that individual failures *moved*, because many different flows produce the same before and after rates. For the encoding and decoding arms the question is undefined, because a changed prompt makes the token streams diverge from the first sampled token. Self-consistency instead *selects* among the draws that define its baseline, so within a trial a uniformly random member of the k drawn is what one sample would have given and the voted member is what the intervention gave: both real generations, paired. Row masses reproduce the baseline rates to within 0.010, which checks the pairing.

Many would-be decision errors become state errors; few become correct. Transition shares are expectations over 400 resampled $k{=}5$ vote subsets per step on 20 steps, reported as point estimates without intervals. At the 1.7B on $N{=}10$, of the samples that would have been a decision error, 0.507 come back a state-realization error and only 0.121 come back correct; at the 4B on $N{=}12$ it is 0.556 against 0.045. The third cell, the 1.7B on $N{=}12$, is smaller: 0.249 become state errors, 0.129 correct, and 0.623 remain decision errors. The state channel also takes mass from samples that would have been correct, 0.068 and 0.082 of all samples in those two cells. Voting settles on a move the samples agree about and then ships a board that is wrong, which is a mechanism and not a bookkeeping artefact. For this arm the word *relocation* is earned at the sample level.

8 Limitations

Registration. The staged measurement of Section 6 and the prediction rule of Section 7 are pre-registered. The composition findings across scale, the cross-environment, transition and model-campaign analyses, and the finding that most state-realization errors violate a task invariant (Appendix G) are post hoc, and none of their p-values is confirmatory (Appendix M).

Families, tasks and serving. The reallocation results rest on two families in one regime; composition on nine, most hosted points resting on tens of content errors from routed providers at unpinned quantizations on sparse grids, so N^* is task-specific and grid-bounded. The inclusion rule of Section 5.4 was fixed after the data, and two families are unmeasurable at a 32,768-token budget. Within the Qwen ladder $4\text{B} \to 8\text{B}$ is flat wherever we can resolve it, and we do not know whether that is a plateau or a local feature. The residual matrix is one model at one rung.

What the transition results do not settle. Self-consistency is not equal-compute, the abstention-free subsets are underpowered, and transitions are defined for that arm only.

What we declined, and did not resolve. Three intervention regimes are *declined* because their target channel is too rare to resolve at any affordable n (Appendix D.1), which is not "no effect". The constrained-decoding column rests on five clusters and is a mechanism, not an inferential claim. All thresholds are labelled choices (Appendix H).

9 Conclusion

An aggregate score does not identify how a system fails. Across the regimes we measured, short reliable horizons go with non-optimal moves and longer ones with incorrect resulting states. Interventions change that composition in ways a target-only evaluation misreports: separately sampled arms show marginal reallocation, and self-consistency turns a large share of removed decision errors into state errors sample by sample. Within our scope, evaluations should report the full channel distribution with aggregate correctness.

REPRODUCIBILITY STATEMENT

Every number in this paper is re-derived from the stored per-sample records by one documented command, and the supplement contains a test that fails if any number in the prose disagrees with the analysis output. The test caught a contrast written as $+0.288$ where the value rounds to $+0.287$, and an interval endpoint written as $+0.068$ where it is $+0.069$, both introduced by transcribing a console print instead of the stored result, which is exactly the error class no reader can detect.

What the anonymous supplement contains. The analysis and serving code; the exact prompt corpus with its hash pins; every configuration; both freeze documents and the intervention-prediction registration file; the decision and observation logs; a machine-readable table of every reported effect with its interval, its n, its number of independent step clusters and whether it was predeclared or post hoc; the CSV of plotted values beside every figure; and the test suite, including the test that binds prose numbers to analysis output.

What it does not contain. The per-sample records (about 540 MB) exceed the upload limit and are archived separately; they will be released with the de-anonymised version. Without them a reader can check that the prose matches the released tables and figure values but cannot re-derive those values. Version-control history is withheld during review because it identifies the author; it is what dates each registration relative to its data, and it will be released on de-anonymisation. If a run directory is missing, each analysis script and test skips it without failing, and a test checks this behaviour. Sampling is not bit-for-bit reproducible across serving stacks. The supplement includes a parity check that reports the smallest difference it could detect. A result of "no difference detected" does not mean the two stacks are equivalent.

Seeds and protocol. All seeds are fixed, and none was re-drawn. All three models run the same twenty steps per board size, so they are compared at the same difficulty. Screening and measurement use separate samples and different seeds, so no step is selected and measured on the same draw. Each arm and device has its own base seed, because identical seeds would return identical completions and silently duplicate samples. The output ceiling is the same for every arm. We never lowered it to save context or because an arm's answer format is shorter, because either change would alter the budget-exhaustion channel.

Pre-registration. Two freeze documents govern this work. Each was written before the data it governs existed, and each states, per arm, what had already been seen. Neither claims to be a blind registration. The confirmatory arm of the scale ladder used a model that had not been run when we formed the hypothesis. The matched-error analysis is labelled a post-hoc sensitivity analysis everywhere, including in its own program output, and it cannot change the frozen verdict. Each deviation is recorded with a reason and a condition for reversing it.

AI USE STATEMENT

We used an LLM coding assistant to help implement experiment and analysis code and to help draft and edit the text. All code is covered by automated tests, every reported number is produced by scripts from the stored records, and the author reviewed all code, results and text and takes full responsibility for them.

REFERENCES

Milad Aghajohari, Kamran Chitsaz, Amirhossein Kazemnejad, Sarath Chandar, Alessandro Sordoni, Aaron Courville, and Siva Reddy. The Markovian thinker. In *The Fourteenth International Conference on Learning Representations*, 2025.

Shraddha Barke, Arnav Goyal, Alind Khare, Avaljot Singh, Suman Nath, and Chetan Bansal. AgentRx: Diagnosing AI agent failures from execution trajectories. *arXiv preprint arXiv:2602.02475*, 2026.

Hyungjoo Chae, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. Web agents with world models: Learning and leveraging environment dynamics in web navigation. *arXiv preprint arXiv:2410.13232*, 2024.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A. Huerta, and Hao Peng. Context length alone hurts LLM performance despite perfect retrieval. *arXiv preprint arXiv:2510.05381*, 2025.

Arian Hosseini, Alessandro Sordoni, Daniel Toyama, Aaron Courville, and Rishabh Agarwal. Not all llm reasoners are created equal. *arXiv preprint arXiv:2410.01748*, 2024.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems*, 2025.

Fangzheng Li, Aimin Zhang, and Chen Lv. Constraint tax in open-weight LLMs: An empirical study of tool calling suppression under structured output constraints. *arXiv preprint arXiv:2606.25605*, 2026.

Huayang Li, Pat Verga, Priyanka Sen, Bowen Yang, Vijay Viswanathan, Patrick Lewis, Taro Watanabe, and Yixuan Su. Alr^2: A retrieve-then-reason framework for long-context question answering. *arXiv preprint arXiv:2410.03227*, 2024.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Shukai Liu, Jian Yang, Bo Jiang, Yizhi Li, Jinyang Guo, Xianglong Liu, and Bryan Dai. Context as a tool: Context management for long-horizon SWE-agents. *arXiv preprint arXiv:2512.22087*, 2025.

Solomon Messing. Hidden measurement error in LLM pipelines distorts annotation, evaluation, and benchmarking. *arXiv preprint arXiv:2604.11581*, 2026.

Elliot Meyerson, Giuseppe Paolo, Roberto Dailey, Hormoz Shahrzad, Olivier Francon, Conor F. Hayes, Xin Qiu, Babak Hodjat, and Risto Miikkulainen. Solving a million-step LLM task with zero errors. *arXiv preprint arXiv:2511.09030*, 2025.

C. Opus and A. Lawsen. The illusion of the illusion of thinking. *arXiv preprint arXiv:2506.09250*, 2025.

Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D'Antoni. Grammar-aligned decoding. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2405.21047.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Denys Pushkin and Emmanuel Abbe. LEAD: Breaking the no-recovery bottleneck in long-horizon reasoning. *arXiv preprint arXiv:2603.06870*, 2026.

Nikolai Rozanov and Marek Rei. StateAct: Enhancing LLM base agents via self-prompting and state-tracking. *arXiv preprint arXiv:2410.02810*, 2024.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*, 2025.

Akshit Sinha, Arvindh Arun, Shashwat Goel, Steffen Staab, and Jonas Geiping. The illusion of diminishing returns: Measuring long horizon execution in LLMs. In *International Conference on Learning Representations*, 2026.

Maria Thomas, Kristina Gligoric, and Nihar B. Shah. Mitigating LLM-based p-hacking by preregistering for the next LLM. *arXiv preprint arXiv:2606.27687*, 2026.

Xinming Tu, Tianze Wang, Yingzhou Lu, Kexin Huang, Yuanhao Qu, and Sara Mostafavi. BenchGuard: Who guards the benchmarks? automated auditing of LLM agent benchmarks. *arXiv preprint arXiv:2604.24955*, 2026.

Blerta Veseli, Julian Chibane, Mariya Toneva, and Alexander Koller. Positional biases shift as inputs approach context window limits. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*, 2023.

Qian Wang, Tianyu Wang, Zhenheng Tang, Qinbin Li, Nuo Chen, Jingsheng Liang, and Bingsheng He. Megaagent: A large-scale autonomous llm-based multi-agent system without predefined sops. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4998–5036, 2025.

Xinyu Jessica Wang, Haoyue Bai, Yiyou Sun, Haorui Wang, Shuibai Zhang, Wenjie Hu, Mya Schroder, Bilge Mutlu, Dawn Song, and Robert D. Nowak. The long-horizon task mirage? diagnosing where and why agentic systems break. *arXiv preprint arXiv:2604.11978*, 2026.

