WHEN DOES REDUNDANT REASONING PAY? A MEASURABLE CONDITION FOR ERROR-TOLERANT REASONING IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models are often made more reliable by having them re-derive their own intermediate results. This redundancy can repair a corrupted step, but it also costs accuracy when nothing goes wrong, so it sometimes helps and sometimes hurts, and it is unclear in advance which. We model a reasoning chain as a sequence of noisy steps in which both a brittle procedure (*sharp*, which states each result once) and a redundant one (*flat*, which re-derives the previous result before using it) can repair a corrupted step. The model has three parameters, each measured directly on the language model, namely the two procedures' per-step recovery rates c_s and c_f, found by injecting errors into correct reasoning traces, and the per-step accuracy cost of redundancy κ. We prove that the redundant procedure can overtake the brittle one if and only if $c_f > \kappa\, c_s$, and give in closed form the per-step error rate beyond which it does. The usual assumption that a brittle chain never recovers ($c_s{=}0$) makes this condition trivially true, yet on six of our seven task families the brittle procedure repairs 10% to 63% of corrupted steps unaided. On nine settings spanning three models and two chain lengths, eight of them modular arithmetic, predictions from the measured parameters alone, with nothing fitted to the crossover, match the observed crossover with $R^2{=}0.839$ (mean absolute error 0.021). The condition also agrees with all six full noise sweeps, including a pre-registered prediction of no crossover that held. Before any compute is spent on redundancy, a few short measurements therefore say whether it can pay off at all, a 64-problem screen flags tasks where it cannot because the model's errors recur at every step, and the same parameters bound the sample size needed to detect the gain. Code is available at `https://anonymous.4open.science/r/reasoning-noise-budget-DF74`.

1 INTRODUCTION

Large language models reason in chains, and chains accumulate error: each step can corrupt the state passed to the next, so the probability that an entire chain is correct decays as it grows (Gan et al., 2025; Zhang et al., 2024; Sinha et al., 2025; Kwa et al., 2025; Chen et al., 2026c). Chain-of-thought adds genuine computation (Wei et al., 2022; Li et al., 2024; Merrill & Sabharwal, 2024), though not always to the benefit of accuracy (Liu et al., 2024), and its reliability is acutely sensitive to per-step accuracy (Gan et al., 2025; Wu et al., 2025); agentic systems formalize this as a noisy process whose errors "compound silently across steps" (Yan et al., 2026). One response is to make each step more reliable, through verification, process rewards, self-refinement, repeated sampling, search or runtime denoising (Lightman et al., 2024; Wang et al., 2024; Cobbe et al., 2021; Weng et al., 2023; Ling et al., 2023; Madaan et al., 2023; Wang et al., 2023; Brown et al., 2024; Yao et al., 2023; Besta et al., 2024; Snell et al., 2024; Yan et al., 2026). Another is to make the *procedure* tolerant of its own slips: have the chain re-derive an intermediate result rather than trust it, so a corrupted step can be repaired downstream. That is not free, since redundancy costs tokens and, empirically, noiseless accuracy, so the practical question is not whether robustness is good but *when it is worth its price*. Boundaries of that shape now exist: a verify-and-repair loop has a stopping rule computed from measured verifier and repairer rates (Wu et al., 2026), intrinsic self-correction has a stability threshold in baseline accuracy (Liu & Meng, 2026), and verification helps below an

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.

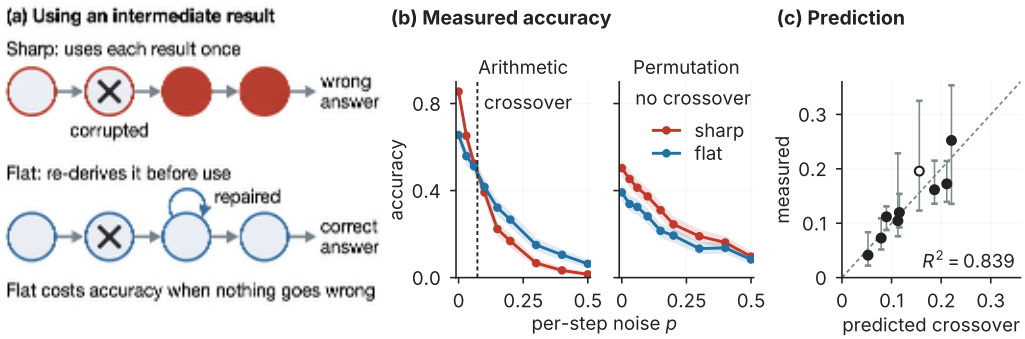


Figure 1: **Redundancy pays only when it repairs more than the brittle procedure already does.** (a) Schematic. A *sharp* procedure uses each intermediate result once, so a corrupted step reaches the answer; a *flat* procedure re-derives the previous result before using it, which can repair the step but costs accuracy when nothing is corrupted. (b) Measured accuracy under injected per-step noise, with 95% bootstrap bands. On arithmetic (Qwen2.5-7B-Instruct, L=8) flat repairs 0.414 of corrupted steps against sharp's 0.007, so $c_f > \kappa c_s$ holds and flat overtakes at p^*=0.073, as in eight of the nine calibration settings. On permutation tracking (L=4) flat repairs fewer than sharp (0.191 against 0.255), the condition fails, and flat never overtakes; this is the only swept family that violates the condition, and its null was recorded before the sweep (Section 4.4). (c) Closed-form prediction from independently measured parameters against the measured crossover, with 95% intervals, over the nine calibration settings (Table 2); the open marker is the stack family, whose reversal does not resolve.

upstream accuracy and hurts above it (Yasir et al., 2026; Verma, 2026; Lu et al., 2025). None of them locates a boundary in the variable that drives the decay itself, the per-step error rate, or tests whether such a boundary exists at all for a given pair of procedures.

This paper derives both (Figure 1). We cast a reasoning chain as transmission over a noisy *step-channel*: a *sharp* schema transmits with little redundancy, a *flat* schema with re-derivations and cross-checks that can repair a corrupted step at the cost of a noiseless accuracy tax. This is the classical picture of reliable computation from unreliable components (von Neumann, 1956; Pippenger, 1985; Shannon, 1948; Cover & Thomas, 2006), in its modern reading of a model stage as a channel (He et al., 2025), and the regime where coding beats no-coding is the reasoning analogue of "survival of the flattest" in quasispecies evolution, where high mutation rates favor a mutationally robust genotype over the fittest one (Wilke et al., 2001; Codoner et al., 2006; Bull et al., 2005; van Nimwegen et al., 1999; Lauring et al., 2013; Berger & Cerf, 2025); at finite genome length the Eigen threshold is a smooth crossover of width $\sim 1/L$ (Eigen, 1971; Eigen et al., 1988; Swetina & Schuster, 1982; Campos & Fontanari, 1998; 1999; Alves & Fontanari, 1998), and the same is true here.

The scope claim. An earlier, unpublished version of this line of work established that such a crossover exists on iterated modular arithmetic and gave a closed form $p^* = (\kappa - 1)/(\kappa - 1 + c)$ in the per-step tax ratio κ and the flat schema's recovery c. That form silently assumes the sharp schema recovers nothing, c_s=0: true by construction on arithmetic, where we measure c_s=0.007, but a property of that task's state representation rather than of brittle procedures in general. Across six further families the sharp schema recovers 0.100 to 0.625 on the cells we measure, since a stack, a program state or a kinship chain is redundant enough that a corrupted intermediate is often caught with no deliberate check; on the one such family we swept, a stack machine, the earlier form is low by a factor of four and outside the empirical confidence interval. The corrected law, derived and proved here, is

$$q^* = \frac{\kappa - 1}{\kappa(1 - c_s) - (1 - c_f)}, \qquad \text{existing in } (0, 1) \text{ if and only if } c_f > \kappa\, c_s, \tag{1}$$

which reduces to the earlier expression term for term at c_s=0. It carries an existence condition the earlier form could not state (Figure 1b), a new comparative static in c_s, and a companion result on when a crossover that exists is nevertheless too small or too low to be measured. The same pass

replaced the criterion for calling a reversal real, requiring a significantly negative flat-minus-sharp difference followed by a significantly positive one rather than the first point-estimate sign change; two crossovers reported in that earlier version do not survive it, on a register automaton and on arithmetic at $L{=}16$, and both are withdrawn (Section 4.4).

Three lines that surround this question without asking it. *Absolute decay* measures how a *single* procedure decays with length, *noise mitigation* pushes accuracy up, and *fragility* studies what perturbing a step destroys. We order *two length-matched* procedures instead, so decay common to both cancels, and we credit recovery only on steps whose corruption determines the answer, which separates a repair from a lucky survival. Appendix D gives each line with its citations and the distinction from flat minima in weight space (Foret et al., 2021; Tsipras et al., 2019).

The recovery measurement is not new; its comparative use is. Phan et al. (2026) insert arithmetic errors into intermediate reasoning steps and measure how often the model still reaches the correct answer without verbalized correction, on our task type and with the ablations that rule out the model discarding the corrupted trace and re-solving; observational versions appear in Lee et al. (2026); Pham et al. (2026). That is the quantity of Section 3, measured first and measured well, and we claim no priority for it. What is new is comparative: the identical probe is run on two length-matched schemas, and the resulting pair (c_s, c_f) is not a diagnostic but a *parameter*, the two numbers deciding through (1) whether the redundant schema can win and where. That work varies model scale and intervention context but never the schema, so no ordering between procedures follows.

Break-even points, and what an existence condition adds. The closest formal relative is the two-rate channel of Reitich (2026): auditing an operation before and after splits its net accuracy change into a correction and a corruption rate, with a break-even requirement rising with baseline success, the dependence our κ carries. Two things separate a crossover from a break-even: it orders two named procedures at matched length rather than an operation against its absence, its rates entering per step through the exponent L; and a break-even always exists somewhere, whereas (1) can have no root in $(0, 1)$ at all. That last point is also what we add to the "redundancy can hurt" literature, which is by now quantitative: Wu et al. (2026) stop a noisy verify-repair loop on the sign of the marginal gain of one more round under a four-parameter noise model, Liu & Meng (2026) give a stability threshold in baseline accuracy with a prompt intervention that flips a degrading model as predicted, and further boundaries in upstream accuracy, model size and task appear elsewhere (Yasir et al., 2026; Verma, 2026; Lu et al., 2025; Wang et al., 2026; Stav et al., 2026). That redundancy has a boundary is their result, not ours; what $c_f > \kappa c_s$ adds is the parameterisation that subsumes it, since their rules say when to stop repairing and ours says whether repairing could have paid at all, from two probes run before the loop; a scoping condition of the same shape, for reflection-driven search, is proved by Wolf et al. (2026).

Is the step the right unit? Arbuzov et al. (2026) argue that it is not: inside a bounded domain, reliability is catalogue discovery and intervention coverage over *hard decisions*, with a sufficient per-decision budget growing polylogarithmically in sequence length rather than an error compounding exponentially in it. Our channel prices every step as an independent draw, so wherever steps and hard decisions come apart the exponent L in (3) counts the wrong thing. We take this as a scope condition, not a refutation, and one we can check: our families are built so that each of the L steps is one hard decision, exactly verified and load-bearing by construction, and that correspondence is what breaks on the derived families, where our screen then says the step-channel does not apply (Section 4.3). Where hard decisions grow with task length, the right L counts decisions, which the same probes measure. Relatedly, the per-step against per-problem dichotomy has been reached independently from token-level signatures (Thoria et al., 2026) and correlated inter-step errors are known to dominate long-horizon agent failure (Khanal et al., 2026; Yang, 2026): ours is not that distinction but its use as a *task-family* property deciding whether a crossover can exist.

A consistency check on independent data. Chapra et al. (2026) vary reasoning depth at fixed semantic content and report geometric per-step decay, one ceiling per task family, and that 14.5% of correct answers come through incorrect intermediate reasoning, an independent measure of the lucky survival our filter removes. They also report that forcing verbose state-tracking does not move

the ceiling (McNemar p=1.000). If that manipulation adds tokens and restatement without a re-derivation whose result is acted on, it leaves c_f at c_s, (1) then fails for any positive tax, and no depth or noise level can make the verbose variant win; our token-matched padded control behaves so where we can measure it, never producing a reversal. This is a consistency check, not evidence: κ, c_s and c_f were not measured in their setup.

Contributions. Two adjacent results are not this one, a budget-driven crossover between thinking and non-thinking modes (Nie et al., 2026) and a risk decomposition for a single procedure (Zhang et al., 2026), neither ordering two procedures along a noise axis nor carrying an existence condition (Appendix D). We claim that no prior work states a crossover in *per-step noise* between two reasoning procedures, that none gives a condition for whether redundancy can pay at all, and that the quasispecies reading is unclaimed, three independent searches for survival-of-the-flattest, Eigen error thresholds or mutational robustness transferred to LLM reasoning having returned nothing. That, and nothing more.

