Conflicting Supervision Moves Commitment, Not Capability

Anonymous authors
Paper under double-blind review

Abstract

Post-training corpora routinely contain the same problem written under several equally correct conventions, and practitioners ask whether the order in which such data is presented matters. The published record is divided: large ordering effects and none at all are both reported, and the disagreement is usually attributed to task or scale. We show it is attributable instead to the learning-rate schedule – something every run fixes and no run reports as a variable. We prove a bound in which the arrangement and the learning-rate schedule enter an ordering effect as separate multiplied factors – the arrangement only through the period of its alternation, the schedule only through how much weight the endpoint can place on any one moment of the run – so the schedule is not a background condition for the question but the averaging operator that decides it. Tested on ten orderings of one conflicted corpus, run twice under families differing in lr_scheduler_type and nothing else, the interior spans 11.63 contrast floors at a constant rate and is monotone in how blocked the arrangement is, while under the single cosine every published arm uses the same ten arms occupy two distinguishable states where their own resolution would allow about ten; three mechanisms we had preregistered were falsified. What the path writes is not capability but commitment: across twelve arms accuracy summed over both conventions is constant to within 9.7% while the allocation share runs 0.04 to 0.87, so a 12.29σ arrangement effect is exactly zero under a convention-agnostic score. An ordering result is therefore uninterpretable without the schedule it was measured under, and marking the convention in the prompt collapses the effect and recovers 87.5% of the union ceiling.

1 Introduction

Train a model on the same problems written under two incompatible conventions, both correct, and one thing is obvious in advance: something is lost. The interesting question is what.

The answer is that nothing need be lost from what the model can solve, and a great deal moves in which correct form it commits to. Twelve arrangements of one corpus hold $\mathrm{acc}_A + \mathrm{acc}_B$ constant to within 9.7% while the share written under one convention runs from 0.04 to 0.87. The same measurement is a 12.29σ effect on one coordinate and exactly zero on the other.

Whether that commitment survives training at all is decided by the learning-rate schedule, and this is the paper's central result. We prove that the ordering-dependent part of an endpoint is bounded by a product of two factors that do not communicate: the arrangement enters only through the period of its alternation, the schedule only through how much weight the endpoint can place on any one moment of the run (Proposition 2): a decaying schedule cannot put a large step size and an uncontracted remainder at the same moment, and a constant one does exactly that at the last step. The prediction is a knob and the measurement is that knob at two settings: ten arrangements of one corpus, one budget, the same three seeds, two families differing in lr_scheduler_type and in nothing else, span 2.20 contrast floors under a cosine and 11.63 under a constant rate. "Order matters" and "order

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.

does not matter" are the two ends of one knob, which is what the divided empirical record on data ordering looks like from here.

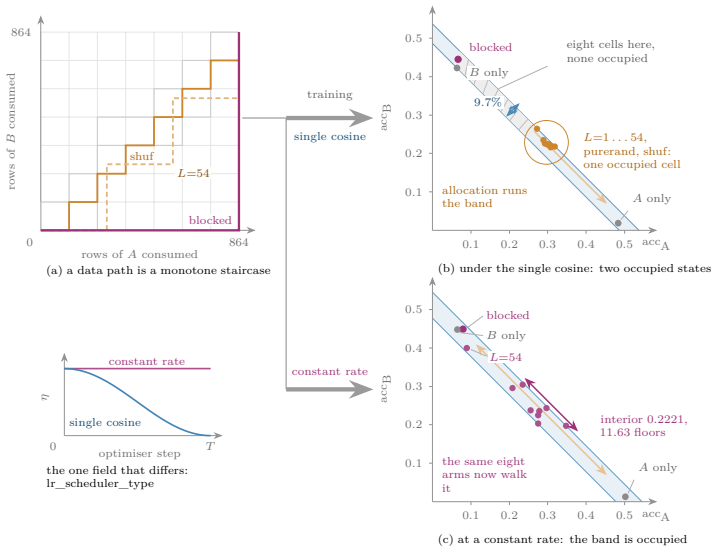
The path space, and the limit that collapses it. Every arrangement of a two-source corpus sits on one ladder: finish A entirely, then B (blocked); alternate blocks of L optimiser steps ($A^L B^L \cdots$); alternate every step (L=1); mix within each batch (shuf). Only the order varies, and the order is a message carrying $\log_2 \binom{2n}{n}$ bits whose receiver is the parameter vector at the end of training. At the ideal end of the ladder that receiver is deaf: as $\eta \to 0$ with alternation frequency $\to \infty$, the averaging theorem (Bogoliubov & Mitropolsky, 1961; Kapitza, 1951) sends the trajectory to the mean-field flow $\dot{\theta} = -\nabla[\frac{1}{2}L_A + \frac{1}{2}L_B]$, whose cross-entropy minimiser at a contested input is $q^\star = (p_A + p_B)/2$ — under conflict the infinitely divided limit installs the Bayes-optimal mixture, and every order lands there. The gap between that limit and the budget anyone actually trains at is what this paper measures. Across every arm that differs only in path (L=1 to 54, purerand, shuf, blocked; three seeds) the allocation runs 0.407 to 0.869, a range that would separate about ten values at the seed dispersion's $2\hat{\sigma}$ resolution. It separates two: nine arms in $[0.407, 0.492]$ and blocked alone at 0.869, the largest gap inside the occupied group being 8.7× smaller than the gap to blocked (20.7× on a second pretraining family). Every arm just counted trains under a cosine, and that is the point rather than a caveat: the same ten paths at a constant rate occupy four states. The two-state count is not a fact about what a path can carry; it is what survives after a decaying schedule has integrated the path away.