Tianshi Xu, Huifeng Wen, and Meng Li. Adapting the interface, not the model: Runtime harness adaptation for deterministic LLM agents. *arXiv preprint arXiv:2605.22166*, 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023.

Eric Zelikman, Qian Huang, Gabriel Poesia, Noah Goodman, and Nick Haber. Parsel: Algorithmic reasoning with language models by composing decompositions. *Advances in Neural Information Processing Systems*, 36:31466–31523, 2023.

Kaituo Zhang, Zhen Xiong, Mingyu Zhong, Zhimeng Jiang, Zhouyuan Yuan, Zhecheng Li, and Ying Lin. Are tools all we need? unveiling the tool-use tax in LLM agents. *arXiv preprint arXiv:2605.00136*, 2026.

Wentao Zhang, Liang Zeng, Yuzhen Xiao, Yongcong Li, Ce Cui, Yilei Zhao, Rui Hu, Yang Liu, Yahui Zhou, and Bo An. AgentOrchestra: Orchestrating multi-agent intelligence with the tool-environment-agent (TEA) protocol. *arXiv preprint arXiv:2506.12508*, 2025.

Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023a.

Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization. *arXiv preprint arXiv:2310.16028*, 2023b.

Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. Gsm-infinite: How do your llms behave over infinitely increasing context length and reasoning complexity? *arXiv preprint arXiv:2502.05252*, 2025.

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, Xiaoteng Ma, Xiaodong Yu, Gowtham Ramesh, Jialian Wu, Zicheng Liu, Pan Lu, James Zou, and Jiaxuan You. Where LLM agents fail and how they can learn from failures. *arXiv preprint arXiv:2509.25370*, 2025.

Kunlun Zhu, Xuyan Ye, Zhiguang Han, Yuchen Zhao, Bingxuan Li, Weijia Zhang, Muxin Tian, Xiangru Tang, Pan Lu, James Zou, Jiaxuan You, and Heng Ji. AgentDebugX: An open-source toolkit for failure observability, attribution, and recovery in LLM agents. *arXiv preprint arXiv:2607.18754*, 2026.

A Reproducing every number in this paper

Every quantitative claim above is bound to a run directory and a command. The binding lives in `reports/paper_claims.csv`, one row per sentence, with the sentence as it appears in the draft, the run directories, the recomputation command, the value that recomputation returns, the value the corresponding internal report printed, and a verdict of VERIFIED, VERIFIED-DIFFERS or UNSUPPORTED. That file is generated by calling the recomputation, so a row cannot drift by being hand-edited.

what	command	output
all claims	`python -m scripts.paper_recompute`	26 recomputed claims
one claim	`python -m scripts.paper_recompute -claim C-002`	one
the ledger	`python -m scripts.make_paper_claims`	`reports/paper_claims.csv`
the figures	`python -m scripts.paper_figures`	`paper/figures/*.pdf` + `*.csv`
the tables	`python -m scripts.paper_tables`	this appendix

Each figure ships a CSV of the exact plotted values beside it. Where the prose and the CSV disagree, the CSV is right and the prose is stale.

Where we disagree with our own earlier analysis. Seven rows carry VERIFIED-DIFFERS. None changes a conclusion and we list them rather than quietly adopting whichever number we prefer. Two are third-decimal Monte Carlo differences between resampling seeds, a paired bootstrap interval of [+0.412, +0.588] against [+0.414, +0.588], and permutation p values of 0.162 and 0.079 against 0.155 and 0.076. One is a rounding slip in an exact interval: the Clopper-Pearson upper on $23/500$ is 0.0682 and an earlier note printed 0.069. In two rows the compared numbers were different quantities. One quoted a control-step rate from an 8192-ceiling cell next to the 13312 cell it was compared with. The other computed a budget share on 79 non-tie steps and compared it with one computed on all 80. In one row the corpus grew: the reader residual was measured over 25,965 completions, and the corpus now holds 29,765. In the last row a per-step monotonicity claim holds for 6 of the 7 budget-dominated steps, not for all of them.

Four rows are UNSUPPORTED. None of them is a claim the paper makes. They are the invariant analysis on a second task, an oracle bound on what a controller could recover, a frontier non-compliance *rate* rather than a phenomenon, and an 8B horizon located under the ladder protocol. Section 8 lists each as a gap.

Pre-registration. Each registered artifact was committed to version control before the data it governs existed: the measurability thresholds, the two ladder-gate branches, the realized baseline band, the two co-primary readings with their reconciliation rule, and the intervention-prediction map of Section 7. The artifacts themselves are in the anonymous supplement. The commit history that dates them is withheld during review because it identifies the author, and will be released on de-anonymisation.

B Why the channel is called state-realization error

The channel's definition is exactly *the move was optimal and the board is wrong*. That is consistent with a copying failure, but equally with applying the transition incorrectly, losing track of source and target cells, or interference from having to emit an action and a next state in one object. Section F.1 narrows the mechanism: the single-side ablation places the label effect mainly on the output side of the interface rather than on reading the board, which still does not separate copying from applying the transition wrongly while emitting, and two of our own results argue against the narrowest reading. The position-labelled arm's three residual errors are empty-cell *label* slips rather than wrong boards, and changing the empty glyph from _ to `E` moves `non_compliant` from 0.040 to 0.008, a notation effect. We therefore reserve "transcription-like failure" as a named hypothesis and use the weaker, definitional name for the channel throughout.

The other two encodings. *typed* gives each checker a stable identity label; *alternation-only* alternates letter case within each run of one color, supplying a parity cue and neither position nor identity. Both are derived from the verbatim prompt by editing the same four enumerated sites. The encoding variants also add two sentences to the verbatim prompt, in all four arms alike, saying that a cell's color is its symbol and that writing a color preserves the label.

C THE INSTRUMENT, IN FULL

Two thirds of a headline error rate was a token ceiling. At a 6144-token ceiling the thinking-on cell scores total per-step error 0.270 [0.240, 0.302], of which 0.186 [0.160, 0.215] is budget exhaustion, 0.690 of the error. Raising the ceiling separates the channels: on the 19 steps that cells at 6144, 8192 and 10240 all share, budget error falls 0.253 [0.193, 0.321] $\rightarrow$ 0.101 [0.075, 0.132] $\rightarrow$ 0.016 [0.003, 0.045] while state-realization does not move (0.211, 0.192, 0.216). Per step, 6 of the 7 budget-dominated steps are monotone non-increasing; the exception reverses by less than the half-width of an n=10 cell (Figure 4).

Two parser defects, one of which manufactured errors in the headline channel. The brace scan that locates a model's answer is quote-aware, so a model writing *it's* opens a string literal that never closes and every later brace becomes invisible, the answer included. Separately, the fallback that recovers bare `new_state = [...]` assignments took the *first* occurrence; a model reasoning aloud writes "copy the current state" before it writes the moved state, so the fallback recorded the pre-move board and the grader called it a state-realization error. Re-deriving both affected cells with the fixed reader recovers 38 fabricated state-realization errors (11 and 27; `wrong_state` $0.018 \rightarrow 0.011$ and $0.008 \rightarrow 0.004$). Both defects were found by reading completions, not by reasoning about code. All 28,155 samples were re-graded behind a gate requiring that any sample the two readers reach by the same path re-grade to the outcome on disk; 258 changed, *none in any Checkers arm*.

The reader's residual error is measured. The frame is every sample where the reader extracts no answer and the generation was not cut off. At the last census it held 15 completions, all read, of which 8 were misses: 0.00031 [0.00013, 0.00061] per sample over 25,965 completions, or 0.00027 [0.00012, 0.00053] over the 29,765 the corpus now holds. The frame has since grown to 35; the 20 additions are all one defect described in Section 6, and adjudicating them is a reader judgment we do not make. Counting every one as a miss bounds the residual at 0.00094 [0.00063, 0.00136], which needs no judgment. A further positional fix (take the last list literal) would recover six frontier cases and is refused: it is exactly the mechanism that fabricated the 38. A reader that misses an answer costs a visible non-compliant sample; a reader that guesses wrong costs an invisible state-realization error that looks like a finding.

A card effect that disappeared under permutation. Two GPUs served a screen of 80 steps, 40 steps each. Their total-error rates differ by $+0.097$, and Newcombe's interval [+0.038, +0.157] excludes zero. *That interval is the wrong test here.* It assumes independent draws, but each card served a different set of steps, so the contrast compares two groups of different steps. A permutation test over 20,000 re-splits gives $p = 0.162$ for total error and $p = 0.079$ for state-realization error. Both observed gaps fall inside the null's central range. When we re-ran the screen with the split made *within* each step, so that card and step are unrelated, the mean difference was -0.004, with 10 positive and 10 negative signs and $p = 1.000$.

The winner's curse, measured rather than assumed. Selection and measurement use independent seeds, so we can observe the selection bias directly. The screen's in-sample rate on the 20 measured steps is 0.625 [0.554, 0.692]. A fresh arm on the same steps gives 0.506 [0.461, 0.551], so the realized curse is $+0.119$ [+0.037, +0.197]. A split-half cross-fit had forecast $+0.037$ to $+0.075$. The forecast differs because the two procedures select differently. The cross-fit kept steps above a *threshold*, which returned 27 and 24 steps, while the measurement took the *top 20 by rank*. Selecting by rank is stronger, so a larger realized curse is expected.