1. **A generalized law with an existence condition** (Theorem 1, restating (1)) and a **resolvability result** (Proposition 1) giving a necessary sample size below which a real crossover cannot be detected, plus a new comparative static: the budget is κ times more sensitive to the *sharp* schema's recovery than to the flat schema's.
2. **Calibration on nine distinct cells** (Figure 1c; Sections 4.1, 4.2) varying schema, model and depth on iterated arithmetic plus one cross-family cell: parameter-free predictions track measurement with $R^2 = 0.839$ about $y{=}x$, and a fourteen-variant continuum reproduces the governing comparative static on a controlled axis. Eight of the nine cells are arithmetic, so the cross-family axis carries one point.
3. **An error-process taxonomy** (Section 4.3) turning a known distinction into a task-family property that decides which families can support the phenomenon at all: redundancy repairs independent per-step slips and is powerless against a correlated per-problem misunderstanding, which a one-minute pre-screen detects.
4. **Accounted-for negatives, including a pre-registered null** (Section 4.4): on all six swept cells the law calls correctly whether a resolvable reversal appears, three further families are screened but not swept, and on a permutation family we recorded "no crossover at any noise level" *before* running the sweep.

2 Theory: when redundancy pays, and when it can be seen

Every quantity below is defined so it can be measured on a real model (Section 3); proofs and numerical verification are in Appendix A.

2.1 Reasoning as a noisy step-channel, with recovery on both sides

A *reasoning schema* S emits a trace with L *load-bearing* steps: steps whose corruption, propagated without correction, changes the terminal answer, which an exact verifier scores, decorative steps being excluded by the filter of Section 3. A noise kernel of intensity p perturbs step boundaries; we summarize it by the *aligned effective per-step error rate*

$$q \;=\; \Pr[\text{a single step is corrupted such that, propagated naively, it flips the answer}], \qquad (2)$$

measured per (kernel, model, schema), so kernels share an axis. Each schema carries a *per-step recovery capability*, the conditional probability that a corrupted load-bearing step is repaired before the terminal answer. The earlier model grants recovery only to the flat schema; we grant it to both, $c_s \in [0, 1)$ and $c_f \in [0, 1]$, by the identical probe. Under length-matching ($L_{\text{sharp}}{=}L_{\text{flat}}{=}L$), so that any reversal is attributable to robustness structure rather than length, per-step survival is $1{-}q(1{-}c_S)$ and

$$A_{\text{sharp}}(q) = a_s\big(1-q(1-c_s)\big)^L, \qquad A_{\text{flat}}(q) = a_f\big(1-q(1-c_f)\big)^L, \qquad 0 < a_f \le a_s \le 1. \quad (3)$$

The noiseless tax of robustness (Tsipras et al., 2019) appears twice below, as the accuracy gap $a_s - a_f$ and, per step, as the ratio $\kappa := (a_s/a_f)^{1/L} \ge 1$, with $\delta := \ln(a_s/a_f) = L \ln \kappa$. Two derived quantities carry the section:

$$M := \kappa(1-c_s) - (1-c_f) \;=\; (\kappa-1) + (c_f - \kappa c_s) \;=\; (\kappa-1)(1-c_s) + (c_f - c_s), \qquad (4)$$

the *margin*, written first through the *net recovery* $c_{\text{net}} := c_f - \kappa c_s$ and then as a tax term plus a *recovery gap*; which of the last two dominates decides whether the crossover is measurable (Section 2.3).

2.2 THE GENERALIZED CROSSOVER LAW

Theorem 1 (Generalized reasoning noise budget)**.** *Let $\kappa > 1$, $c_s \in [0, 1)$, $c_f \in [0, 1]$. Then A_{sharp} and A_{flat} cross at an interior point $q^* \in (0, 1)$* if and only if

$$c_f > \kappa\, c_s \qquad \Big(\textit{equivalently } c_{\text{net}} > 0,\ \textit{equivalently } \ln c_f - \ln c_s > \delta/L\Big), \tag{5}$$

read as $c_f > 0$ when c_s=0. The crossing is then unique, sharp is strictly better on $[0, q^*)$ and flat strictly better on $(q^*, 1]$, and

$$\boxed{q^* = \frac{\kappa - 1}{\kappa(1 - c_s) - (1 - c_f)} = \frac{\kappa - 1}{M} = \frac{\kappa - 1}{\kappa - 1 + c_{\text{net}}}} \tag{6}$$

If $c_f \le \kappa c_s$ there is no interior crossing and $A_{\text{sharp}}(q) \ge A_{\text{flat}}(q)$ on all of $[0, 1]$; in particular the flat schema never overtakes when $c_f \le c_s$. At c_s=0, (6) and (5) reduce term for term to the earlier $q^* = (\kappa - 1)/(\kappa - 1 + c_f)$ and $c_f > 0$.

Read (5) as: *the flat schema's recovery must beat the sharp schema's recovery marked up by the per-step noiseless tax*; if the brittle baseline already self-corrects, there may be nothing left to buy. The weaker condition $M > 0$ is necessary but *not* sufficient, satisfied without a crossing by 105 of 2000 randomized admissible configurations while (5) has zero violations, and Table 1 reports a family in exactly that gap (GSM, M=+0.490 with $c_f < \kappa c_s$).

Comparative statics. Differentiating (6),

$$\frac{\partial q^*}{\partial c_f} = -\frac{\kappa - 1}{M^2} < 0, \qquad \frac{\partial q^*}{\partial c_s} = +\frac{\kappa(\kappa - 1)}{M^2} > 0, \qquad \frac{\partial q^*}{\partial \kappa} = \frac{c_f - c_s}{M^2}, \tag{7}$$

the last positive wherever a crossover exists. The first survives in sign but not magnitude, since $M \le \kappa - 1 + c_f$ with equality only at c_s=0. The second is new: a brittle baseline that quietly self-corrects *raises* the budget, and $\partial q^*/\partial c_s = -\kappa\, \partial q^*/\partial c_f$ exactly, so the budget is κ times more sensitive to the sharp schema's recovery. The third holds only on the existence region; stated unconditionally, as the earlier theory did, it is false, since a pair with $c_f < c_s$ has $\partial q^*/\partial \kappa < 0$ and no crossover.

2.3 EXISTENCE IS NOT ENOUGH: RESOLVABILITY

A statistically honest criterion reports a budget only when the reversal resolves on both sides of the crossing (Section 3), and families can satisfy (5) and still fail that test. The same closed form says when. Write $\varepsilon(q) := A_{\text{flat}}(q) - A_{\text{sharp}}(q)$ and define the *altitude factor* $w := 1 - q^*(1 - c_s) = (c_f - c_s)/M \in (0, 1)$, by (4) the share of the margin contributed by the recovery gap rather than the tax.

Proposition 1 (Separation, altitude, and the attainable advantage)**.** *Assume* (5)*. Writing $D := \ln A_{\text{sharp}} - \ln A_{\text{flat}}$, the crossing has separation **(R1)** $|D'(q^*)| = LM/(\kappa w)$, altitude **(R2)** $A^* := A_{\text{sharp}}(q^*) = a_s w^L$, slope **(R3)** $\varepsilon'(q^*) = (a_s L/\kappa) M w^{L-1}$ and attainable advantage **(R4)** $G_{\max} := \max_{q \ge q^*} \varepsilon(q) \le a_s w^L M/(1 - c_f)$. Consequently **(R5)** no grid point can exhibit a significant flat advantage at level α unless $N \ge z^2_{\alpha/2}\, \sigma^2/G^2_{\max}$, with σ^2 the sampling variance of the estimated difference at the maximizing noise level.*

A real crossover can therefore be invisible in two independent ways: a *thin margin* (M small at healthy altitude), which separates the curves slowly with required N growing like M^{-2}, and a *floor* (w small, so $A^* = a_s w^L$ sits near chance), which leaves nothing to measure whatever the margin. The same quantity $M/(\kappa w)$ sets the localization width of the empirical crossover (Proposition 5), so a wide confidence interval and an unresolved reversal are one phenomenon. (R5) is necessary and not sufficient, a planning instrument rather than a power calculation: it rules out the two withdrawn cells (N_{req}=1937 and 2832 against 600 and 200 collected) but admits the stack family at 114, which then failed to resolve at N=600, so on our cells it should be inflated by roughly a factor of two before use.

2.4 WHAT THE c_s=0 THEORY GOT RIGHT

The model-free single-crossing theorem (Theorem 2), the flatness–hazard bridge and its two-sided form (Lemma 1), the reliable depth $L^*(q) \propto 1/q$ (8) (Theorem 3), the $\Theta(L)$ sharpness of the crossing (Theorem 4) and the optimal amount of redundancy (Proposition 4) all survive, with the earlier c becoming the recovery *gap* $c_f - c_s$ and the bridge limit becoming $(1 - c_f)/(1 - c_s)$ (9), which dropping c_s misstates by 0.75% on arithmetic and 83% on the stack family; statements, constants and proofs are in Appendix A. Two of them change what can be claimed. Existence is *per-kernel*, $c_{f,k} > \kappa c_{s,k}$ (Proposition 3), so the earlier claim that it is a kernel-independent property of a schema is an artifact of c_s=0. And because $\kappa = e^{\delta/L}$ falls with L while (5) does not, existence is depth-dependent: at fixed δ and $c_f > c_s > 0$ a crossover exists in $(0, 1)$ if and only if $L > L_{\min} := \delta/\ln(c_f/c_s)$ (Corollary 1), so a self-correcting sharp baseline can have no crossover at short depths and acquire one by deepening the chain. Computed from Table 1, $L_{\min}$ runs 0.07 to 2.17 across our families and does not bind, except on GSM, where $L_{\min}$=$5.86 > L$=2 and the null is an existence-depth failure.

3 MEASURING THE LAW

The theory references four quantities per cell: the noiseless accuracies a_s, a_f (hence κ), the two recoveries c_s, c_f, and the aligned effective error q of (2). All are measured, none fitted, so a prediction of q^* is parameter-free with respect to the crossover it predicts (construction in Appendix C, run settings in Appendix E).

Families and schemas. Seven families share one interface: an initial state, L verifiable single-step transitions, an exact terminal verifier, and a corruption operator on an intermediate value. Four are synthetic (iterated modular arithmetic, the family the earlier theory was built around; a stack machine; a register automaton of two-digit rewrites; permutation composition) and three are *derived* from public corpora (code execution (Gu et al., 2024), kinship composition (Sinha et al., 2019), GSM-Symbolic linear chains (Mirzadeh et al., 2025)), covering only the step-verifiable subset of each. **Sharp** emits one minimal line per step and trusts it; **flat** is length-matched but re-derives the previous committed value and reconciles before use, overriding a mismatch with the re-derivation, a check digit in the spirit of reliable computation from unreliable components (von Neumann, 1956; Shannon, 1948); **sharp+padded** is a length control, equally verbose with no genuine cross-check.

Kernels and probes. At each step boundary, **K1 (semantic)** replaces the committed value with a near-miss with probability p; **K2 (activation)** adds Gaussian noise scaled to the residual-stream RMS to a mid-block hidden state; **K3 (sampling)** generates the step at high temperature. Accuracy is plotted against the knob p; one perturbed step is load-bearing with probability q_{eff}, so $q = q_{\text{eff}}\, p$ and every prediction below is mapped to knob units as $p^* = q^*/q_{\text{eff}}$ at the *sharp* schema's measured q_{eff} for that sweep. The recovery probe runs on every step of a clean, verifier-correct trace: the *load-bearing* test injects a corruption, propagates it deterministically, and keeps the step only if the answer changes, filtering the "lucky survivals" in which a perturbed trace reaches the right answer through a broken step (Jiang et al., 2025; Gur-Arieh et al., 2026) (on arithmetic the unfiltered rate overstates the sharp schema's recovery by a factor of 4.1, 0.031 against 0.007); the model then continues under its own schema, and the recovery rate, the load-bearing neighborhood-survival rate (NSR) of the earlier version of this work, is the fraction of load-bearing steps still reaching the correct answer. We run it on *both* schemas, which the generalized theory requires and the earlier version omitted. A separate pre-screen measures, on 64 zero-injection traces, the per-trace distribution of intrinsic slips against the $\mathrm{Binomial}(L, q_0)$ an independent step-channel predicts, reporting the variance ratio as an overdispersion factor.