Contributions. (i) The ordering effect factors, and the schedule is one of the two factors (Proposition 2): the π-dependent part of an endpoint is bounded by $\tau_L \cdot \sup_t \|w\|$. We claim no new averaging theorem — the $\tau_L \to 0$ corner is classical incremental-gradient theory — and no prediction of magnitude; what is new is that the two factors separate, which turns a divided empirical record into one knob. (ii) The knob, measured at both settings (§5.2), monotone in how blocked the arrangement is, with the corner case $\tau_L = T$ measured against the bound. (iii) A decomposition separating what a model can do from which form it writes (Definition 2), computable from the per-problem scores an evaluation already produces; under it this paper's largest effect and a null are one measurement on two axes. (iv) The stopping phase, and a query-time key: blocked is the lowest-frequency alternation stopped at maximum swing, a mirror test predicts a sign and a symmetry in advance and finds them on two corpora (§5.3), and marking the convention in the prompt collapses the arrangement effect (§5.4). (v) A record of what did not work: three pre-registered mechanisms died, one positive control inverted, one came back void. Appendix A states five things this list does not claim, and §6 the scope limits on the results themselves.

2 Related Work

A conflict is an error — the premise almost everyone shares. The knowledge-conflict literature names our object exactly (Xu et al., 2024 call it intra-memory conflict) and every remedy it surveys is a repair (Xue, 2026), which presupposes that one of the conflicting forms is wrong. Our natural corpus is built so that neither is, and Proposition 1 then makes the mixture the loss-minimising policy rather than a malfunction; Krestnikov (2026) come nearest to dropping the premise, and the same premise inverted organises the evaluation side (Plank, 2022), where perspectivist repairs also need a target unavailable here. Schaeffer et al. (2023) is the closest precedent and the closest warning — emergence turned out to be a property of the metric — and §5.1 performs that move on a different axis. Holtzman et al. (2021) names the mechanism by which surface forms compete for probability mass, and that competition is our S; Janeiro et al. (2026) average the surface-form term away, which is right when phrasing is nuisance, where we keep it as a coordinate because in a conflicted corpus the phrasing is exactly what arrangement moves.

Where the propositions come from, and what sits inside the wall. In the deterministic cyclic case Proposition 1 is the central dichotomy of the incremental-gradient literature (Bertsekas, 2011; Nedić & Bertsekas, 2001) — decaying step to the summed minimiser, constant step to a limit cycle whose position depends on the order — with Proposition 2 interpolating



(a) a data path is a monotone staircase

(b) under the single cosine: two occupied states

(c) at a constant rate: the band is occupied

Figure 1. The space of data paths, and its image under each of the two schedules. (a) A data path is a monotone lattice path: shuf hugs the diagonal, a rung is a coarser staircase, blocked is the corner. (b,c) Its image in the read-out's coordinates (Definition 2); the panels differ in one field of the trainer configuration and in nothing else, and the annotated spans are measured (§5.2).