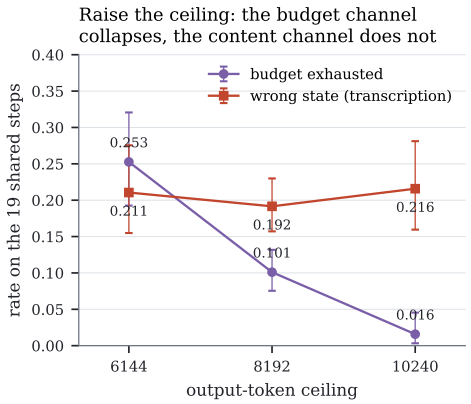

Figure 4: Raising the output ceiling removes most budget exhaustion but leaves state-realization error unchanged. The plot uses the 19 steps shared by three cells of the same model and prompt, with exact 95% intervals. Values in `paper/figures/fig3_budget_artifact.csv`.

D Additional scale results

Deliberation moves the mass off move selection

This first dissociation needs only one model. We hold the model and the task fixed and turn thinking on, on the full 80-step path at $N{=}8$:

	`wrong_move`	`wrong_state`	`budget_exhausted`
thinking off, ceiling 2048, $n{=}2000$	0.602 [0.580, 0.623]	0.118 [0.104, 0.133]	0.059 [0.049, 0.070]
thinking on, ceiling 6144, $n{=}800$	0.004 [0.001, 0.011]	0.080 [0.062, 0.101]	0.186 [0.160, 0.215]

Move selection collapses by two orders of magnitude and state-realization does not. We report this as a dissociation of channels and not as a measured decline from 0.118 to 0.080, because the second figure is a censored floor: the thinking-on cell truncates 0.186 of its samples, the two arms sit at unequal ceilings, and truncation here is informative censoring. A single mixed error rate would show $0.779 \to 0.270$ and would say nothing about which failure had moved.

The per-cell composition values and the paired contrasts

$D(m, N)$	$N{=}4$	$N{=}6$	$N{=}8$	$N{=}10$	$N{=}12$
Qwen3-1.7B	0.081	0.339	0.347	0.557	0.418
Qwen3-4B	0.667	0.333	0.833	0.845	0.719
Qwen3-8B	undefined	1.000	0.941	0.860	0.674

paired ΔD	$N{=}4$	$N{=}6$	$N{=}8$	$N{=}10$	
8B − 1.7B	n/a	**+0.661** [+0.542, +0.770]	**+0.594** [+0.350, +0.766]	**+0.303** [+0.112, +0.452]	**+0.257**
4B − 1.7B	+0.586 [-0.108, +0.976]	−0.006 [-0.364, +0.409]	**+0.486** [+0.260, +0.612]	**+0.287** [+0.125, +0.435]	**+0.301**
8B − 4B	n/a	+0.667 [+0.231, +1.000]	+0.108 [-0.142, +0.391]	+0.015 [-0.097, +0.114]	−0.045

The dominant channel at each model's own horizon

model	N^*	`wrong_move`	`wrong_state`	dominant channel
Qwen3-1.7B	4	**0.170** [0.121, 0.229]	0.015 [0.003, 0.043]	move selection, 11:1
Qwen3-4B	8	0.030 [0.011, 0.064]	**0.150** [0.104, 0.207]	state-realization, 5:1

A THIRD CHANNEL AT THE TOP, REPORTED AS A PHENOMENON AND NOT A RATE

One observation sits outside the ladder and we keep it there. Qwen3-235B-A22B-Thinking, on three of the 8B's hardest steps at a 32768 ceiling: `wrong_state` 0 observed in 30, exact upper bound 0.116; `budget_exhausted` 0 observed in 30 with zero truncation; `non_compliant` 5 of 30, 0.167 [0.056, 0.347].

That last figure is a phenomenon and not a rate: five events, four of them at a single step, and we never quote it without its n and its interval. It is not a fourth rung and is not used as one: these steps were screened on the 8B and transplanted, which is exactly the selection confound the three-model ladder was built to avoid, and no horizon is located for this model under a comparable protocol.

What the samples contain matters more than their rate. The model writes an essay and a program, narrates the correct answer in prose, and emits as its only object a template `{'move': selected_move, 'new_state': new_state}` bound to identifier names. In 8 of 12 such completions the correct board is present in the text as a literal. The model solved the step and declined to state it in the requested shape. A harness that pools this with running out of tokens scores it as incapacity, and an earlier version of our own analysis did exactly that and reached a conclusion we have since retracted, which is the concrete reason the taxonomy separates the two unusable-answer channels rather than reporting one `invalid` rate.

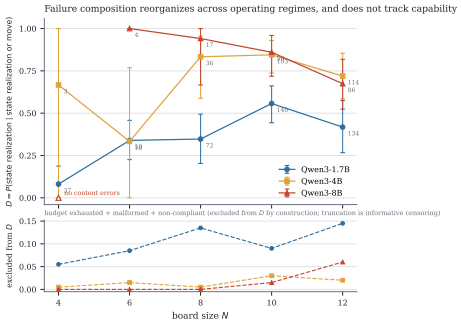

Figure 5: D against board size for the three models, with 95% step-cluster bootstrap intervals and the content-error count annotated at every point. The 8B's $N{=}4$ cell is $200/200$ correct, so D is undefined and drawn open rather than imputed. The lower strip shows the channels that D excludes by construction (budget exhaustion, malformed output and non-compliance), so readers can see how large they are. Values in `fig_failure_composition.csv` and `fig_failure_composition_excluded.csv`.

D.1 AN INTERVENTION CAN BE UNMEASURABLE, NOT MERELY WEAK

An intervention aimed at one channel does not weaken smoothly as the channel thins out. Past a point it stops being measurable at all, and the two outcomes look identical in a results table unless someone computes the difference first. Targeting a halving of state-realization error at $\alpha = 0.05$ with power 0.80, by exact enumeration of the joint binomial, fixes a minimum resolvable baseline rate of $\mathbf{0.087}$ at $n{=}500$ per arm, written down as a number before any state-realization aggregate was built from the ladder. **At the smaller model's own horizon the target channel is** 0.015, three events in 200 samples. **So the arm was dropped before it was run, and that is the result**: not that the intervention fails at 1.7B scale, nor that the result was null, since a null implies an interval excluding a large effect and there is no interval (Appendix E).

D.2 THE HORIZON FITS

model	$N^*(0.30)$	$H_{0.30}$	AUC	logistic fit
Qwen3-1.7B	4	5.02 [4.14, 5.77]	0.452	logit $e(N) = -2.645 + 0.358N$
Qwen3-4B	8	8.86 [8.24, 9.56]	0.715	logit $e(N) = -5.158 + 0.486N$
Qwen3-8B	10	10.55 [9.93, 11.34]	0.846	logit $e(N) = -7.442 + 0.625N$

D.3 Per-cell values, the dissociation, and the frontier observation

The per-cell D values and the full paired-contrast table, the thinking-on/off dissociation that fixes the regime we measure in, the channel mix at each model's own horizon, and a frontier observation at 235B that we report as a phenomenon and not a rate (five events at one step, never quoted without its n and interval) are all in Appendix D. None of them changes the reading above; they are the evidence behind it.

D.4 Matching on overall error, as a robustness check

The obvious objection to the composition result is that fixed N is fixed *difficulty*, not a fixed point on each model's own capability curve: perhaps composition only looks reorganized because the 1.7B is being read near collapse while the 8B is read in comfort. We therefore re-select, for each model, the rung whose cell error is closest to a common target q, and recompute D there.

This is a sensitivity analysis and not a discovery. The cell means were inspected before it was designed, so the rungs it selects were foreseeable; what was fixed in advance is the selection rule, the target grid, and the treatment of uncertainty. In particular the matching is *redone inside every bootstrap replicate*, because the rung is chosen using an estimated error rate and conditioning on that choice would report an interval that is too narrow.

At the primary target $q{=}0.25$ the rule selects the 1.7B at $N{=}4$ (error 0.240, $D{=}0.081$), the 4B at $N{=}8$ (error 0.185, $D{=}0.833$) and the 8B at $N{=}10$ (error 0.265, $D{=}0.860$); each rung is the one chosen in 87% to 96% of replicates. The contrasts against the 1.7B are $+0.752$ [+0.071, +0.898] for the 4B and $+0.779$ [+0.521, +0.919] for the 8B, and the 8B$-$4B contrast is $+0.027$ [-0.161, +0.689]. Both larger-model contrasts stay positive at all five targets in the grid.

So the reorganization survives matching, and so does the flat middle step. The robustness check corroborates the result as scored; it does not upgrade it, and no outcome of a post-hoc analysis on the same corpus could.

E Measurability, in full

The design fixes a minimum resolvable baseline. Our contrast targets a halving of state-realization error at $\alpha = 0.05$ two-sided with power 0.80, deciding by whether the difference interval excludes zero, with power obtained by exact enumeration of the joint binomial rather than a normal approximation. That target and those parameters fix a threshold on the baseline rate: a cell whose state-realization error is below it cannot yield a resolvable halving at that n, whatever the intervention does.

per-arm n	minimum resolvable baseline	cost at the measured rate
250 (10 steps $\times$ 25)	0.162	3.4 GPU-hours
500 (20 steps $\times$ 25)	$\mathbf{0.087}$	6.8 GPU-hours

Both thresholds were written down as numbers before any state-realization-only aggregate was built from the ladder. At the smaller model's own horizon the target channel is 0.015: three events in 200 samples, power to resolve a halving $\mathbf{0.079}$ at $n{=}250$ and $\mathbf{0.171}$ at $n{=}500$ (Figure 6). Doubling again does not rescue it; the rate is below the threshold at any n this project could buy. A second criterion, registered separately, also rules out this arm. At every 1.7B rung where state-realization error *is* measurable ($N \geq 6$), between 70% and 100% of steps count as hard. The model fails broadly there, not at a few identifiable steps. The only rung where selecting steps is meaningful is $N{=}4$, and there the rate is 0.015.