When a reversal counts. A grid point is significantly negative if the paired bootstrap 95% CI of $A_{\text{flat}} - A_{\text{sharp}}$ lies entirely below zero, significantly positive if entirely above. We report a p^* only when some significantly negative point is followed by a later significantly positive one, a *sustained* reversal; otherwise the cell is unresolved and its point-estimate sign change, if any, is not a budget. A functional floor rejects any cell whose flat schema never clears $A_{\min} = \max(5A_{\text{chance}}, 0.05)$, for A_{chance} a measured copy-the-first-quantity baseline. This bar is harder than the first sign change, is

applied identically to every family and depth, and is what withdrew two crossovers reported in an earlier version of this work (Section 4.4).

Models. The primary model is Qwen2.5-7B-Instruct (Team, 2025), with Phi-4-mini-instruct ($\sim$ 3.8B) and Phi-4 (14B) for a size and family axis and Qwen2.5-1.5B-Instruct as a below-capability control. A format pre-flight over a ten-model ladder finds a parse rate of 1.000 (Wilson $[0.893, 1.0]$, n=32) for *both* schemas on every model, so nothing below is a formatting artifact. Decoding is greedy except where a kernel injects sampling. Sample sizes are per cell, pooled over seeds $\{42, 123, 456\}$, running from N=160 per grid point on the two Phi cells and the below-capability control to N=600 on the depth sweep, stack, permutation and continuum, with N=64 for the screen-only families; each cell's N is in Table 2.

4 RESULTS

Three evidence lines calibrate the law: one varies the schema at a fixed task and model, one varies the model and the depth, one varies the task family with a single point. A fourth tests the law where it says nothing will be found.

4.1 THE LAW PREDICTS THE BUDGET WITHOUT EVER FITTING IT

Figure 1c is the paper's central claim; Table 2 in Appendix B lists its cells with their sample sizes. Of ten measurements, nine on arithmetic and one on the stack machine, two are the same condition: the continuum anchor and the cross-model Qwen2.5-7B L=8 cell are the same schema, task, model and depth through two pipelines on independent draws. We fit over the *nine distinct conditions*, spanning a factor of 4.2 in predicted budget, each predicted from three numbers measured on data that never sees the crossover: $R^2 = 0.839$ about y=x, mean absolute error 0.0208, budgets from 0.041 to 0.252. The duplicated pair is informative on its own: the two pipelines measure p^*=0.072 and 0.073 against predictions of 0.068 and 0.079.

The cross-model line reproduces both comparative statics of (7): Phi-4-mini pays the same 0.20 noiseless accuracy gap as Qwen2.5-7B at a near-identical tax ratio (κ=1.029 against 1.034) but recovers more (c_f=0.561 against 0.414) and crosses *sooner* (0.041 against 0.073), which is $\partial q^*/\partial c_f < 0$; Phi-4 recovers less and is taxed more (κ=1.042) and crosses *later* (0.105), which is $\partial q^*/\partial \kappa > 0$, though its sharp schema is at ceiling (160 of 160 at p=0) so that κ is a boundary estimate that can only move down. The phenomenon is not a length effect: the padded control never reverses on any sweep (Appendix G).

The cross-family point separates the corrected law from the earlier one. The stack machine's sharp schema recovers c_s=0.454 against arithmetic's 0.007, a stack's state being naturally redundant. Under the mapping convention of Section 3, the earlier c_s=0 form predicts a crossing at p^*=0.047 against an empirical bootstrap interval of $[0.123, 0.325]$, low by a factor of four and *outside* it, where the generalized form predicts 0.156, inside it; that interval comes from a cell whose reversal does not resolve (Section 4.4). On arithmetic, where $c_s \approx 0$, the two forms differ by 0.0013 at L=8 and 0.0016 at L=12, at least 38 times narrower than the interval on either measurement: **no arithmetic budget from the earlier version moves**.

4.2 A CONTROLLED CONTINUUM: THE COMPARATIVE STATIC, BY CONSTRUCTION

Cross-model comparisons vary many things at once. A continuum of fourteen flat-schema variants isolates one term of the denominator of (6): one model, one family, L=8, differing only in how often and how completely the schema re-derives its previous committed value, over 135 schema-by-noise grid points at N=600 each (405 runs of N=200, three seeds), with measured recovery spanning $c_f \in [0.006, 0.537]$.

Three things follow. First, the anchor replicates: the variant reproducing the earlier flat schema measures c_f=0.434 against 0.414 and crosses at p^*=0.072 against 0.073. Second, five variants produce sustained reversals and all five are predicted parameter-free; the generalized form gives Pearson r=0.908 with RMSE 0.027 against r=0.900 and 0.028 for the earlier form, a gain of 0.008 in r, small because this family has $c_s \approx 0$ and is where the earlier form was already right. Third,

Table 1: Predicted against observed verdicts, from measured κ, c_s, c_f alone. "Exists" is condition (5); N_{req} is the per-grid-point sample size (R5) requires, against 600 collected on arithmetic, stack and permutation and 200 on the automaton. Predictions are in knob units p^*, or in effective-error units q^* for the screen-only GSM row, which has no sweep to map through. Provenance is when the verdict was fixed: *retrodiction* (sweep first, bound after, so a consistency check), *pre-registered* (prediction and falsifier recorded before the sweep), *prospective* (untested). The last three rows were screened only, at $N{=}64$, the source of their κ.

cell	κ	c_s	c_f	exists	provenance	predicted	observed
arithmetic $L{=}8$	1.034	0.007	0.414	yes	retrodiction	$p^*{=}0.079$, $N_{\text{req}}{=}78$	crossover, $p^*{=}0.073$
arithmetic $L{=}12$	1.041	0.005	0.332	yes	retrodiction	$p^*{=}0.116$, $N_{\text{req}}{=}379$	crossover, $p^*{=}0.120$
arithmetic $L{=}16$	1.043	0.007	0.273	yes	retrodiction	$p^*{=}0.144$, $N_{\text{req}}{=}1937$	no resolvable reversal
stack $L{=}8$	1.026	0.454	0.640	yes	retrodiction	thin margin $M{=}0.200$	no resolvable reversal
automaton $L{=}6$	1.121	0.100	0.281	yes	retrodiction	$p^*{=}0.418$ at $A^*{=}0.040$	no resolvable reversal
permutation $L{=}4$	1.065	0.255	0.191	**no**	**pre-registered**	no crossover at any noise level	none; sharp better at $7/9$
code execution $L{=}4$	1.107	0.312	0.431	yes	prospective	$p^*{=}0.554$, beyond the grid	screened, not swept
kinship comp. $L{=}3$	1.313	0.429	0.625	yes	prospective	$p^*{=}0.887$, beyond the grid	screened, not swept
GSM linear chain $L{=}2$	1.899	0.625	0.778	**no**	prospective	$q^*{=}1.84$, none in $(0, 1)$	screened, not swept

the re-derivation frequency series is monotone in both variables (Figure 2 in Appendix G): recovery $0.537, 0.386, 0.214, 0.133$ maps to budgets $0.112, 0.162, 0.172, 0.252$, endpoints separated at 95% and the middle pair overlapping, while κ is non-monotone across the series, so the tax cannot be producing the trend. A screening rule stated in advance, a resolvable reversal when the crossover exists and $p^*_{\text{pred}} \leq 0.30$, calls 12 of 14 variants correctly; both misses are variants it predicted resolvable that did not resolve, and they carry margins $M{=}0.080$ and 0.174, the smallest and the third smallest of the seven it predicted resolvable. It is a heuristic with an undefended threshold, not Proposition 1.

4.3 Which families can support the phenomenon at all

The model (3) assumes per-step errors are independent, testable in about a minute per family. On iterated arithmetic ($q_0{=}0.018$, accuracy 0.859 at $N{=}64$) the slip-count histogram tracks the independent prediction with overdispersion $0.87\times$, which at this sample size means no detectable overdispersion rather than demonstrated independence: only 9 of 64 traces carry any slip. On code execution ($q_0{=}0.465$, accuracy 0.312) the overdispersion is $4.72\times$ and the distribution is U-shaped, with mass at zero slips and at the maximum at once: 26.6% of traces are perfectly clean where independence predicts 0.7% (Figure 3, Appendix G). The screen detects gross violation, which is how it is used.

The mechanism explains the negatives that follow. A flat schema repairs a corrupted step by *re-deriving* it; when the error is a per-problem misunderstanding rather than a slip, the re-derivation reproduces the misunderstanding, so recovery is structurally near zero, and broken code-execution traces show exactly that, repeating an identical wrong operation at every step. The screen also separates two failure modes that look alike in aggregate accuracy: kinship composition is close to independent ($1.03\times$) but at an intrinsic rate ($q_0{=}0.755$) so high that both schemas floor, a capability problem rather than a correlation one. Stack and permutation screen as independent and are the two families we swept in full.

4.4 Negative results and withdrawn claims

Table 1 records when each verdict was fixed relative to its sweep. Five of the six swept cells are retrodictions, because their sweeps ran before the generalized theory existed, and the resolvability bound was derived after the $L{=}16$ and automaton cells failed to resolve, so for these cells the closed form is a consistency check on independently measured parameters. The sixth was predicted in advance and tests existence rather than location.

Two crossovers reported in an earlier version of this work do not survive the sustained-reversal criterion. A register-automaton crossover at $p^*{=}0.35$ ($L{=}6$), previously presented as evidence of generality beyond arithmetic, has no significantly positive grid point at $L{=}4$ or $L{=}6$; the largest difference favoring flat is $+0.025$ $[-0.010, +0.065]$. The arithmetic $L{=}16$ point, previously reported

at p^*=0.200, likewise has none at N=600. Both are withdrawn, recomputed from unmodified raw records.

A family that violates the existence condition provides its most direct test. On permutation tracking we measured a_s=0.503, a_f=0.392 (hence κ=1.065 at L=4), c_s=0.255 and c_f=0.191, so the flat schema recovers *less* than the sharp one, because re-deriving a permutation is harder than applying it, so redundancy incurs its tax without adding recovery. Condition (5) fails, $0.191 < 1.065 \times 0.255 = 0.271$, and the weaker margin condition fails as well, M=−0.016. The family separating the two conditions is GSM, with M=+0.490 > 0 but c_f=0.778 $< \kappa c_s$=1.187, which was screened but not swept. The prediction of no crossover at any noise level, with any resolvable reversal at N=600 as its falsifier, was recorded before the sweep was run. The difference is significantly negative at seven of nine noise levels and never significantly positive, with its maximum of -0.013 at p=0.5, so the prediction is confirmed.

The resolvability bound accounts for both withdrawals retrospectively. Part (R5) of Proposition 1 requires $N \geq 2832$ per grid point on the automaton, against 200 collected, and $N \geq 1937$ at L=16, against 600. Across the three arithmetic depths the ratio of collected to required sample size falls from 7.7 to 1.6 to 0.31, and the number of significantly positive grid points falls from 5 to 2 to 0. The budget therefore rises with depth, from 0.073 to 0.120, as the tax ratio grows (κ from 1.034 to 1.041) and recovery falls (c_f from 0.414 to 0.332), and it can no longer be resolved at L=16.

5 Discussion and conclusion

Our results indicate that whether redundancy can help is decided by quantities measurable before any noise sweep is run. The theory assumes independent per-step errors, and a 64-problem overdispersion screen detects gross violations, as on code execution, where a model's own errors recur throughout a trace and re-deriving a step reproduces them (Section 4.3). Because existence requires $c_f > \kappa c_s$, the recovery of the brittle baseline must be measured, not only that of the redundant procedure. Proposition 1 then yields a necessary sample size and a crossover altitude, and a crossover located near the accuracy floor cannot be observed at practical sample sizes. Two conditions bound the scope of the result. The model must be able to execute the redundant procedure (on Qwen2.5-1.5B-Instruct the flat schema never exceeds 0.037), and the noise must corrupt content, since existence is kernel-specific (Proposition 3) and activation or sampling noise never yields a significantly positive difference (Appendix G).

The idealized model predicts $q^* \propto 1/L$; on a capacity-limited model the budget instead rises with depth (Section 4.4), as the flat schema's tax grows and its recovery falls. Proposition 2 recovers the sign of this rise and about 60% of its magnitude, so redundancy gains with depth only where its per-step cost and recovery remain stable.

These results have several limitations. All families are synthetic or corpus-derived so that each step admits an exact verifier; transfer to open-ended reasoning remains open. Eight of the nine calibration settings are iterated arithmetic, and the only other, the stack machine, does not resolve at N=600. The empirical evidence that the c_s=0 form misstates the budget, as distinct from the measurements showing $c_s > 0$, therefore rests on one unresolved cell, and three families were screened but not swept. Proposition 1 is necessary rather than sufficient and is not tight here, since it admits the stack family at $N \geq 114$, which did not resolve at N=600, and the independent-error model overpredicts that family's largest flat advantage by a factor of 3.3. Recoveries under non-semantic noise remain unmeasured.

In conclusion, whether a redundant reasoning procedure can outperform a brittle one under per-step noise is governed by an existence condition, $c_f > \kappa c_s$, in quantities measurable directly on a model, with the crossover location in closed form when the condition holds. Crediting the brittle procedure with its own recovery is essential, because outside iterated arithmetic it repaired 10% to 63% of corrupted steps unaided. Extending these measurements to open-ended reasoning without step-level verifiers is the principal open problem.

AI USE STATEMENT

During the preparation of this manuscript, the authors used AI tools to assist with document formatting, LaTeX syntax, and technical code maintenance. In particular, an AI coding agent, Antigravity, was used to support refactoring and organizing the experimental codebase for release as supplementary material, with the aim of improving clarity and reproducibility for reviewers.

ETHICS STATEMENT

This work studies the robustness of reasoning procedures on synthetic and derived, exactly-verifiable tasks using publicly available open-weight models; it involves no human subjects, no private or personally identifiable data, and no new data release. The intended impact is positive: a measurable criterion for when redundant, error-tolerant reasoning is preferable, and an explicit statement of when it is not, can make long-horizon LLM systems more reliable and can prevent redundancy being added where it cannot help. We see no specific risk of harm beyond those already inherent to general LLM deployment.

REPRODUCIBILITY STATEMENT

All formal claims are proved in Appendix A; the closed form, the existence condition, the comparative statics and every part of Proposition 1 are additionally checked numerically against root-finding and finite differences over randomized configurations (Appendix A). The task generators, exact verifiers, noise kernels, schema prompts, the two-sided flatness (NSR) probe, the error-process prescreen, and the crossover criterion are described in Section 3, with construction details, coverage fractions and family-specific caveats in Appendix C and run settings in Appendix E. Per-cell sample sizes are given in Table 2 and range from $N{=}160$ (both Phi cells and the below-capability control) to $N{=}600$ per grid point (arithmetic depth sweep, 1000 at $p{=}0.1$ and 400 at $p{=}0.05$; stack; permutation; continuum), with paired bootstrap 95% confidence intervals; the three screen-only families are measured at $N{=}64$. Predictions of p^* are computed from independently measured tax and recovery and are never fitted to the crossover. Claims withdrawn relative to the earlier version of this work, and the criterion that withdrew them, are stated in Section 4.4 and Appendix C.

The complete experiment and analysis code is available in two forms: an anonymized repository at `https://anonymous.4open.science/r/reasoning-noise-budget-DF74`, browsable during review, and a supplementary zip archive. Both contain the task-family protocol and all seven families, the evaluation harness, the noise kernels, the load-bearing recovery probe, the error-process pre-screen, the sweep driver, the analysis pipeline, and the per-cell records from every sweep reported here, together with 269 tests. Because those records are included, every number, table and figure in this paper regenerates without a GPU: after installing the CPU requirements, `pytest tests/` reproduces the test suite, `experiments/sim/verify_generalized.py` rechecks the closed form against numerical root-finding, and `analyze_llm.py` and `analyze_continuum.py` regenerate the crossover table, the calibration fit and the figures. Collecting new measurements additionally requires a GPU that fits a 7B model in bfloat16 and the open-weight checkpoints named in Section 3; model and dataset locations are read from the environment rather than hardcoded, and the archive's README gives the exact sweep commands.

REFERENCES

Domingos Alves and Jose F. Fontanari. Error threshold in finite populations. *Physical Review E*, 57 (6):7008, 1998.

Ashwath Vaithinathan Aravindan and Mayank Kejriwal. Fragile thoughts: How large language models handle chain-of-thought perturbations. *arXiv preprint arXiv:2603.03332*, 2026.

Mikhail L. Arbuzov, Sisong Bei, Ziwei Dong, Dmitri Kalaev, and Alexey A. Shvets. Beyond exponential decay: Rethinking error accumulation in large language models. *arXiv preprint arXiv:2505.24187*, 2025.

Mikhail L. Arbuzov, Lee Mosbacker, Sisong Bei, Ziwei Dong, Dmitri Kalaev, and Alexey Shvets. The architecture of errors: From universal impossibility to patch-local LLM reliability. *arXiv preprint arXiv:2605.30628*, 2026.

Maxime Berger and Raphaël Cerf. Survival of the flattest in the quasispecies model. *Advances in Applied Probability*, 57:566–606, 2025. arXiv:2306.09221.

Maciej Besta, Nils Blach, Ales Kubicek, et al. Graph of thoughts: Solving elaborate problems with large language models. In *AAAI Conference on Artificial Intelligence*, 2024. arXiv:2308.09687.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, et al. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

James J. Bull, Lauren Ancel Meyers, and Michael Lachmann. Quasispecies made simple. *PLoS Computational Biology*, 1(6):e61, 2005.

Paulo R. A. Campos and Jose F. Fontanari. Finite-size scaling of the quasispecies model. *Physical Review E*, 58(2):2664, 1998.

Paulo R. A. Campos and Jose F. Fontanari. Finite-size scaling of the error threshold transition in finite populations. *Journal of Physics A: Mathematical and General*, 32(1):L1, 1999.

Shubh Chapra, Dhruv Kumar, Murari Mandal, and Yash Sinha. The complexity ceiling benchmark: A multi-domain evaluation of sequential reasoning under depth scaling. *arXiv preprint arXiv:2606.29278*, 2026.

Kuan-Yen Chen, Fang-Yi Su, Shih-Yen Lin, Bao Li, and Jung-Hsien Chiang. The self-correction illusion: Role relabeling gates explicit error flagging in large language models. *arXiv preprint arXiv:2606.05976*, 2026a.

Mingguang Chen, Bo Qu, and Licheng Wang. The calibration floor: Format repair can masquerade as self-correction at small-to-mid scale. *arXiv preprint arXiv:2608.04355*, 2026b.

Mingguang Chen, Licheng Wang, and Bo Qu. The horizon gap: Planning, memory, execution, training, and evaluation for long-horizon LLM agents. *arXiv preprint arXiv:2608.06663*, 2026c.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research (TMLR)*, 2023. arXiv:2211.12588.

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.08939.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Francisco M. Codoner, Jose-Antonio Daros, Ricard V. Sole, and Santiago F. Elena. The fittest versus the flattest: experimental confirmation of the quasispecies effect with subviral pathogens. *PLoS Pathogens*, 2(12):e136, 2006.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

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

Manfred Eigen. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 58(10):465–523, 1971.

Manfred Eigen, John McCaskill, and Peter Schuster. Molecular quasi-species. *The Journal of Physical Chemistry*, 92(24):6881–6891, 1988.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.01412.

Zeyu Gan, Yun Liao, and Yong Liu. Rethinking external slow-thinking: From snowball errors to probability of correct reasoning. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2501.15602.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. arXiv:2104.08821.

Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I. Wang. CRUXEval: A benchmark for code reasoning, understanding and execution. *arXiv preprint arXiv:2401.03065*, 2024.

Yoav Gur-Arieh, Ana Marasović, and Mor Geva. Faithfulness metrics don't measure faithfulness: A meta-evaluation with ground truth. *arXiv preprint arXiv:2605.25052*, 2026.

Alex Havrilla and Maia Iyer. Understanding the effect of noise in llm training data with algorithmic chains of thought. *arXiv preprint arXiv:2402.04004*, 2024.

Shizhe He, Avanika Narayan, Ishan S. Khare, Scott W. Linderman, Christopher Ré, and Dan Biderman. An information theoretic perspective on agentic system design. *arXiv preprint arXiv:2512.21720*, 2025.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Jie Huang, Xinyun Chen, Swaroop Mishra, et al. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.01798.

Enyi Jiang, Changming Xu, Nischay Singh, et al. Robust answers, fragile logic: Probing the decoupling hypothesis in llm reasoning. *arXiv preprint arXiv:2505.17406*, 2025.

Ryo Kamoi, Yusen Zhang, Nan Zhang, et al. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics (TACL)*, 2024. arXiv:2406.01297.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1609.04836.

Aaditya Khanal, Yangyang Tao, and Junxiu Zhou. Beyond pass@1: A reliability science framework for long-horizon LLM agents. *arXiv preprint arXiv:2603.29231*, 2026.

Thomas Kwa, Ben West, Joel Becker, et al. Measuring ai ability to complete long software tasks. *arXiv preprint arXiv:2503.14499*, 2025.

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

Adam S. Lauring, Judith Frydman, and Raul Andino. The role of mutational robustness in rna virus evolution. *Nature Reviews Microbiology*, 11(5):327–336, 2013.

Jinu Lee, Shivam Agarwal, Amruta Parulekar, Siddarth Madala, Dilek Hakkani-Tur, and Julia Hockenmaier. ReasoningFlow: Discourse structures for understanding LLM reasoning traces. *arXiv preprint arXiv:2606.05402*, 2026.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2402.12875.

Xiaobo Liang, Lijun Wu, Juntao Li, et al. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.14448.

Hunter Lightman, Vineet Kosaraju, Yura Burda, et al. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.20050.

Zhan Ling, Yunhao Fang, Xuanlin Li, et al. Deductive verification of chain-of-thought reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.03872.

Aofan Liu and Jingxiang Meng. Self-correction as feedback control: Error dynamics, stability thresholds, and prompt interventions in LLMs. *arXiv preprint arXiv:2604.22273*, 2026.

Ryan Liu, Jiayi Geng, Addison J. Wu, et al. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*, 2024.

Jack Lu, Ryan Teehan, Jinran Jin, and Mengye Ren. When does verification pay off? a closer look at LLMs as solution verifiers. *arXiv preprint arXiv:2512.02304*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.17651.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1706.06083.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.07923.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.05229.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. arXiv:1704.03976.

Wenhua Nie, Junlin Liu, Jianan Wu, Zijie Meng, Yilong Fan, Zhang Zijian, Haoran Zheng, and Jyh-Shing Roger Jang. The coupling tax: How shared token budgets undermine visible chain-of-thought under fixed output limits. *arXiv preprint arXiv:2605.07686*, 2026.

Hoang Pham, Dong Le, and Anh Tuan Luu. GRACE: Step-level benchmark for faithful reasoning over context. *arXiv preprint arXiv:2606.16151*, 2026.

Hoang Phan, Quang H. Nguyen, Hung T. Q. Le, Xiusi Chen, Heng Ji, and Khoa D. Doan. Decoding the critique mechanism in large reasoning models. *arXiv preprint arXiv:2603.16331*, 2026.

Nicholas Pippenger. On networks of noisy gates. In *Symposium on Foundations of Computer Science (FOCS)*, pp. 30–38, 1985.

Fernando Reitich. Correction and corruption: A two-rate view of error flow in LLM protocols. *arXiv preprint arXiv:2604.18245*, 2026.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27 (3):379–423, 1948.

Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, et al. Hybridflow: A flexible and efficient rlhf framework. In *European Conference on Computer Systems (EuroSys)*, 2025. arXiv:2409.19256.

Freda Shi, Xinyun Chen, Kanishka Misra, et al. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2302.00093.

Akshit Sinha, Arvindh Arun, Shashwat Goel, Steffen Staab, and Jonas Geiping. The illusion of diminishing returns: Measuring long horizon execution in llms. *arXiv preprint arXiv:2509.09677*, 2025.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. doi: 10.18653/v1/D19-1458.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2001.07685.

Elroy Stav, Dvir Berlowitz, Maayan Orner, and Sarit Kraus. When does intrinsic self-correction help? a task-sensitive analysis. *arXiv preprint arXiv:2606.23196*, 2026.

Jorg Swetina and Peter Schuster. Self-replication with errors: A model for polynucleotide replication. *Biophysical Chemistry*, 16(4):329–345, 1982.

Yefan Tao, Gerald Friedland, Madhusudhanan Chandrasekaran, and Luyang Kong. Reflection or re-generation? why LLM revision fails where human revision succeeds. *arXiv preprint arXiv:2607.28908*, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Tanvi Thoria, Kiana Jafari, Marc R. Schlichting, and Mykel J. Kochenderfer. How language models fail: Token-level signatures of committed and persistent reasoning failures. *arXiv preprint arXiv:2606.06635*, 2026.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory. *arXiv preprint arXiv:2411.11984*, 2024.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1805.12152.

Gladys Tyen, Hassan Mansoor, Victor Cărbune, et al. Llms cannot find reasoning errors, but can correct them given the error location. In *Findings of ACL*, 2024. arXiv:2311.08516.

Erik van Nimwegen, James P. Crutchfield, and Martijn Huynen. Neutral evolution of mutational robustness. *Proceedings of the National Academy of Sciences*, 96(17):9716–9720, 1999.

Yuvraj Verma. Try again, don't look back: Blind resampling outperforms self-repair in small code models. *arXiv preprint arXiv:2607.26117*, 2026.