A dropped arm can be written up three ways, and two of them are false.

- *The intervention does not work at 1.7B scale.* False. Nothing was measured.
- *The result was null.* False, and worse, because a null implies an interval that excludes a large effect. There is no interval; the experiment could not have produced evidence either way.

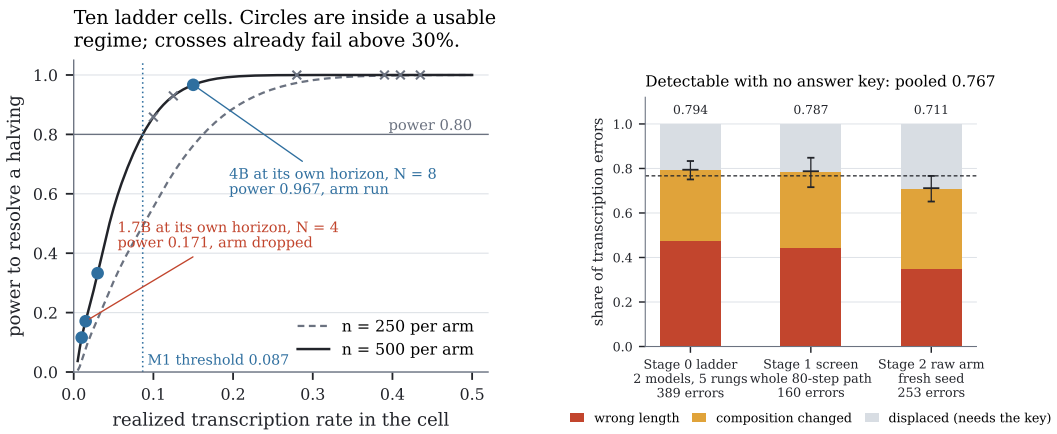

Figure 6: Left (Section D.1): power to resolve a halving of state-realization error against the rate each of the ten ladder cells realized. Circles are cells inside a usable regime; crosses are cells already failing above 30% overall. Only one circle clears the threshold. Right (Appendix G): composition of state-realization errors in three independently seeded sample sets; the lower two bands are boards that could not be any state of the task. Values in `fig4_measurability.csv`, `fig4_power_curve.csv` and `fig6_invariant.csv`.

- *At the scale where this model still has a usable regime, the channel the intervention targets is nearly absent, so the question is unanswerable within this budget.* This one.

The same shape has already occurred on a different axis. A second task, long addition at $N{=}12$, produced 0 state-realization errors in 1200 samples, exact upper bound 0.003. The encoding intervention had nothing to move there either. Task dependence and scale dependence of the target channel are the same category of finding, and neither is evidence about the intervention.

F THE INTERVENTION CORPUS, IN FULL

WHY ONLY THREE CELLS IN THE RESIDUAL MATRIX CARRY AN ADJUSTED p

Counting the action-only column, the full matrix has 28 cells, and testing all of them at face value would produce stars everywhere. Only the three predeclared target channels are treated as hypotheses; the rest is effect estimation. The action-only arm's predeclared target is executed-system correctness, which is not one of the six taxonomy channels, so it has no cell in the matrix and is evaluated in Appendix J instead against its own registered comparator and criterion.

AT THE 8B, ON THE 8B'S OWN HARD STEPS

Four encodings at a shared ceiling of 13312, on the eight steps where the unmodified encoding fails most, $n{=}200$ each. Transcription error is 0.405 $[0.336, 0.477]$ for raw, 0.115 $[0.074, 0.168]$ for alternation-only, 0.050 $[0.024, 0.090]$ for typed, and 0 observed in 200 for position-labelled, exact 95% upper bound 0.018. Truncation is 0.005 for three arms and 0.020 for alternation-only, so we do not claim 0.005 for four. On the five control steps, where raw is already clean at 0.024 $[0.005, 0.069]$, alternation-only is *worse* at 0.160 $[0.101, 0.236]$, the cost of an encoding that must be recomputed on every move. Full five-channel table in Appendix K.

Three encodings that share only the property of making a one-cell displacement locally detectable each remove most of the error, which rules out absolute position and stable identity as *necessary* without identifying the mechanism. This cell also has a selection problem the 4B measurement does not: its eight steps were chosen for being hard in the raw arm and all four arms were then measured on those same steps, which is why the 4B result and not this one is the central measurement.

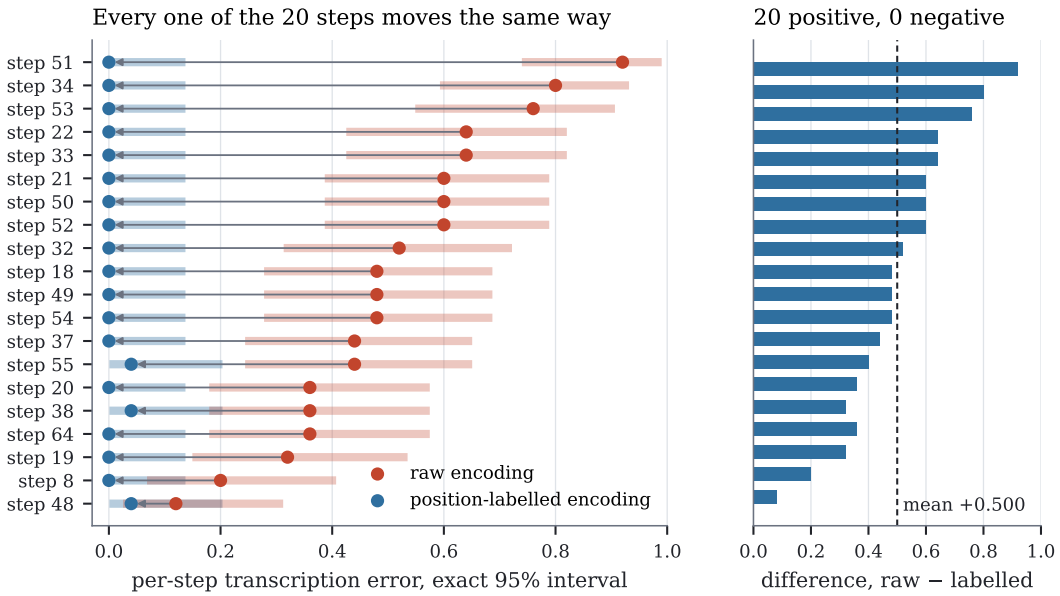

Figure 7: Qwen3-4B at its own horizon, on its own screened steps, measured on a third independent seed. Left: per-step state-realization error under each encoding with exact 95% intervals, steps ordered by the raw arm. Right: the within-step difference. Values in `fig1_paired_steps.csv`.

The format defect the position labels create, in full

All 20 non-compliant samples in the position-labelled arm are one defect. The model writes the empty cell's label and drops the closing quote, so `'14:_'` becomes `'14:_,` so the state literal will not parse while the move literal beside it parses fine. `finish_reason` is `stop` for all 20 and none is truncated, so it is not budget. The same sequence appears in 30 of the 480 samples that *did* parse, there inside the reasoning trace. The arm's only three state-realization errors are the same defect milder: a bare `'_'` with the label omitted, so the state does not decode to a board at all.

For the glyph arm the mechanism is identified rather than inferred: conditional on what each answer wrote, 453 answers using `E` produced 0 non-compliant samples and the 10 that reverted to `'_'` produced 3, and every non-compliant sample in the arm carries an unterminated `'<i>:_,`.

Two further facts about the single-side arms

A safeguard that did not fire. A single-side arm shows the model one notation and asks for the other, so it may echo the notation it was shown where the plain form was requested a sample with the task right and the notation wrong. We registered before either arm existed that such a sample is graded `wrong_state` and *additionally flagged*, never rescored correct, and that ρ is reported twice: on the channel as graded, and with flagged samples excluded. The observed rate is 0 of 500 in both arms, so the two are identical, 0.348 and 0.940 either way, and both select the same branch of the map. Registering the check was still the right call: ρ is computed on exactly the channel an unflagged notational echo would inflate, and the check could only have been added honestly before the data existed.

What would overturn this. A third arm whose output labels are deliberately *wrong* would separate the index from the mere act of emitting a per-cell token: if ρ stayed near 0.940 there, the mechanism would be the emission and not the index it carries. We did not run that arm.

The token accounting

Both arms use the same output-token ceiling of 13312, so their truncation rates are comparable. They did not use equal tokens. The position-labelled arm generates $4{,}526$ mean completion tokens per

sample against raw's 5,583, 18.9% fewer. It truncates at 0.012 against 0.058, and its prompts are 8.0% longer. We therefore describe the arms as sharing a ceiling, never as using equal compute.

F.1 Which side of the interface carries the label effect

The labelled encoding changes two things at once: the board the model reads and the board it must write. Two orthogonal single-side arms separate them, each 500 samples on the same 20 screened steps, with the recovery fraction, its threshold and the outcome map registered before either arm was sampled. **Labelling what the model must *write* recovers** 0.940 **[0.905, 0.970] of the effect; labelling what it *reads* recovers** 0.348 **[0.202, 0.477].** Under the outcome map fixed in advance that is the *output alignment* branch: the error is not mainly a failure to locate the changed cell but a failure to hold position while emitting a long list, a statement about generation, not perception. It is not a cheaper encoding and we do not net the cost out: the output-only arm has the highest `wrong_move` rate of any arm in this corpus, 0.080 against 0.014 (Appendix F).

Table 1: Residual matrix: change in each channel's rate, in percentage points, against each intervention's own matched baseline. ⋆ is the channel predeclared as that intervention's target; those three carry Holm-corrected p, every other cell is effect estimation with no adjusted p. CREATED marks a zero baseline, where a ratio is undefined; † marks a derived quantity. Values in `fig_intervention.csv`.

baseline →	position labels raw	glyph _→E position-labelled	constrained decoding raw, same 5 steps
`wrong_move`	−3.4	−1.0	−2.4
`wrong_state`	**−50.0** ⋆	+0.6	−2.4
`non_compliant`	CREATED	**−3.2** ⋆	0.0
`malformed`	0.0	0.0	0.0
`budget_exhausted`	−4.6	−0.2	+1.6
`invariant_violation`†	−36.0	CREATED	**−22.4** ⋆
`invariant_unassessable`†	CREATED	−0.6	0.0
step clusters	20	20	5

F.2 The staged arm table

	`correct`	**`wrong_state`**	`non_compliant`
raw	0.388 [0.345, 0.432]	**0.506** [0.461, 0.551]	0 of 500, upper 0.007
position-labelled	0.928 [0.902, 0.949]	**0.006** [0.001, 0.017]	**0.040** [0.025, 0.061]

F.3 The single-side arms, in full

The labelled encoding changes two things at once: the board the model reads and the board it must write. Two orthogonal single-side arms separate them, each 500 samples on the same 20 screened steps with the empty-cell glyph held at _ so the label mechanism is not confounded with the glyph effect above. The recovery fraction $\rho_{\text{side}} = (p^{\text{state}}_{\text{raw}} - p^{\text{state}}_{\text{side}})/(p^{\text{state}}_{\text{raw}} - p^{\text{state}}_{\text{indexed}})$, its threshold and the map from outcome to conclusion were registered before either arm was sampled; neither had ever been run.

arm	input	output	**`wrong_state`**	ρ
raw	plain	plain	0.506	n/a
`labels_input_only`	labelled	plain	0.332	**0.348** [0.202, 0.477]
`labels_output_only`	plain	labelled	0.036	**0.940** [0.905, 0.970]
position-labelled	labelled	labelled	0.006	1 by construction

Labelling what the model must *write* recovers 0.940 **of the effect; labelling what it *reads* recovers** 0.348**.** Under the outcome map fixed in advance that is the *output alignment* branch, selected by the rule rather than after looking; the 0.8 bar is a predeclared interpretive threshold, not a null-hypothesis cutoff, and intervals are given so a reader preferring another bar can apply it. The error is therefore not mainly a failure to locate the changed cell: it is a failure to hold position while emitting a long

list, since an index written beside each cell *as the model produces it* prevents the drift while the same index on the input leaves two thirds of the error in place. That is a statement about generation, not perception.

It is not a cheaper encoding, and we report the cost rather than netting it out. The output-only arm has the highest `wrong_move` rate of any arm in this corpus, 0.080 against the labelled arm's 0.014, and its overall correctness of 0.848 is below the labelled arm's 0.928. It isolates a mechanism; it does not replace the full encoding. A pre-registered safeguard against the obvious confound (a model shown labelled cells echoing them where plain output was asked) recorded 0 of 500 in both arms, and ρ is identical with such samples excluded. Per-sample accounting is in Appendix F.

F.4 Horizon, as a geometric equivalent

Per-step rates are hard to compare against intuition, so we convert them, and label the conversion, because it is an assumption and not a measurement. Writing

$$H_{\text{geo}} \;=\; \frac{1}{1 - p_{\text{correct}}},$$

the raw arm sits at 1.6 [1.5, 1.8] expected steps before its first error, the position-labelled arm at 13.9 [10.2, 19.6], the glyph-fixed arm at 29.4 [18.6, 50.2], and constrained decoding at 2.1 [1.8, 2.6]. Intervals come from transforming each bootstrap draw of p_{correct}.

H_{geo} is a *geometric-equivalent* horizon, not a measured run length. Reading it literally assumes a constant and independent per-step failure hazard, and no trajectory was run. We report it because trajectory success is uninformative here. The canonical path at $N{=}8$ has 80 steps, so even the glyph-fixed arm's 0.966 per-step correctness gives $0.966^{80} \approx 0.06$. The position-labelled arm gives $0.928^{80} \approx 2.5 \times 10^{-3}$, and raw gives 1.3×10^{-33}.

G The invariant: what establishing it costs, and where it does not apply

Most of these failures are not silent. We call a failure *silent* if neither interface validation nor a task-invariant check detects it. On this task most state-realization errors are not silent. The board's symbol multiset (N reds, N blues, one empty) is invariant along the canonical path, so a claimed board with a different length or multiset cannot be any state of this task, which one pass over the claimed state establishes with no answer key. Classified by rules needing no judgment, 0.767 [0.736, 0.796] of state-realization errors violate the invariant (802 pooled over three seeded sets whose rates differ: 0.794, 0.787, 0.711), which is what makes the reallocation signature in Figure 3A observable: one sub-mode empties while the other rises. The invariant belongs to this task's state space and we have not demonstrated an analogue elsewhere (Appendix G).

The verification cost. We verified the conserved multiset at every state of every path used here: 24, 48, 80, 120 and 168 states at $N = 4, 6, 8, 10, 12$, one distinct multiset each. Establishing that a claimed board violates it costs one pass over the claimed state: no answer key, no simulation of the trajectory, no reference implementation of the rules, no learned verifier. It is linear in the size of the state and constant in the length of the path. The three state-realization errors of the position-labelled arm do not decode to a board at all, which is why its genuinely silent share is 0 observed in 500.

Where it does not apply. The general condition is stated rather than shown: where the state space has a conserved quantity that is cheap to evaluate, a large fraction of execution error is detectable without an oracle. We checked the one other task in our corpus that could have carried the analysis, and it cannot: on Tower of Hanoi, state-realization error re-graded with the current reader is 0.0108 [0.0063, 0.0172] at $N{=}6$ and 0.0038 [0.0024, 0.0056] at $N{=}8$, roughly eight times below the rate at which our design could resolve anything, so we classified nothing. That is not an absence of evidence but an instance of the pattern: Hanoi's state assigns a unique numerical identifier to each disk, which is the intervention of Section 6 already built into the task, and the task that has it shows almost no state-realization error to classify. Long addition cannot test it either, having produced 0 state-realization errors in 1200 samples.

H LIMITATIONS, IN FULL

The ladder and intervention results rest mainly on one family. The scale ladder, the residual matrix and the prediction test use Qwen3 only. The reallocation pattern is repeated on one other family, DeepSeek-R1-Distill-Llama-8B (Section 7.1), and composition at the failure boundary is measured on nine families (Section 5.4). For the ladder and the interventions, an effect specific to Qwen3's architecture or training recipe is therefore not ruled out. A direct test is small: run the same three stages (ladder, screen, contrast) on a 4B-class model from another family, about $1{,}800$ samples and roughly 12 GPU-hours at this project's measured rate. If the horizon and the dominant channel land in the same place, the ladder result is about model scale. If they do not, it is about Qwen3. That replication also needs a check on completion tokens. A family whose default reasoning emits a few hundred tokens, against Qwen3's roughly five thousand, would be measuring a different regime.

Three rungs, and the top step of the ladder is flat. The scale dimension has three points because 24 GiB cards will not hold Qwen3-14B or 32B, so the ladder stops where the hardware does. Within those three, the composition result is carried by the largest capability gap: $4\text{B} \rightarrow 8\text{B}$ is flat at every rung where we can resolve it, and separates at exactly one rung, $N{=}6$, where the 8B has *four content errors in total* and $D{=}1.000$ with an exact interval of [0.398, 1.000]. **We do not know whether that flatness is the onset of a plateau or a local feature of these two models**, and a fourth rung is the single observation that would most change the reading. A local 32B ladder under the same protocol has not been run, so we list it as future work. More samples at the existing rungs would tighten $N{=}6$ but cannot produce a monotone chain: the point estimates at $N{=}8, 10, 12$ are $+0.108$, $+0.015$ and -0.045, and two of the three have the wrong sign for one.

The matched-error analysis is a robustness check and nothing more. The cell means were inspected before it was designed, so the rungs its rule selects were foreseeable. What was frozen in advance is the selection rule, the target grid, and the decision to re-match inside every bootstrap replicate. It supports the result as scored but cannot upgrade it. No post-hoc analysis on the same corpus could, and the program prints this in its own output.

The intervention corpus is one model at one operating point. The residual matrix is measured on Qwen3-4B at $N{=}8$, on twenty screened hard steps. It is not a claim about interventions in general, and the three regimes we *decline* are part of why: their target channel is too rare for any affordable sample size to resolve a halving, so we report them as not measurable rather than as measured nulls. "We could not resolve this" and "there is no effect here" are different statements, and our data support only the first.

The constrained-decoding column rests on five step clusters. Its guarantee is exact and its reallocation signature has a clear mechanism. But five clusters allow only 2^5 sign assignments, so no randomization p below about 0.06 is possible, and the interval on its total state-realization effect spans 0.32. We present that column as a mechanism and never as an inferential claim, and we did not extend the arm: the question it would buy precision on is already decided, since even a resolved reduction of 0.05–0.10 leaves it an order of magnitude behind the notation change.