John von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. In C. E. Shannon and J. McCarthy (eds.), *Automata Studies*, pp. 43–98. Princeton University Press, 1956.

Alexander von Recum, Leander Girrbach, and Zeynep Akata. Are reasoning llms robust to interventions on their chain-of-thought? In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2602.07470.

Kaiyang Wan, Lang Gao, Honglin Mu, Preslav Nakov, Yuxia Wang, and Xiuying Chen. A fano-style accuracy upper bound for llm single-pass reasoning in multi-hop qa. *arXiv preprint arXiv:2509.21199*, 2025.

Peiyi Wang, Lei Li, Zhihong Shao, et al. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2312.08935.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2203.11171.

Zhixiang Wang, Ziliang Hong, and Ulas Bagci. A decodability criterion predicts when hidden-state selection beats majority voting in large language models. *arXiv preprint arXiv:2608.17124*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2201.11903.

Yixuan Weng, Minjun Zhu, Fei Xia, et al. Large language models are better reasoners with self-verification. In *Findings of EMNLP*, 2023. arXiv:2212.09561.

Claus O. Wilke, Jia Lan Wang, Charles Ofria, Richard E. Lenski, and Christoph Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844): 331–333, 2001.

Yotam Wolf, Noam Wies, and Amnon Shashua. When does in-context search help? a sampling-complexity theory of reflection-driven reasoning. *arXiv preprint arXiv:2607.06720*, 2026.

Yitao Wu, Si Shen, Rui Yang, Hong Peng, and Bin Hu. Verify, repair, repeat, or stop? robust stopping for noisy verify-repair loops in LLM agents. *arXiv preprint arXiv:2607.17641*, 2026.

Yuyang Wu, Yifei Wang, Ziyu Ye, et al. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*, 2025.

Yandong Yan, Junwei Peng, Shijie Li, Chenxi Li, Yifei Shang, Can Deng, Ruiting Dai, Yongqiang Zhao, Jiaqi Zhu, and Yu Huang. Denoiseflow: Uncertainty-aware denoising for reliable LLM agentic workflows. *arXiv preprint arXiv:2603.00532*, 2026.

Zhanfu Yang. Why retrying fails: Context contamination in LLM agent pipelines. *arXiv preprint arXiv:2605.08563*, 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, et al. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.10601.

Tahreem Yasir, Sutapa Dey Tithi, Benyamin Tabarsi, et al. When verification hurts: Asymmetric effects of multi-agent feedback in logic proof tutoring. *arXiv preprint arXiv:2603.27076*, 2026.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019. arXiv:1901.08573.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2305.13534.

Yue Zhang, Zhiyi Dong, Tommaso Cesari, and Yongyi Mao. On the cost and benefit of chain of thought: A learning-theoretic perspective. *arXiv preprint arXiv:2605.21260*, 2026.

Zhanke Zhou, Rong Tao, Jianing Zhu, et al. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2410.23856.

A Proofs and numerical verification

Throughout, $A_{\mathrm{sharp}}(q) = a_s(1-qs)^L$ and $A_{\mathrm{flat}}(q) = a_f(1-qf)^L$ with $s := 1 - c_s$, $f := 1 - c_f$, $0 < a_f \le a_s \le 1$, $L \ge 1$, $\kappa = (a_s/a_f)^{1/L} \ge 1$, $\delta = \ln(a_s/a_f) = L \ln \kappa$, $M = \kappa s - f$ and $c_{\mathrm{net}} = c_f - \kappa c_s$.

A.1 THEOREM 1 (GENERALIZED CLOSED-FORM CROSSOVER)

Proof. Both curves are positive on $[0, 1)$ and nonnegative at q=1, so on $[0, 1)$ the equation $A_{\text{sharp}}(q) = A_{\text{flat}}(q)$ is equivalent, after taking L-th roots of the positive quantities, to

$$\Gamma(q) := \kappa(1 - qs) - (1 - qf) = 0.$$

Γ is affine with $\Gamma(0) = \kappa - 1 > 0$ and slope $\Gamma' = -(\kappa s - f) = -M$, so $\Gamma(q) = (\kappa - 1) - qM$ and $\Gamma(1) = \kappa c_s - c_f$.

Existence. An affine function with $\Gamma(0) > 0$ has a root in $(0, 1)$ if and only if $\Gamma(1) < 0$, i.e. if and only if $c_f > \kappa c_s$; and $\Gamma(0) > 0 > \Gamma(1)$ forces $M > 0$, so that root is $q^* = (\kappa - 1)/M$, which is (6). Conversely if $\Gamma(1) \geq 0$ then, by affinity, $\Gamma \geq \min(\Gamma(0), \Gamma(1)) \geq 0$ on $[0, 1]$, so $A_{\text{sharp}} \geq A_{\text{flat}}$ throughout, with equality only where $\Gamma = 0$, which for $\Gamma(1) = 0$ is the single endpoint q=1.

Uniqueness and ordering. Apply Theorem 2 on $[0, q_{\max}] = [0, 1]$. Its hypotheses hold: (A1) is $a_s \geq a_f$, i.e. $\kappa \geq 1$. For (A2), the base-model hazards are $h_S(q) = L(1 - c_S)/(1 - q(1 - c_S))$ and

$$h_{\text{sharp}}(q) - h_{\text{flat}}(q) = L\Big[\tfrac{s}{1-qs} - \tfrac{f}{1-qf}\Big] = \frac{L\,(s - f)}{(1 - qs)(1 - qf)},$$

because $s(1 - qf) - f(1 - qs) = s - f$ identically, the cross terms qsf cancelling; hence (A2) holds if and only if $c_f > c_s$, independently of κ, L, q. For (A3) at $q_{\max}$=1, $A_{\text{sharp}}(1) = a_s c_s^L < a_f c_f^L = A_{\text{flat}}(1)$ is $(c_f/c_s)^L > a_s/a_f = \kappa^L$, i.e. $c_f > \kappa c_s$ (read directly as $c_f > 0$ when c_s=0). Since $\kappa \geq 1$, (A3) implies $c_f > \kappa c_s \geq c_s$ and hence (A2), so the single condition delivers all three hypotheses.

Reduction. At c_s=0, $s = 1$ and $M = \kappa - 1 + c_f$, so both the formula and the condition reduce to the earlier ones.

Statics. With $dM/dc_f = 1$, $dM/dc_s = -\kappa$, $dM/d\kappa = s$, differentiating $(\kappa - 1)/M$ gives (7). □

$M > 0$ **is necessary but not sufficient.** $M > 0$ only says Γ has negative slope, so its root is at positive q; it does not place the root below q=1. The configuration a_s=0.8014, a_f=0.0552, c_s=0.5628, c_f=0.6292, L=13 has κ=1.2285 and M=0.1662 > 0, yet κc_s=0.691 $> 0.629 = c_f$ and the formula returns q^*=1.375, outside the probability range. The degenerate case is cleanest: with c_s=c_f and any $\kappa > 1$, $M = (\kappa - 1)(1 - c_s) > 0$ but $A_{\text{flat}}/A_{\text{sharp}} = a_f/a_s < 1$ identically.

A.2 THEOREM 2 (SINGLE CROSSING, MODEL-FREE)

Theorem 2 (Single crossing)**.** *Let* $A_{\text{sharp}}, A_{\text{flat}} : [0, q_{\max}] \to (0, 1]$ *be* C^1 *and* $h_S(q) := -\frac{d}{dq} \ln A_S(q)$. *Assume (A1)* $A_{\text{sharp}}(0) \geq A_{\text{flat}}(0)$*; (A2)* $h_{\text{sharp}}(q) > h_{\text{flat}}(q)$ *on* $(0, q_{\max}]$*; (A3)* $A_{\text{sharp}}(q_{\max}) < A_{\text{flat}}(q_{\max})$. *Then there is a unique* $q^* \in (0, q_{\max})$ *with* $A_{\text{sharp}}(q^*) = A_{\text{flat}}(q^*)$ *and the ordering flips exactly once.*

Proof. $D := \ln A_{\text{sharp}} - \ln A_{\text{flat}}$ is C^1 with $D(0) \geq 0$ by (A1), $D' = h_{\text{flat}} - h_{\text{sharp}} < 0$ by (A2), and $D(q_{\max}) < 0$ by (A3). A strictly decreasing continuous function with $D(0) \geq 0 > D(q_{\max})$ has a unique zero, and $\text{sign}(A_{\text{sharp}} - A_{\text{flat}}) = \text{sign}\, D$. □

A.3 PROPOSITION 1 (RESOLVABILITY)

Proof. (R1). Write $\Delta := -D = \ln A_{\text{flat}} - \ln A_{\text{sharp}}$. From the (A2) computation, $\Delta'(q) = L(s - f)/[(1 - qs)(1 - qf)]$. At the crossover the defining relation $\kappa(1 - q^*s) = 1 - q^*f$ gives $1 - q^*f = \kappa w$ with $w = 1 - q^*s$, so $\Delta'(q^*) = L(s - f)/(\kappa w^2)$. Also $w = 1 - s(\kappa - 1)/M = (\kappa s - f - \kappa s + s)/M = (s - f)/M = (c_f - c_s)/M$, giving $\Delta'(q^*) = LM^2/(\kappa(c_f - c_s)) = LM/(\kappa w)$. Here $w > 0$ since $c_f > c_s$ and $w < 1$ since $M - (c_f - c_s) = (\kappa - 1)(1 - c_s) > 0$.

(R2). $A_{\text{sharp}}(q^*) = a_s(1 - q^*s)^L = a_s w^L$, and it equals $A_{\text{flat}}(q^*)$ by definition.

(R3). $\varepsilon' = -A_{\text{flat}} h_{\text{flat}} + A_{\text{sharp}} h_{\text{sharp}}$, and at q^* the two accuracies coincide at A^*, so $\varepsilon'(q^*) = A^* \Delta'(q^*) = (a_s L/\kappa) M w^{L-1}$.

(R4). $\varepsilon'(q) = L[s\,a_s(1-qs)^{L-1} - f\,a_f(1-qf)^{L-1}]$, whose sign is that of $(sa_s)/(fa_f) - r(q)^{L-1}$ with $r(q) := (1-qf)/(1-qs)$. Since $r' = (s-f)/(1-qs)^2 > 0$, r is strictly increasing, so ε' changes sign at most once and only from $+$ to $-$: ε is unimodal. At q^*, $r(q^*) = \kappa$ and $\varepsilon'(q^*) > 0$ reduces to $\kappa s > f$, i.e. $M > 0$, so the maximum lies strictly to the right of q^*. Setting $\varepsilon'(q_m) = 0$ gives $r(q_m) = r_m := ((s/f)\kappa^L)^{1/(L-1)}$ and $q_m = (r_m - 1)/(r_m s - f)$. At the stationary point $a_f(1-q_m f)^{L-1} = (s/f)a_s(1-q_m s)^{L-1}$, so

$$\varepsilon(q_m) = a_s(1-q_m s)^{L-1}\Big[\tfrac{s}{f}(1-q_m f) - (1-q_m s)\Big] = a_s(1-q_m s)^{L-1}\tfrac{c_f - c_s}{1-c_f},$$

the q_m terms cancelling. Since $q_m > q^*$ and $q \mapsto (1-qs)^{L-1}$ decreases, $G_{\max} \le a_s w^{L-1}(c_f - c_s)/(1-c_f) = a_s w^L M/(1-c_f)$.

(R5). $\varepsilon(q) \le G_{\max}$ everywhere, and a two-sided test at level α rejects only if the observed difference exceeds $z_{\alpha/2}$ standard errors; taking the most favorable grid point gives the bound. With uncorrelated per-problem outcomes, $\sigma^2 = 2A_m(1-A_m)$ with $A_m := A_{\mathrm{sharp}}(q_m)$; measured on the relevant grid points the paired bootstrap standard error equals this value to within 20% (ratios 0.93 to 1.19). □

A.4 Depth: reliable depth, existence depth, and the depth derivative

Theorem 3 (Reliable depth). *$A_{\mathrm{flat}}(q) > A_{\mathrm{sharp}}(q)$ if and only if $L > L^*(q)$, where*

$$L^*(q) = \frac{\ln(a_s/a_f)}{\ln\big[(1-q(1-c_f))/(1-q(1-c_s))\big]} \overset{q\to 0}{\approx} \frac{\ln(a_s/a_f)}{(c_f-c_s)\,q} \;\propto\; \frac{1}{q}, \tag{8}$$

and dually $q^*(L) \approx \delta/((c_f - c_s)L) \propto 1/L$.

Proof. $A_{\mathrm{flat}}/A_{\mathrm{sharp}} = (a_f/a_s)\rho^L$ with $\rho = (1-qf)/(1-qs) > 1$ exactly when $c_f > c_s$, so the ratio is strictly increasing in L and equals 1 at $L\ln\rho = \delta$. For small q, $\ln\rho = (c_f - c_s)q + O(q^2)$. □

The earlier statement of (8) is the $c_s{=}0$ case: the recovery c becomes the recovery *gap*.

Corollary 1 (Existence depth). *With δ, $c_s > 0$, $c_f > c_s$ fixed, a crossover exists in $(0,1)$ if and only if $L > L_{\min} := \delta/\ln(c_f/c_s)$.*

Proof. Condition (5), $c_f > e^{\delta/L}c_s$, reads $\ln(c_f/c_s) > \delta/L$, i.e. $L > \delta/\ln(c_f/c_s)$; at $c_s{=}0$ the logarithm diverges and $L_{\min} = 0$, recovering unconditional existence. □

Proposition 2 (Depth dependence under capacity drift). *With differentiable $\kappa(L), c_s(L), c_f(L)$ and $M' = \kappa' s - \kappa c_s' + c_f'$,*

$$\frac{dq^*}{dL} = \frac{\kappa'(L)\,(c_f - c_s) + (\kappa(L)-1)\big(\kappa(L)c_s'(L) - c_f'(L)\big)}{M(L)^2}.$$

Proof. Differentiate $q^* = (\kappa - 1)/M$ through L by the quotient rule and collect terms. At $c_s \equiv 0$ this reduces to the earlier $[c\kappa' - (\kappa-1)c']/(\kappa - 1 + c)^2$. □

For the measured arithmetic transition $L{=}8 \to 12$, κ rises ($1.034 \to 1.041$), c_f falls ($0.414 \to 0.332$) and c_s falls negligibly ($0.0075 \to 0.0051$), so the first term is positive, the $-c_f'$ part of the second is positive and dominant, and the new $\kappa c_s'$ part contributes about -0.0006 per unit L against $+0.021$, under 3%. Hence $dq^*/dL > 0$: the budget rises with depth as measured. Evaluated at the $L{=}8$ parameters and integrated over $\Delta L{=}4$, the derivative predicts a rise of $+0.029$ against the $+0.047$ measured, so it gives the sign and about 60% of the magnitude.

A.5 SMOOTHNESS, HAZARD BRIDGE, KERNELS, AND OPTIMAL REDUNDANCY

Theorem 4 (Smooth crossover)**.** *At finite L, $D = \ln A_{\text{sharp}} - \ln A_{\text{flat}}$ is real-analytic on $[0, 1)$ and q^* is a simple transversal zero with $|D'(q^*)| = LM^2/(\kappa(c_f - c_s)) = \Theta(L)$. Holding δ fixed and letting $L \to \infty$, $\kappa \to 1$ and $M \to c_f - c_s$, so $|D'(q^*)| \to L(c_f - c_s)$ and $A^* \to a_s \exp(-\delta(1 - c_s)/(c_f - c_s))$.*

Proof. Both curves are degree-L polynomials in q, hence analytic; transversality is (A2) strictly, which needs $c_f \neq c_s$, guaranteed by existence; the value of $|D'(q^*)|$ is (R1). The limits follow from $w = (c_f - c_s)/M = 1/(1 + \delta(1 - c_s)/((c_f - c_s)L))$ and $A^* = a_s w^L$. □

The earlier limit constant was c_f; the correct one is the recovery gap. The interpolation to the Eigen error threshold as $L \to \infty$ (Eigen, 1971; Campos & Fontanari, 1998) is unaffected, so this is a crossover with a characteristic depth, not a phase transition at finite L (Campos & Fontanari, 1998; 1999).

The first-order flatness–hazard bridge referenced in Section 2 is

$$h_{\text{flat}}(q)/h_{\text{sharp}}(q) \xrightarrow[q\to 0]{} (1 - c_f)/(1 - c_s), \tag{9}$$

obtained from $h_S(q) = L(1 - c_S)/(1 - q(1 - c_S))$ at $q \to 0$; the earlier version of this work stated it as $1 - \text{NSR}$, the $c_s{=}0$ case. It is two-sided:

Lemma 1 (Two-sided flatness–hazard bound)**.** *For all $q \in [0, q^*]$, $\dfrac{1 - c_f}{\kappa(1 - c_s)} \leq \dfrac{h_{\text{flat}}(q)}{h_{\text{sharp}}(q)} \leq \dfrac{1 - c_f}{1 - c_s}$.*

Proof. Write $h_{\text{flat}}/h_{\text{sharp}} = (f/s)g(q)$ with $g = (1 - qs)/(1 - qf)$. Then $g(0) = 1$ and $g' = (f - s)/(1 - qf)^2 < 0$, so g decreases on $[0, q^*]$ from 1 to $g(q^*) = w/(\kappa w) = 1/\kappa$. □

The upper bound is the first-order law (9). The two bounds differ by exactly the factor κ, so the first-order law is accurate to relative error $(\kappa - 1)/\kappa \leq \delta/L$ on the whole interval, which is stronger than the earlier asymptotic $O(cq^*)$ and makes the $1/(1 - c_s)$ scaling unmissable. At $c_s{=}0$ it is the earlier lemma.

Proposition 3 (Kernel dependence)**.** *For each kernel k, a crossover exists on the aligned q-axis if and only if $c_{f,k} > \kappa\, c_{s,k}$, and then $q_k^* = (\kappa - 1)/(\kappa(1 - c_{s,k}) - (1 - c_{f,k}))$.*

Proof. Theorem 1 applied to the pair $(c_{s,k}, c_{f,k})$ that kernel k induces. □

Existence is therefore a condition on the *ratio* $c_{f,k}/c_{s,k}$ relative to the per-step tax, not on $c_{f,k}$ alone, and is not automatic across kernels. The earlier claim that existence and direction are kernel-independent is an artifact of $c_s{=}0$. This is a weakening, and it also supplies a second, measurable mechanism for a kernel null (a large $c_{s,k}$ rather than a small q), which the same probe can decide.

Proposition 4 (Optimal flatness)**.** *Let the flat schema's recovery, written c here as a design variable rather than the measured c_f, cost steps $L_f(c) = L + \beta(c)$ with $\beta(0) = 0, \beta' > 0$, and ceiling $a_f(c)$ with $a_f' \leq 0$. With $\Phi(c; q) = \ln a_f(c) + (L + \beta(c)) \ln(1 - q(1 - c))$ and $c^*(q) = \arg\max_{c\in[0,1]} \Phi$, we have $c^*(0) = 0$, and under the cost-regularity condition $L + \beta(c) > \beta'(c)(1 - c)(1 - q(1 - c))$, $c^*(q)$ is nondecreasing in q in the strong set order.*

Proof. $\partial_c \Phi = a_f'/a_f + \beta' \ln(1 - q(1 - c)) + (L + \beta)q/(1 - q(1 - c))$. At $q = 0$ the last term vanishes and the first two are ≤ 0, so the maximum is at $c = 0$. With $u := 1 - q(1 - c)$, the flat schema's own per-step survival rather than the altitude factor w, the cross-partial is $\partial^2_{cq}\Phi = (L+\beta)/u^2 - \beta'(1-c)/u$, positive exactly under the stated condition, so Φ has increasing differences in (c, q) and Topkis' theorem applies. Concavity in c is not needed. □

c_s does not enter Φ, which is the flat schema's own log-accuracy in its own design variable, so this result is untouched by the generalization; maximizing the difference or the log-ratio changes the objective only by an additive term independent of c_f. Only the threshold it references moves, to (6).

Lemma 2 (Exact interval for the budget)**.** *$q^*(\kappa, c_s, c_f)$ is strictly increasing in κ and in c_s and strictly decreasing in c_f. Hence if the three marginal intervals have level $1-\alpha$ each and the rectangle stays inside the existence region (it suffices that $c_{f,\mathrm{lo}} > \kappa_{\mathrm{hi}} c_{s,\mathrm{hi}}$), then $[\, q^*(\kappa_{\mathrm{lo}}, c_{s,\mathrm{lo}}, c_{f,\mathrm{hi}}),\ q^*(\kappa_{\mathrm{hi}}, c_{s,\mathrm{hi}}, c_{f,\mathrm{lo}})\,]$ covers q^* with probability at least $1-3\alpha$.*

Proof. Monotonicity is (7); the endpoints are the min and max of a monotone function over the rectangle, and a union bound over three marginals gives the coverage. The earlier two-argument version is the $c_s{=}0$ case, where the c_s marginal is degenerate and coverage improves to $1-2\alpha$. □

Lemma 2 is exact, not asymptotic. On arithmetic a Wilson interval on $c_s{=}0.0075$ from 670 load-bearing trials is $[0.003, 0.017]$, which moves q^* by at most 0.003, against a reported 95% interval of half-width 0.029 on the same cell.

Proposition 5 (Concentration of the empirical crossover)**.** *With N i.i.d. problems per grid point and accuracies $\geq A_{\min} > 0$ near q^*,*

$$|\hat{q}^* - q^*| = O_p\Big(\frac{\kappa(c_f - c_s)}{A_{\min} L M^2 \sqrt{N}}\Big) = O_p\Big(\frac{\kappa w}{A_{\min} L M \sqrt{N}}\Big).$$

Proof. Hoeffding and a union bound over the finite grid give $\sup |\hat{A}_S - A_S| = O_p(N^{-1/2})$; the delta method with the lower bound $A_{\min}$ gives $|\hat{D}(q^*) - D(q^*)| = O_p(1/(A_{\min}\sqrt{N}))$; divide by $|D'(q^*)|$ from Theorem 4. □

The same quantity $M/(\kappa w)$ therefore governs both the separation rate of Proposition 1 and the localization width, so a wide interval and an unresolved reversal are one phenomenon. Numerically the bound predicts the stack family's interval should be about $2.3\times$ wider than arithmetic's at matched L and N; the half-widths in Table 2 are 0.101 against 0.029, a ratio of $3.5\times$. The bound is therefore in the right direction but understates the widening by a factor of 1.5, consistent with the sample-size bound of Proposition 1 not being tight on this family.

A.6 Numerical verification

A verification script checks the whole of this appendix against direct root-finding and finite differences over randomized configurations, with a deterministic subset pinned in continuous integration. Headline residuals: the closed form against bisection root-finding of $A_{\mathrm{sharp}} - A_{\mathrm{flat}}$, maximum absolute error 2.3×10^{-14} over 2000 admissible configurations; reduction to the earlier form at $c_s{=}0$, 2.1×10^{-15} over 2000; the existence condition $c_f > \kappa c_s$ against a dense grid scan for a sign change, zero violations in 2000 draws, of which 105 satisfy $M > 0$ with no crossing and none of the 870 draws with $M \leq 0$ had one; the three comparative statics against central differences, maximum relative error 5.3×10^{-9}; Proposition 1 (R1) to (R4) against numerical differentiation and grid maximization, maximum relative error 5.6×10^{-9} and maximum absolute error 2.6×10^{-8}; the two-sided hazard bounds, zero violations over 50000 (configuration, q) pairs; the reliable depth, residual 7.4×10^{-15}, and the existence depth $L_{\min}$, zero mismatches over 17700 pairs; and dq^*/dL under drift, maximum relative error 2.5×10^{-7}, at the finite-difference truncation floor.

B The calibration cells

Table 2 lists the ten calibration measurements behind Figure 1c, with the measured parameters each prediction is computed from and the sample size per grid point. Nine of the ten are the arithmetic family and one is the stack machine. The first row and the Qwen2.5-7B $L{=}8$ row are the same condition, that is the same schema, task, model and depth, measured through two pipelines on independent problem draws; their noiseless sharp accuracies are 0.852 and 0.855. The headline fit counts the nine distinct conditions and treats the pair as a replication, giving $R^2 = 0.839$ and mean absolute error 0.0208 about $y{=}x$ over $n{=}9$; counting the pair twice gives $R^2 = 0.855$ and

0.0191 over n=10, the difference being the near-exact agreement of a replication with itself. The corresponding accuracy curves for the four cross-model cells are in Appendix G, Figure 4.