One task carries the headline, and one earlier cell has a selection problem. Appendix G is a Checkers result. Hanoi sits in the same corpus with the same interfaces and grading, and we have now classified it: all six channels are computable on $7{,}950$ samples, so the decomposition itself is not Checkers-specific. What that corpus can support is limited: it exists only thinking-OFF at a 2048 ceiling, so it is compared against the regime-matched thinking-OFF Checkers cells rather than against the ladder, there is no Hanoi ladder across models, and the cells are not paired. It is therefore a demonstration that the instrument transfers, not a second-environment test of Section 5. One result from it is worth stating, because it is the methods claim of Section 4 made concrete: under an identical harness, model and ceiling, the two tasks fail through *opposite* unusable-answer channels. Hanoi shows `malformed` at 0.020–0.025 with budget exhaustion at exactly 0.000, Checkers shows budget exhaustion at 0.042–0.059 with `malformed` at exactly 0.000. A single pooled `invalid` rate would report these as the same failure. Both tasks are move-dominant in that regime, which locates the state-realization dominance this paper builds on as a property of the thinking-ON regime rather than of the task. Long addition cannot test any of this, having produced zero state-realization

errors in $1{,}200$ samples, so that second-domain line is closed rather than pending. The 8B four-arm cell selected its steps in the arm it then measured; the 4B result does not, which is why it and not the 8B cell is the central measurement. The 235B figures are five events at one step, reported as a phenomenon with their interval and never as a rate.

The mechanism is identified on one axis only. The two orthogonal single-side arms registered for this purpose have now run, and they do separate the sides: labelling the board the model must write recovers 0.940 $[0.905, 0.970]$ of the indexed reduction in state-realization error, labelling the board it reads recovers 0.348 $[0.202, 0.477]$, and the frozen outcome map selects output alignment (Section F.1). Three things it does *not* settle. It does not say *what* the label does: an arm whose output labels are deliberately wrong would separate the index from the act of emitting a per-cell token, and it was not run. It does not make the input side null, since labelling the input alone still recovers about a third of the effect, so "output alignment dominates" is a statement about shares and not an exclusion, and the two fractions are not additive. And it does not come free: the output-only arm has the highest `wrong_move` rate of any arm in this intervention corpus, so the mechanism is isolated at a cost we report rather than net out. The observational version of the test is still unavailable on this task: the natural covariate is collinear with position along the optimal path, and a position-matched, adequately powered version refuted the hypothesis it was built to confirm. **The mechanism claim rests entirely on within-step manipulation and has no observational support here**, and like every intervention result in this paper it is one model at one operating point. Separately, we have no measured bound on how much of the residual a controller with perfect error detection could recover; the decision not to build one rests on a kill criterion that fired, which is a weaker argument than a measured ceiling.

Thresholds are choices. The 0.30 bar defining the competence horizon is a judgement, not a fact about the task, and is labelled as such everywhere it appears; we report the crossing across the whole family $q \in \{0.20, \ldots, 0.40\}$ and a threshold-free area precisely so no conclusion depends on it. The same applies to the 0.8 recovery-fraction threshold in the mechanism ablation, which is an interpretive cut and not a hypothesis test, and to the three-percentage-point non-inferiority margin for the interface arm.

Truncation is informative censoring. Cut-off samples are disproportionately the harder ones, so any rate measured under material truncation is a lower bound and conditioning on the samples that answered under-corrects. We quote every rate with its ceiling, audit the censoring, and report a worst-case bound, but that bound holds the counts fixed and attributes censoring adversarially, so it answers only what censoring alone could do. Where it looks tighter than the bootstrap, that is not evidence a sign is established, and we never report it as inference.

Intervals are within-run. They describe sampling error given one run and exclude run-to-run variability by construction. The measured component is $1.38\times$ sampling with an interval containing 1, so no correction is applied; the honest upper bound on the variance ratio is $2.20\times$, about $1.5\times$ on a half-width, and that estimate is substantially carried by a single step.

I Related work, in full

LEAD (Pushkin & Abbe, 2026) supplies the task, the prompt text and the framing we argue with; its per-step formulation is what makes any of this measurable, since a failure at step k is not the accumulated consequence of a failure at step $k-1$.

Failure taxonomies and diagnosis, which we do not claim as novel. Barke et al. (2026) localise the critical failure step over a grounded-theory failure taxonomy, using an LLM-based judge; Wang et al. (2026) build a cross-domain diagnostic study with a human-validated LLM-as-judge pipeline; Zhu et al. (2025) define a modular taxonomy spanning memory, reflection, planning, action and system operations and pair it with corrective feedback; and Zhu et al. (2026) continue that line into automated recovery. **We therefore claim no taxonomy, no diagnosis method, and no first coupling of diagnosis to repair.** What differs is the *kind* of evidence: those studies are trajectory-level, and the two whose abstracts describe their attribution use an LLM judge, while ours is per-step and adjudicated by an exact simulator with no model in the loop, on a task where the correct action set and the correct resulting state are both decidable. That makes channel membership a property of the

run rather than of a judge, which is what lets an intervention's effect on one channel be attributed rather than inferred.

The interface as the locus of failure, which is also not novel. Zhang et al. (2026) separate prompt-formatting cost, protocol overhead and genuine tool gain, name the residue a "tool-use tax", and propose an inference-time gate; Xu et al. (2026) argue that many practical agent failures are addressed by evolving the runtime interface around a frozen model; Rozanov & Rei (2024) have the model track state explicitly through a chain of states, and Chae et al. (2024) have an agent choose actions using outcomes simulated by a learned world model. **"Some agent failures are interface failures rather than capability failures" is established** (Zhang et al., 2026; Xu et al., 2026), and our action-only arm, in which the simulator computes the resulting state, applies that factorisation rather than introducing it, run for completeness and reported against the strongest representation intervention we have, not against the baseline that would flatter it. This literature also contains the published form of the controller we declined: the gate of Zhang et al. (2026) is structurally the channel-aware router our own data argues against.

Measurement validity, where our discipline is shared rather than distinctive. Messing (2026) report naive standard errors 40–60% smaller than ones corrected for measurement error in LLM annotation, evaluation and benchmarking pipelines, a related concern about understated uncertainty to the within-step dependence our step-cluster resampling addresses; Tu et al. (2026) formalise benchmark-validity checks; Thomas et al. (2026) treat pre-registration as the remedy for analytic flexibility. **We claim no novelty for auditing measurement validity.** We do claim that *measurability decides which arms are worth running*: three intervention regimes here are declined because their target channel is too rare to resolve at any affordable sample size, reported as part of the result rather than as a gap. We found generic power analysis and minimum-detectable-effect practice, but not this use of per-channel measurability to drop a specific arm at a specific operating point.

Constrained decoding, long-horizon scaling, and state representation. Grammar-aligned decoding (Park et al., 2024) supplies the mechanism, and a measured suppression of tool calling under structured-output constraints in open-weight models (Li et al., 2026) shows that constraining output is not free; our contribution is not the method but the shape of its effect, a null on a total target from a method whose guarantee is sound. Work asking what degrades as a task gets longer reports one accuracy curve per model (Sinha et al., 2026; Zhou et al., 2025; Meyerson et al., 2025); our result is orthogonal to whether those curves are right, because the curve is a mixture and two models with similar curves can be failing for different reasons where the curves cross. Shojaee et al. (2025) and the reply of Opus & Lawsen (2025) argue about whether an observed collapse is capability or format; the channel split makes that a measurement rather than an argument, and at 235B we find both. Positional and retrieval effects in long inputs (Liu et al., 2024; Du et al., 2025; Veseli et al., 2025; Li et al., 2024) concern finding information in a context, whereas here the state is short, present and adjacent to the answer and the model still writes it wrong; length-generalization work (Zhou et al., 2023b) makes the closest prediction, that the difficulty lies in the representation of the intermediate object, and our encoding arms test it at prompt level rather than at training level.

Decomposition, control, and heterogeneity. Least-to-most prompting (Zhou et al., 2023a), plan-and-solve (Wang et al., 2023), tree search (Yao et al., 2023), algorithmic composition (Zelikman et al., 2023) and the agent-orchestration line (Liu et al., 2025; Zhang et al., 2025; Wang et al., 2025) address long-horizon failure by adding control. Appendix G argues that on tasks with a conserved quantity the verification step of such a loop can be a single pass over the state rather than a model call, and Section D.1 that the prior question, which channel a controller would be catching and at which scale, is usually not answered before it is built. Hosseini et al. (2024) and Press et al. (2023) show that models with equal aggregate scores differ in what they get wrong, and Jiang et al. (2025) and Aghajohari et al. (2025) study how deliberation is allocated. We add a scale axis and a measurement discipline: read each model at its own competence horizon rather than at a shared difficulty, and the dominant channel differs by enough that an intervention can be unmeasurable at one end of the range it is claimed for.

Where that leaves the contribution. Not a taxonomy, not diagnosis-to-repair, not the interface thesis, and not measurement auditing, since each has 2026 prior art named above. What we claim is narrower: on a task where action correctness and state realization are separately and exactly decidable, the composition of per-step failure reorganizes across operating regimes rather than tracking capability, and the interventions we tested act on that composition selectively, nearly eliminating one mechanism,

leaving another untouched, and creating a third from nothing with measurability deciding in advance which of those effects can be resolved at all.

J THE ACTION-ONLY INTERFACE ARM: A PRE-REGISTERED NEGATIVE RESULT

The pre-registered decision rule placed this arm in the appendix. Comparator, non-inferiority margin, Pareto criterion, the secondary endpoints and the placement rule itself were written down before any action-only sample existed (no run directory, script or config for it was in the repository when the rule was frozen), and the rule says placement follows the outcome and not the outcome's attractiveness. It lost, so it is here, and it is here in full rather than in a file drawer: a pre-registered loss constrains the interface-failure account, which is the reason the arm was worth running.

Design. The joint-output interface asks the model for an action *and* the resulting state. The action-only interface asks for the action, and the deterministic simulator computes the transition: $s \to \text{LLM} \to a \to T(s,a) \to s'$. Qwen3-4B, the same 20 pinned hard steps at $N{=}8$, $n{=}25$ per step, 500 samples, all of them usable. The input is the *raw* unlabelled board, so the arm carries the interface change and not the notation change, which is what makes raw, position-labelled and action-only three arms on one baseline rather than a nested ablation of the labelled arm. The output ceiling stays at 13312: lowering it because the answer format got shorter would entangle an interface intervention with a budget intervention and move the `budget_exhausted` channel the whole taxonomy is defined against, so the token saving is *measured* and not assumed. Both `model_action_correct` and `system_executed_state_correct` are logged per sample, so a reader can see that the architecture changed rather than that the parser got more forgiving.