Table 2: The ten calibration measurements behind Figure 1c, which plots the nine distinct conditions. κ, c_s and c_f are measured independently of the sweep; p^*_{pred} is (6) evaluated at them and mapped from effective-error units to the kernel knob as $p^* = q^*/q_{\text{eff}}$ at that sweep's sharp schema q_{eff} (Section 3). N is problems per grid point, pooled over seeds $\{42, 123, 456\}$; the depth sweep uses 1000 at p=0.1 and 400 at p=0.05. Measured p^* is the sustained-reversal crossover with its paired bootstrap 95% CI, except for the stack row, whose point-estimate crossing is reported because the location prediction is testable against the bootstrap interval even though the reversal itself does not resolve (Table 1). The anchor row repeats the Qwen2.5-7B L=8 condition through a second pipeline and is excluded from the n=9 fit. Phi-4's sharp schema is at ceiling (160/160 correct at p=0), so its κ (†) is a boundary estimate that can only move down.

line	cell	N	κ	c_s	c_f	p^*_{pred}	measured p^* (95% CI)
continuum	re-derive every step	600	1.051	0.006	0.537	0.090	0.112 [0.087, 0.131]
	re-derive every 2nd step	600	1.084	0.006	0.386	0.187	0.162 [0.136, 0.215]
	re-derive every 3rd step	600	1.054	0.006	0.214	0.211	0.172 [0.139, 0.214]
	re-derive every 4th step	600	1.035	0.006	0.133	0.221	0.252 [0.136, 0.353]
cross-model	Qwen2.5-7B-Instruct, L=8	600	1.034	0.007	0.414	0.079	0.073 [0.052, 0.109]
	Qwen2.5-7B-Instruct, L=12	600	1.041	0.005	0.332	0.116	0.120 [0.092, 0.154]
	Phi-4-mini-instruct, L=8	160	1.029	0.018	0.561	0.053	0.041 [0.022, 0.083]
	Phi-4 (14B), L=8	160	1.042†	0.011	0.359	0.113	0.105 [0.076, 0.229]
cross-family	stack machine, L=8	600	1.026	0.454	0.640	0.156	0.196 [0.123, 0.325]
replication	anchor = Qwen2.5-7B L=8	600	1.030	0.006	0.434	0.068	0.072 [0.051, 0.104]

C TASK FAMILIES: CONSTRUCTION, COVERAGE, AND CAVEATS

All seven families expose the same interface (initial state, L verifiable transitions, exact terminal verifier, corruption operator), and every family invariant is checked in tests: |states| = L + 1, the initial state equals `states[0]`, the answer equals `states[-1]`, and chaining the single-step transition reproduces the state sequence.

Iterated arithmetic. From v_0, apply $v_{k+1} = (a_k v_k + b_k) \bmod 100$ with a_k=1 with probability 0.7 and a_k=2 otherwise, and $b_k \in \{0, \ldots, 12\}$. The interleaved doublings block a closed-form summation shortcut. One inherent property, not a bug: doubling steps under modulus 100 are four-to-one, so some near-miss corruptions alias back to the correct answer; the load-bearing filter discards such steps, 18 of 688 on the sharp schema at L=8.

Stack, register automaton, permutation. Synthetic generators over, respectively, push/pop traces, two-digit symbolic rewrites (reverse digits, add a constant), and composition of permutations on a small symbol set. The permutation family was retuned from six symbols to four after the six-symbol configuration floored both schemas. Its alphabet size and its swept depth coincide at four, which are distinct parameters: the alphabet is a fixed configuration constant, whereas the depth is the difficulty axis, so the tax ratio for that cell is $\kappa = (a_s/a_f)^{1/4}$.

Code execution (derived). Built from a public code-reasoning corpus (Gu et al., 2024) by tracing loop accumulators. Only 64 of 800 candidate problems (8.00%) admit a length-controllable loop with a step-by-step verifiable accumulator. This family's answer is the accumulator immediately after the loop, which can differ from the corpus's own output field when a function post-processes after the loop, so it is described as derived rather than as the benchmark.

Kinship composition (derived). Chains are generated from the composition algebra in the style of a relational-reasoning corpus (Sinha et al., 2019), which supplies relation frequencies, entity names and surface spellings (an 18-relation vocabulary that contains son- and daughter-in-law and

no spouse relations). Rejection sampling to reach a target hop count grows geometrically, from 2.6:1 at two hops to 4062:1 at ten, so accepted chains at high hop counts are a heavily conditioned subsample and we use two to six hops. Measured compositional-class drift confirms the concern: from four to ten hops the grandparent class falls from 8.7% to 3.3% and the aunt/uncle class from 17.8% to 10.8%, so hop count is not a cleanly isolated difficulty axis at high hops.

GSM-Symbolic linear chains (derived). Only 30.2% of the corpus (Mirzadeh et al., 2025) consists of linear, step-verifiable chains (45.1% of `main`, 25.0% of `p1`, 10.7% of `p2`); the released corpus does not carry inline calculator annotations, so chains are recovered by parsing the narrated prose. The level $(\texttt{p2}, 6)$ has zero eligible instances and is dropped, and a degenerate $L{=}1$ level was removed. Unlike the other six families, this one's first state is computed rather than given, so that state is neither scored nor corruptible and its intrinsic-slip probe compares against the reference value rather than the model's own.

Screen-only cells. The three derived families in Table 1 were screened but not swept, and their tabulated κ comes from the $N{=}64$ family screen. Re-measured at $N{=}200$, the noiseless sharp/flat accuracies are $0.345/0.245$ (code execution), $0.180/0.135$ (kinship) and $0.335/0.105$ (GSM linear chain), so those point estimates are loose even though the qualitative verdict, a crossover beyond the swept noise range, is what the screen was used for.

Withdrawn claims. Two results reported in the earlier version of this work do not survive the sustained-reversal criterion of Section 3 and are withdrawn: a register-automaton crossover ($p^*{=}0.35$ at $L{=}6$ and $p^*{=}0.50$ at $L{=}4$) and the $L{=}16$ point of the arithmetic depth sweep ($p^*{=}0.200$). Both were recomputed from the same raw records under the corrected criterion, with no raw record modified and the pre-correction summary artifacts left on disk as the historical record. At $L{=}12$ the crossover was re-confirmed at matched $N{=}600$ with a *tighter* interval than originally reported. Section 4.4 states both withdrawals, and separately reports the sample-size bound that accounts for them after the fact.

D Extended related work: the three surrounding lines

This section states the three lines summarized in Section 1 with their citations, then distinguishes two recent results that share our vocabulary.

(i) Absolute decay characterizes how a *single* procedure's accuracy falls with length (Gan et al., 2025; Sinha et al., 2025; Kwa et al., 2025; Wu et al., 2025; Arbuzov et al., 2025; Ton et al., 2024; Wan et al., 2025; Chapra et al., 2026); we order *two length-matched* procedures, so decay common to both cancels. *(ii) Noise mitigation* pushes accuracy up through sampling, verification, search and runtime denoising (Wang et al., 2023; Lightman et al., 2024; Wang et al., 2024; Yao et al., 2023; Besta et al., 2024; Chen et al., 2023; Yan et al., 2026); we ask when robustness should be preferred to precision at all. Since unaided self-revision reduces uncertainty about the target by close to nothing (Huang et al., 2024; Tao et al., 2026; Kamoi et al., 2024; Tyen et al., 2024), our flat schema recovers by structural redundancy, not introspection, and cross-checks inside the trace rather than in a post-hoc revision pass, where the format-repair and role-relabeling artifacts of that literature live (Chen et al., 2026a;b). *(iii) Fragility and lucky survival*: perturbing intermediate steps reliably breaks reasoning (Aravindan & Kejriwal, 2026; von Recum et al., 2026; Mirzadeh et al., 2025; Chen et al., 2024; Shi et al., 2023; Havrilla & Iyer, 2024), yet a terminal answer can survive a step it ignored (Jiang et al., 2025; Chen et al., 2025; Lanham et al., 2023; Zhou et al., 2024), which inflates any naive robustness measure, so we credit recovery only on steps whose corruption determines the answer, for *both* schemas. Our flatness is over reasoning *trajectories*, distinct from flat minima over network *weights* (Foret et al., 2021; Keskar et al., 2017; Hochreiter & Schmidhuber, 1997), and the tax is the trajectory-level analogue of the weight-space robustness tradeoff (Tsipras et al., 2019; Zhang et al., 2019; Madry et al., 2018).

Two in-window results share the word crossover without sharing the phenomenon. Nie et al. (2026) locate a closed-form crossover between thinking and non-thinking modes driven by a shared output-token budget, where the mechanism is answer truncation rather than error repair, and Zhang et al. (2026) decompose chain-of-thought risk into an oracle-trajectory term and a trajectory-mismatch

term with bounded, linear and exponential error-growth regimes, a risk decomposition for one procedure rather than an ordering of two. Neither has an existence condition.

E REPRODUCIBILITY DETAILS

The code, the per-cell records behind every number below, and the analysis pipeline that regenerates them are at `https://anonymous.4open.science/r/reasoning-noise-budget-DF74` and in the supplementary archive; its README gives the exact commands for each of the settings stated here. Sample sizes are per cell (Table 2), pooled over seeds $\{42, 123, 456\}$: N=600 per grid point for the arithmetic depth sweep at L=8, 12, 16 (with 1000 at p=0.1 and 400 at p=0.05), for the stack and permutation sweeps and for every continuum variant; N=160 at every grid point for Phi-4-mini-instruct, Phi-4 and the Qwen2.5-1.5B-Instruct control; N=200 for the register automaton and the K3 sampling kernel, and 80 to 96 for K2; N=64 for the family screen. Accuracy intervals are paired bootstrap 95%. Inference is greedy via vLLM except where a kernel injects sampling. K2 adds Gaussian noise of standard deviation 0.15 times the per-token residual-stream RMS at each step boundary with probability p; K3 raises the decoding temperature to 1.3 with probability p. Theory is stated in the effective per-step error q while figures plot the kernel knob p; for K1 on arithmetic the two nearly coincide ($q_{\text{eff}} \approx 0.97$, against 0.82 on the stack family), and predictions in Table 2 are mapped through the sharp schema's measured q_{eff} for that sweep. The hardware and serving stack changed partway through this work, so the arithmetic family was re-baselined: at L=12, p=0 on a fresh seed the new stack gives sharp 0.790 and flat 0.480 against the earlier 0.775 and 0.480, which is why seeds are pooled across machines for arithmetic. A format pre-flight over ten open-weight models finds a parse rate of 1.000 for both schemas on every model (Wilson $[0.893, 1.0]$, n=32).

F TRAINING TOWARD FLAT SCHEMAS

Proposition 4 says the accuracy-maximizing amount of redundancy rises with the noise budget, and it survives the generalization untouched; only the threshold it references moves to (6), which on a family with a self-correcting sharp baseline is later by the factor $(\kappa-1+c_f)/(\kappa-1+c_{\text{net}})$, a factor of 3.3 on the stack family. The training experiment below tests that prescription on the arithmetic family, where the correction is negligible. From a shared pool of self-generated flat-schema traces (Qwen2.5-7B-Instruct, temperature 0.8) we build reward-filtered fine-tuning sets and apply reward-weighted LoRA fine-tuning (rank 16, scaling 32, dropout 0.05, on `q,k,v,o,gate,up,down`; AdamW at learning rate 1.5×10^{-4}, 3 epochs, batch 4 with gradient checkpointing; adapters merged for evaluation). **Flat (ours)** keeps traces that are correct *and* robust under the load-bearing-filtered recovery reward; **B1** keeps terminally-correct traces (outcome-only RLVR) (DeepSeek-AI et al., 2025; Shao et al., 2024; Sheng et al., 2025); **B2** keeps traces agreeing across input paraphrases (Miyato et al., 2018; Sohn et al., 2020; Liang et al., 2021; Gao et al., 2021); **B3** fine-tunes on traces with corrupted intermediates but correct terminal labels (Zhou et al., 2024); **B4** does no training and votes over matched inference samples (Wang et al., 2023).

Results are in Table 3 and Figure 7. The flat reward removes the noiseless tax, reaching 0.778 at low noise against the sharp base's 0.744 while the untrained flat schema sat at 0.512, and it beats every reasoning baseline at high noise (0.380 against outcome-RLVR's 0.260, consistency's 0.249 and self-consistency's 0.333), trained on the fewest examples because the robustness filter is strictest, though each method is a single run without intervals. It loses to corrupted-trace augmentation (0.601). Corrupting a load-bearing intermediate changes B3's answer only 1% of the time (answer-invariance 0.99) against 66% for the flat-trained model, 60% for the base and 82% for outcome-RLVR, while B3's clean traces are as internally consistent as the flat model's (step-fidelity 0.99 against 0.98). B3 therefore writes faithful-looking steps whose values its answer ignores, the "robust-answer, fragile-logic" decoupling (Jiang et al., 2025) that the load-bearing filter deliberately excludes from our reward. Two caveats: the fine-tuning sets are unequal in size because each method keeps only what its filter admits (59 for Flat against 98 for B3), and this is a single-family demonstration. The phenomenon and the theory do not depend on it.