The comparator is the labelled arm, and was fixed in advance. Raw, at 0.388, is contextual only: the review that proposed this arm anchored on a ceiling of 0.894 that the position-labelled arm already exceeds, so raw is not an admissible headline comparator here. The frozen rule is explicit that a loss to the labelled arm is not to be rescued by re-benchmarking against raw, and we do not report that contrast.

	position-labelled (comparator)	action-only
executed-system `correct`	0.928	0.902
`wrong_move`	0.014	0.074
`non_compliant`	0.040	0.000
`malformed`	0.000	0.000
`budget_exhausted`	0.012	0.024
`wrong_state`	0.006	0.000, structural
mean completion tokens	4,526	5,397

Verdict: inconclusive or negative. The paired difference in executed-system correctness is -0.026 [-0.094, +0.032] over the 20 shared step clusters, with a one-sided lower bound of -0.082. Accuracy superiority fails, because the two-sided interval includes zero. The Pareto criterion fails on *both* of its conditions: the one-sided lower bound of -0.082 sits below the -0.03 margin, and completion-token use did not fall by the required 20%: it rose (Figure 8). **The registered conclusion is that a notation change dominates an interface change.** That is a constraint on the hypothesis that these are interface failures, and it is worth printing even though it is not the result the arm was hoped to give.

The structural zero is not a finding. `wrong_state` is 0.000 in this arm, and it could not have been anything else. The environment computes $T(s,a)$, so the model cannot make a state-realization error. The zero is an identity. The analysis output flags it as structural, and it is excluded from the confirmatory family. Reporting it as the elimination of state-realization error would make the experiment vacuous, and the frozen protocol explicitly forbids that.

The tokens went the wrong way, and that is the informative part. The answer the model must produce is far shorter (a move, instead of a move plus a 17-cell board), yet the arm generates *more* completion tokens: $5{,}397$ mean against the position-labelled arm's $4{,}526$, a paired difference of $+871.1$ [+546.7, +1,229.7] on the same 20 clusters. The frozen Pareto condition required a reduction

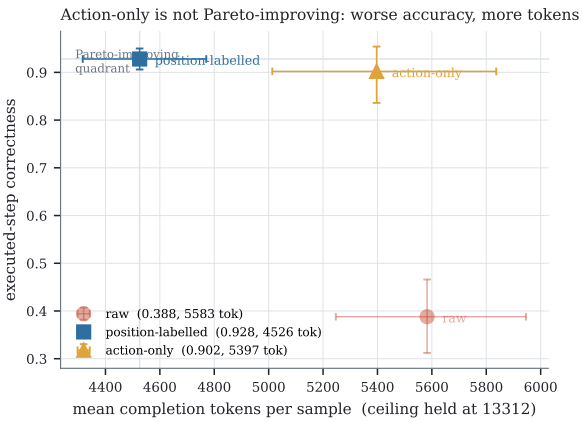

Figure 8: The action-only arm against its registered comparator on the two axes the Pareto criterion is stated over. It sits below and to the right of the position-labelled arm: lower executed-system correctness and more completion tokens, so neither condition of the criterion is met and the arm does not dominate on either axis. Values in `fig_action_only_pareto.csv`.

of at least 0.20. The measured change is -0.1925, which means 19.2% *more* tokens. The ceiling was held constant, so this is a measured change and not a budget artefact. The saving the interface argument predicts did not materialize in the reasoning trace, which is where this model spends nearly all of its completion tokens, and shortening the answer does not shorten the trace.

The arm also relocates two other channels. `non_compliant` goes from 0.040 to 0.000 (the labelled arm's format defect is a property of writing labelled cells, and an arm that writes no board cannot have it) while `wrong_move` goes from 0.014 to 0.074 and `budget_exhausted` from 0.012 to 0.024. These are exploratory secondary endpoints, reported because the protocol requires the full composition change rather than the headline alone, and they carry no adjusted p.

K Full per-cell tables

L The model campaigns, in full

What was run. Two hosted campaigns extend the local Qwen3 ladder with the same task, the hash-pinned prompt, the raw arm and the exact grader: Qwen3-32B, Qwen3-235B-A22B, GPT-5.5 and Claude Opus 5 first, then eight further families chosen as distinct base models, reasoning-capable, and spread across capability rather than concentrated at the frontier. Each model passes a comparability gate on $N{=}4$ and 8 (at least 90% of answers bind the requested keys, at most 2% truncated), then a bracket over $N \in \{4, 8, 16, 32, 64\}$ extended up to 96 when the largest rung is still inside the bar and down to 3 and 2 when even $N{=}4$ is past it, then 36 further samples at the rungs carrying content errors. The second campaign ran at a 32,768-token ceiling and a 40-minute per-call timeout; streaming one call showed that the models timing out were reasoning slowly, not failing, and timed-out calls were verified unbilled. Output ceilings therefore differ across the thirteen models: the local Qwen3 ladder, GPT-5.5 and Claude Opus 5 ran at 13,312 tokens, and Qwen3-32B, Qwen3-235B-A22B and the eight second-campaign families at 32,768. The inclusion rule's worst-case censoring bound is what prevents a truncated answer from being counted toward either content channel.

N^*. N^* is the largest *tested* board with observed per-step error at or below 0.30, for this task, prompt and grader. It can only take grid values and is never interpolated. When the largest tested rung is still inside the bar, N^* is a lower bound.

Inclusion, by rule. We compute three readings of composition for every model: at N^*, at the first tested rung past the bar, and pooled across those two rungs. A reading qualifies if it rests on at least five content errors and its worst-case censoring bound lies on one side of 0.5; the plotted reading is

Table 2: Qwen3-4B at $N = 8$, all 20 measured steps, both arms, $n = 25$ per step per arm. Exact 95% intervals. Recomputed by `python -m scripts.paper_recompute -claim C-003`.

step	raw `wrong_state`	position-labelled	difference
8	0.200 [0.068, 0.407]	0.000 [0.000, 0.137]	+0.200
18	0.480 [0.278, 0.687]	0.000 [0.000, 0.137]	+0.480
19	0.320 [0.149, 0.535]	0.000 [0.000, 0.137]	+0.320
20	0.360 [0.180, 0.575]	0.000 [0.000, 0.137]	+0.360
21	0.600 [0.387, 0.789]	0.000 [0.000, 0.137]	+0.600
22	0.640 [0.425, 0.820]	0.000 [0.000, 0.137]	+0.640
32	0.520 [0.313, 0.722]	0.000 [0.000, 0.137]	+0.520
33	0.640 [0.425, 0.820]	0.000 [0.000, 0.137]	+0.640
34	0.800 [0.593, 0.932]	0.000 [0.000, 0.137]	+0.800
37	0.440 [0.244, 0.651]	0.000 [0.000, 0.137]	+0.440
38	0.360 [0.180, 0.575]	0.040 [0.001, 0.204]	+0.320
48	0.120 [0.025, 0.312]	0.040 [0.001, 0.204]	+0.080
49	0.480 [0.278, 0.687]	0.000 [0.000, 0.137]	+0.480
50	0.600 [0.387, 0.789]	0.000 [0.000, 0.137]	+0.600
51	0.920 [0.740, 0.990]	0.000 [0.000, 0.137]	+0.920
52	0.600 [0.387, 0.789]	0.000 [0.000, 0.137]	+0.600
53	0.760 [0.549, 0.906]	0.000 [0.000, 0.137]	+0.760
54	0.480 [0.278, 0.687]	0.000 [0.000, 0.137]	+0.480
55	0.440 [0.244, 0.651]	0.040 [0.001, 0.204]	+0.400
64	0.360 [0.180, 0.575]	0.000 [0.000, 0.137]	+0.360
pooled	0.506 [0.461, 0.551]	0.006 [0.001, 0.017]	+0.500

Table 3: The ladder in full: five channels per cell, $n = 200$ each, ceiling 13312, thinking on, raw encoding. Recomputed by `python -m scripts.paper_recompute -claim C-011`.

model	N	`corr`	`move`	`state`	`budget`	`malf`	`n.c.`
Qwen3-1.7B	4	0.760	0.170	0.015	0.055	0.000	0.000
	6	0.620	0.195	0.100	0.085	0.000	0.000
	8	0.505	0.235	0.125	0.135	0.000	0.000
	10	0.210	0.310	0.390	0.090	0.000	0.000
	12	0.185	0.390	0.280	0.145	0.000	0.000
Qwen3-4B	4	0.980	0.005	0.010	0.005	0.000	0.000
	6	0.895	0.060	0.030	0.015	0.000	0.000
	8	0.815	0.030	0.150	0.005	0.000	0.000
	10	0.455	0.080	0.435	0.030	0.000	0.000
	12	0.410	0.160	0.410	0.020	0.000	0.000

the qualifying one with the most errors. Where two readings qualify they agree for every model, so the choice among them does not decide any verdict. The rule was fixed after the first campaign's data had been seen and is reported as exploratory.

The trend, and how far it goes. Across the eleven boundary models the rank correlation between N^* and D is $\rho = 0.79$, permutation $p = 0.005$ (200,000 permutations, seeded). Dropping points under ten content errors: $\rho = 0.77$, $p = 0.013$; hosted models only: $\rho = 0.76$, $p = 0.036$; readings with no truncated answer: $\rho = 0.72$, $p = 0.048$. The separation is the stronger statement, and it needs no test: both N^*=4 models are action-dominant and all nine with $N^* \geq 8$ are state-dominant. What we do not claim is the shape of any relationship, a law across tasks, or that N^* measures capability in general.

What the excluded and unusual cases show. GLM-5.3 and Kimi-K2.6 run out of their 32,768-token budget on exactly the boards where they begin to fail. Kimi truncates 38 of 45 answers at N=64. No reading survives censoring, so measuring them would need a larger budget, not more samples. From N=16 upward, Mistral Small stops giving a move and describes an algorithm instead. We classify this as non-compliance. It is a change in how the model answers, not a parser failure.

Table 4: The 8B four-arm cell in full, on the eight hard steps and the five control steps, ceiling 13312. Recomputed by `python -m scripts.paper_recompute -claim C-014 -claim C-015`.

steps	encoding	corr	move	state	budget	malf	n.c.	trunc
hard (8)	raw	0.590	0.000	0.405	0.005	0.000	0.000	0.005
	typed	0.935	0.010	0.050	0.005	0.000	0.000	0.005
	position-labelled	0.990	0.005	0.000	0.005	0.000	0.000	0.005
	alternation-only	0.865	0.000	0.115	0.020	0.000	0.000	0.020
control (5)	raw	0.968	0.000	0.024	0.008	0.000	0.000	0.008
	typed	0.984	0.000	0.008	0.008	0.000	0.000	0.008
	position-labelled	0.992	0.000	0.000	0.008	0.000	0.000	0.008
	alternation-only	0.824	0.000	0.160	0.016	0.000	0.000	0.016

Table 5: Every model in the campaigns and whether it is plotted. Inclusion is decided by rule, not by the direction of the result. D is the state share of content errors with its exact 95% interval; the censoring bound assigns every truncated answer to either channel. *Reading* names the rungs composition is read at.

model	family	category	N^*	reading	act.	state	trunc.	D [95% CI]	cens. bound
Qwen3-1.7B	Qwen3 dense	boundary	4	R3 @[4,6]	73	23	0	0.24 [0.16, 0.34]	–
Mistral Small	Mistral	boundary	4	R3 @[4,8]	10	5	0	0.33 [0.12, 0.62]	–
Qwen3-4B	Qwen3 dense	boundary	8	R3 @[8,10]	22	117	0	0.84 [0.77, 0.90]	–
Qwen3-32B	Qwen3 dense	boundary	8	R3 @[8,12]	2	9	0	0.82 [0.48, 0.98]	–
Granite 4.2 8B	IBM Granite	boundary	8	R3 @[8,16]	14	33	1	0.70 [0.55, 0.83]	[0.69, 0.71]
Nemotron 3.5 Lightning	NVIDIA Nemotron	boundary	8	R1 @[8]	2	4	0	0.67 [0.22, 0.96]	–
Qwen3-8B	Qwen3 dense	boundary	10	R3 @[10,12]	35	101	0	0.74 [0.66, 0.81]	–
Grok 4.6	xAI Grok	boundary	16	R2 @[32]	0	16	15	1.00 [0.79, 1.00]	[0.52, 1.00]
Qwen3-235B-A22B	Qwen3 MoE	boundary	20	R3 @[20,30]	3	12	0	0.80 [0.52, 0.96]	–
MiniMax M3	MiniMax	boundary	32	R1 @[32]	0	10	1	1.00 [0.69, 1.00]	[0.91, 1.00]
GPT-5.5	OpenAI	boundary	56	R3 @[56,64]	1	25	0	0.96 [0.80, 1.00]	–
Claude Opus 5	Anthropic	not reached	$\geq$64	within comp.	0	11	0	1.00 [0.72, 1.00]	–
Gemini 3.8 Flash	Google Gemini	not reached	$\geq$96	within comp.	0	18	0	1.00 [0.81, 1.00]	–
GLM 5.3 Flash	Z.ai GLM	**excluded**	32	failures coincide with the token ceiling					
Kimi K2.6	Moonshot Kimi	**excluded**	32	failures coincide with the token ceiling					

MiniMax-M3 is state-dominant at its boundary (10 of 10 at N=32). Far past the boundary, at N=64 where 93% of steps fail, its errors lean toward wrong actions (21 of 32) under heavy truncation. A reading taken deep in collapse measures something different from a reading at the boundary, so the rule reads at the boundary. The hosted Qwen3-32B is served by a third-party provider at an unpinned quantization and measures N^*=8, level with the locally served 4B; it is plotted as hosted and not joined to the local ladder.

Grader audit. The second campaign stored each answer's text. Wrong-action cases were checked to have a move outside the reference optimal set, and wrong-state cases to have a move equal to an optimal move with a different reported board. No misclassification was found and no grade was changed.

M WHAT WAS REGISTERED, AND WHAT WAS NOT

The staged measurement of Section 6 is pre-registered: target effect, α, power, decision rule, per-arm n, seven kill criteria and the reconciliation rule for its two co-primary readings were each committed before the data they govern. The single-side ablation and the action-only arm are registered on the same terms and reported under the branch their own rule selected, for the interface arm the losing one. The prediction rule of Section 7 is the strongest case; its chronology is given there. The claims this paper leads with were not registered. That composition reorganizes across scale and that most state-realization failures violate a task invariant are post-hoc observations from data collected for another question, as are the cross-environment, transition and model-campaign analyses; none of their p-values is confirmatory. We applied one registered stopping rule in the analysis instead of halting the run, and we disclose this deviation.

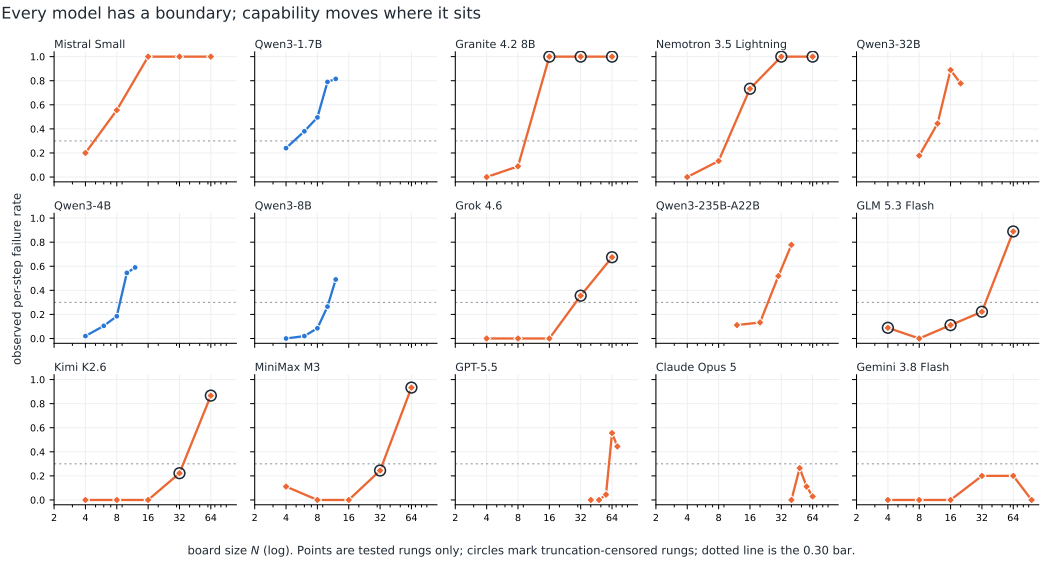

Figure 9: Observed per-step failure rate against board size for every model, including the excluded ones. Points show tested rungs only, circles mark rungs censored by truncation, and the dotted line is the 0.30 bar. Every model that crosses the bar keeps failing at larger boards.

Table 6: Observed per-step error at every tested rung for every model. A † marks rungs where more than 2% of answers hit the token ceiling. No value of N is interpolated. Per-channel counts are in `results/campaign_grid.csv`.

model	N: error rate
Qwen3-1.7B	4: 0.24, 6: 0.38, 8: 0.49, 10: 0.79, 12: 0.81
Mistral Small	4: 0.20, 8: 0.56, 16: 1.00, 32: 1.00, 64: 1.00
Qwen3-4B	4: 0.02, 6: 0.10, 8: 0.18, 10: 0.55, 12: 0.59
Qwen3-32B	8: 0.18, 12: 0.44, 16: 0.89, 20: 0.78
Granite 4.2 8B	4: 0.00, 8: 0.09, 16: 1.00†, 32: 1.00†, 64: 1.00†
Nemotron 3.5 Lightning	4: 0.00, 8: 0.13, 16: 0.73†, 32: 1.00†, 64: 1.00†
Qwen3-8B	4: 0.00, 6: 0.02, 8: 0.09, 10: 0.27, 12: 0.49
Grok 4.6	4: 0.00, 8: 0.00, 16: 0.00, 32: 0.36†, 64: 0.68†
Qwen3-235B-A22B	12: 0.11, 20: 0.13, 30: 0.52, 40: 0.78
GLM 5.3 Flash	4: 0.09†, 8: 0.00, 16: 0.11†, 32: 0.22†, 64: 0.89†
MiniMax M3	4: 0.11, 8: 0.00, 16: 0.00, 32: 0.24†, 64: 0.93†
Kimi K2.6	4: 0.00, 8: 0.00, 16: 0.00, 32: 0.22†, 64: 0.87†
GPT-5.5	40: 0.00, 48: 0.00, 56: 0.04, 64: 0.56, 72: 0.44
Claude Opus 5	40: 0.00, 48: 0.26, 56: 0.11, 64: 0.03
Gemini 3.8 Flash	4: 0.00, 8: 0.00, 16: 0.00, 32: 0.20, 64: 0.20, 96: 0.00