Table 3: Matched-compute comparison on arithmetic (mean terminal accuracy under K1, L=8). Low-noise is $p \leq p^*$, high-noise is $p > p^*$; fine-tuning set sizes in parentheses.

method	low-noise ($p \leq p^*$)	high-noise ($p > p^*$)
base (sharp)	0.744	0.200
base (flat)	0.512	0.250
B1 outcome-RLVR (98)	0.716	0.260
B2 consistency (77)	0.650	0.249
B4 self-consistency	0.578	0.333
Flat-reward, ours (59)	**0.778**	**0.380**
B3 corrupt-aug (98)	0.778	<u>0.601</u>

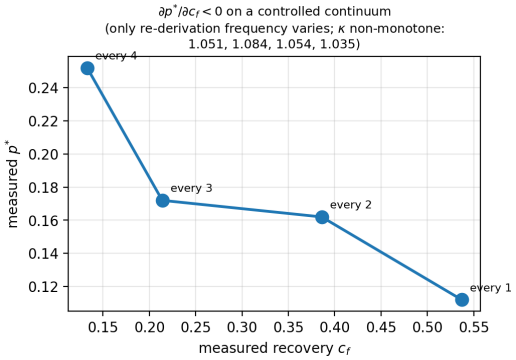


Figure 2: **The comparative static on a controlled continuum.** Changing only how often the flat schema re-derives its previous value moves measured c_f and measured p^* in opposite directions, as $\partial q^*/\partial c_f < 0$ requires (Section 4.2).

G ADDITIONAL FIGURES

Figure 3 shows the error-process pre-screen of Section 4.3, Figure 4 the cross-model accuracy curves and the padded-length control, Figure 5 the two measurement instruments, and Figure 6 the simulation of the survival process free of model confounds.

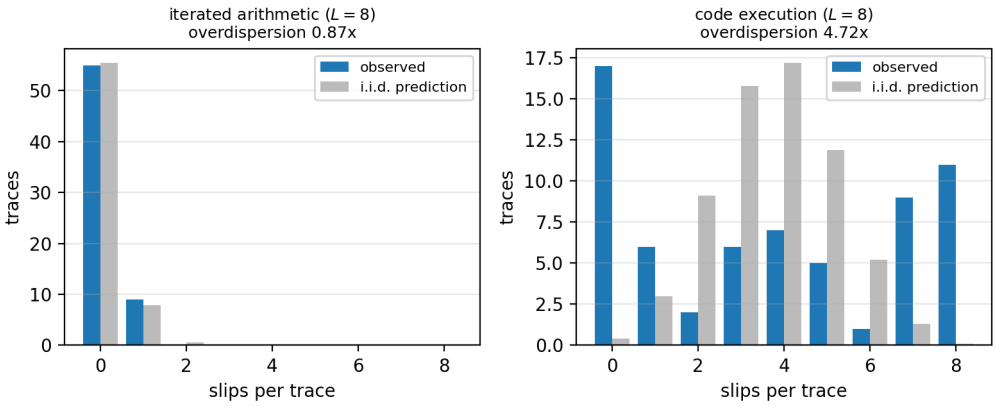
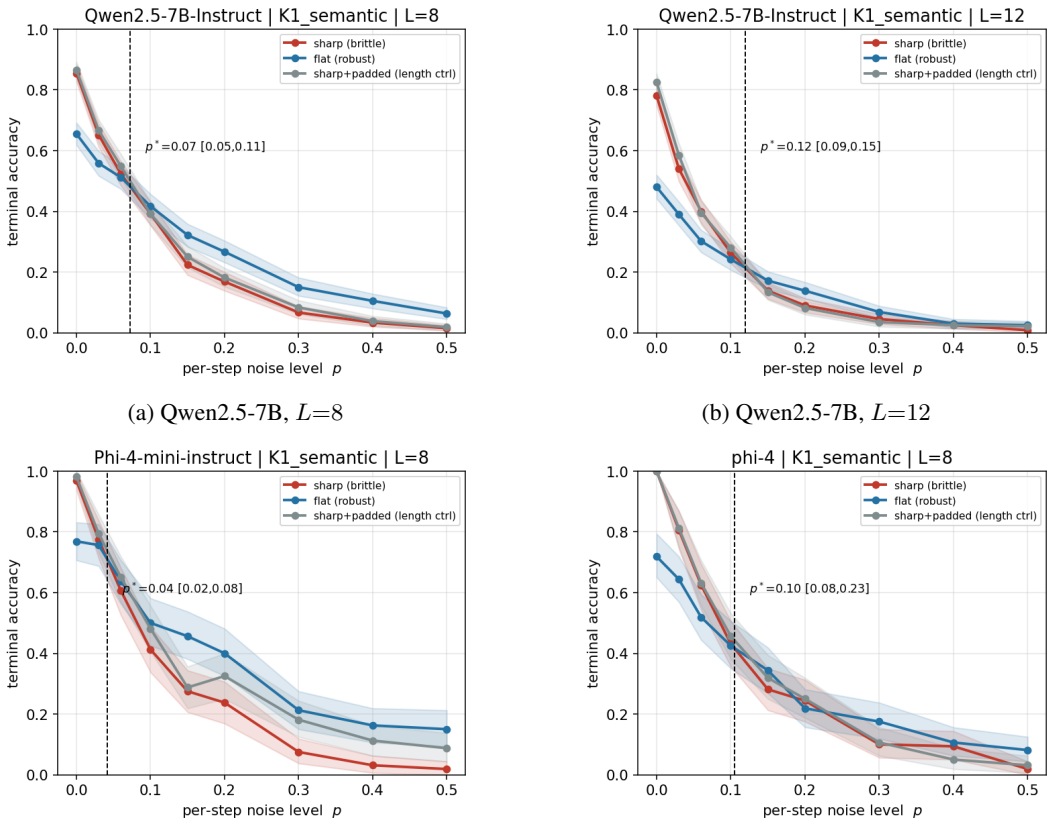


Figure 3: **The error process decides whether redundancy can help.** Per-trace slip counts on zero-injection traces against the $\mathrm{Binomial}(L, q_0)$ an independent step-channel predicts (N=64, Qwen2.5-7B-Instruct, L=8): arithmetic shows no detectable overdispersion at this sample size ($0.87\times$, with only 9 of 64 traces carrying any slip), while code execution is U-shaped ($4.72\times$) with 26.6% perfectly clean traces against the 0.7% independence predicts. The screen is used to detect gross violation, not to certify independence.



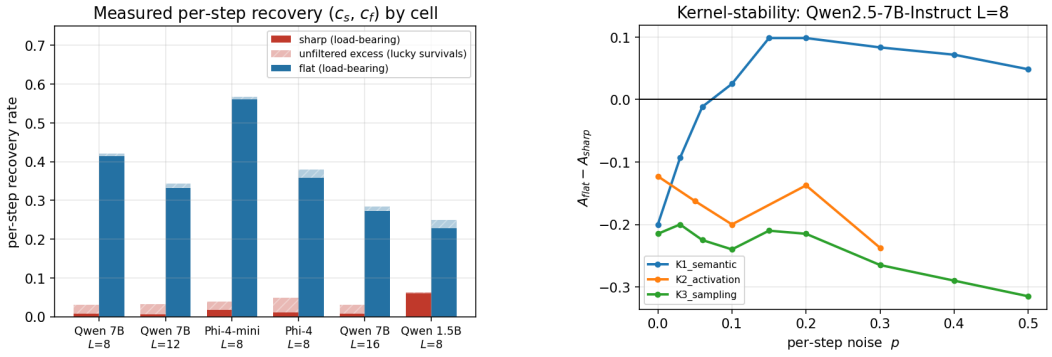
(a) Qwen2.5-7B, L=8

(b) Qwen2.5-7B, L=12

(c) Phi-4-mini-instruct

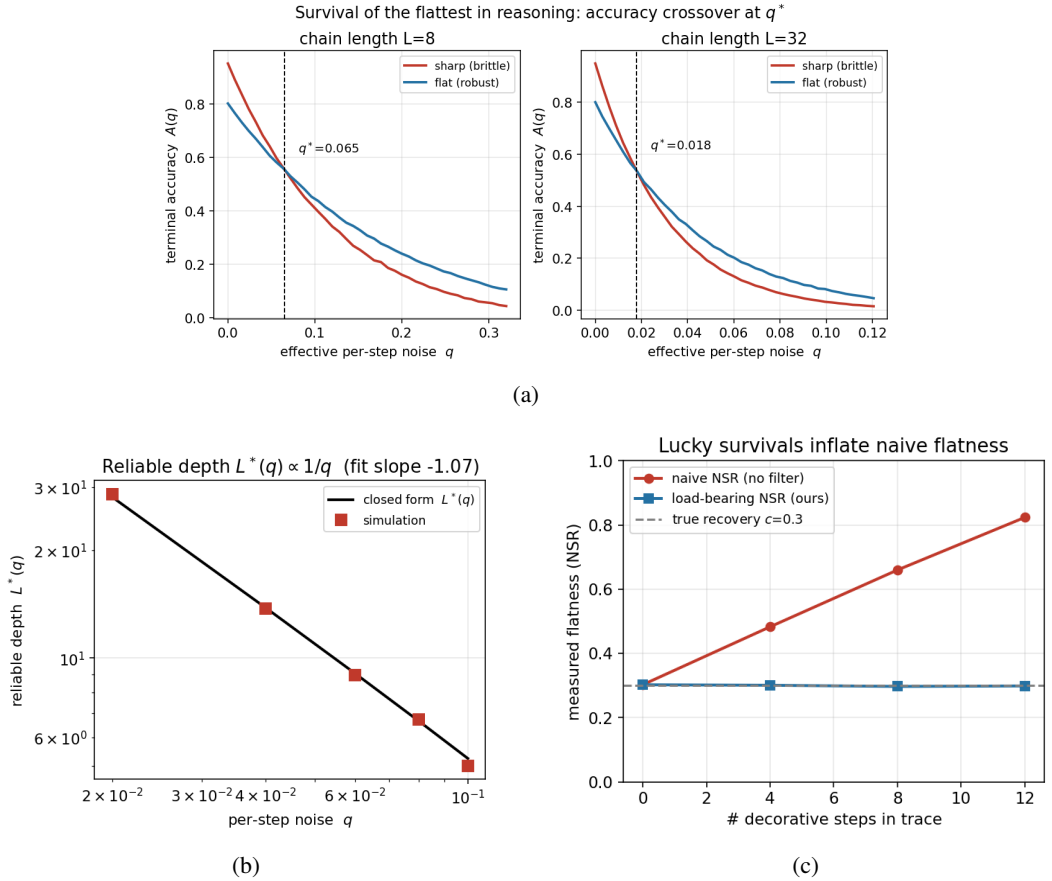
(d) Phi-4 (14B)

Figure 4: **The cross-model line of Table 2.** Terminal accuracy against per-step semantic noise for the four model and depth cells whose budgets the closed form predicts from measured tax and recovery. In every panel the token-matched but brittle padded control never produces the reversal, so the effect is robustness structure and not length; its noiseless accuracy sits at or above the sharp schema's, so verbosity by itself neither costs accuracy nor buys a reversal.



(a) (b)

Figure 5: **Measurement instruments.** (a) Load-bearing per-step recovery, the c_s and c_f the closed form is fed, for both schemas on arithmetic across the six model-by-depth cells. The solid bar is the load-bearing rate the paper reports; the hatched cap on top is the extra the unfiltered rate would have claimed, that is the lucky survivals the load-bearing filter removes, which inflate the sharp schema's apparent recovery several-fold precisely because its true recovery is near zero. (b) $A_{\text{flat}} - A_{\text{sharp}}$ against the noise knob under the three kernels: only the content-corrupting semantic kernel reaches the budget and crosses zero, while activation and sampling noise leave the ordering stable and never resolve a significantly positive difference at any grid point.



(a)

(b) (c)

Figure 6: **Simulation of the survival process**, free of model confounds. (a) accuracy crossover at two depths; (b) reliable depth scaling with the effective error; (c) decorative steps inflate a naive recovery rate while the load-bearing filter recovers the true one.

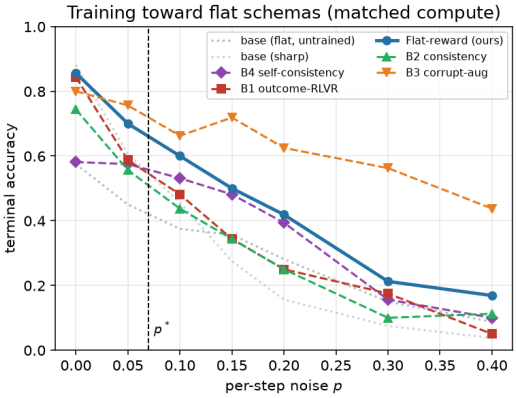


Figure 7: **Training toward flat schemas** (Appendix F). Terminal accuracy against per-step noise after matched-compute fine-tuning, one run per method and no intervals.