and Proposition 3 the limit cycle. What we add is the transport: that literature fixes the objective, and the question here is what the same dichotomy does to a policy when the summed minimiser is a mixture. Batch composition (Rao et al., 2024; Dai et al., 2026; Lu et al., 2022), optimiser state (Sweeney, 2026b), one-shot commutator scores (Sweeney, 2026a) and per-step gradient surgery (Yu et al., 2020; Liu et al., 2021; Navon et al., 2022) all operate strictly inside the averaging wall, whereas methods reweighting proportions (Xie et al., 2023; Fan et al., 2024; Gu et al., 2025) act on the quantity Proposition 1 leaves free; forgetting (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017; Wu et al., 2024) is a movement of C where the conflict switch is a movement of S.

Ordering, its nulls, and the sharpest counter-result. The record on ordering is divided (Bengio et al., 2009; Rohrer & Taylor, 2007; Say et al., 2025; Jia et al., 2025; Elgaar & Amiri, 2026; Luo et al., 2025), and the division is what the averaging wall predicts. Luo et al. (2025) find curriculum advantages that hold at a constant rate and diminish under standard decay, in pretraining at 1.5B parameters; we reproduce that moderator under control at 7B in supervised fine-tuning, and where their reading is that decay wastes the curriculum, ours is that decay is the averaging operator — so a paper reporting an ordering effect without its schedule has not reported a moderator its own effect depends on. Against all of this, LeDoux (2026) train small networks on modular arithmetic from scratch, where two fixed orderings reach 99.5% test accuracy and random ordering does not, wide enough to be named a covert information channel. Both results are right, and what the path can write depends on how much the prior has already fixed: order is a wide channel into an empty state and a narrow one into a full state. Appendix B extends all three groupings.

3 Method: the Path, the Channel, and the Two Coordinates

Definition 1 (Data path and the division ladder). Fix a multiset $\mathcal{D} = \mathcal{D}_A \uplus \mathcal{D}_B$ from two sources. A data path is an ordering π of $\mathcal{D}$; training is T optimiser steps of step size η along π. The division ladder is the family π_L alternating same-source blocks of exactly L consecutive optimiser steps, with L=1 the finest alternation and shuf the within-batch mixture. One source's per-epoch allocation is 54 steps and a rung is one arranged one-epoch file passed over three times (324 steps); blocked is not a rung but three epochs on A then three on B, so its effective block is L=162. Every arm differs from every other only in π, at fixed $\mathcal{D}$, T, η and seed.

Definition 2 (Capability and allocation). Let the two evaluation sets be the same problems under mutually exclusive conventions, so a generation matches at most one gold. Write $C = \mathrm{acc}_A + \mathrm{acc}_B$ (capability) and $S = \mathrm{acc}_B / C$ (allocation), so $(\mathrm{acc}_A, \mathrm{acc}_B) = (C(1-S), CS)$ is a change of coordinates, not a model. With $p(x)$ the probability the model solves x and $s(x)$ the probability it then answers under B, mutual exclusivity gives $C = \mathbb{E}[p]$ for any s whatever, which is why capability is the robust half; the allocation is $S = \mathbb{E}[ps]/\mathbb{E}[p]$, and it equals the convention policy $\mathbb{E}[s]$ exactly when $\mathrm{Cov}(p, s) = 0$ across problems. We measure that covariance, and it does not vanish.

Endpoint states, and a clause that fails. Two endpoints count as distinguishable when separated by $2\hat{\sigma}$; the counts below are single-linkage cluster counts at that resolution, not mutual informations (Appendix G.8 sweeps the constant). Definition 2's covariance clause is testable and fails: problems solved on all four draws are answered under B with probability 0.4478 against 0.3491 for problems solved on some but not all (5.17 floors, $z = 3.56$), so problems the model half-solves drift toward the pretrained convention. Conservation of C is unaffected, being measured rather than derived from the clause; what does not stand is reading S as a policy a capability change cannot touch, and that coupling is bounded at 0.099 against the 0.4621 the path family spans.

Proposition 1 (The infinite-division limit). Along π_L, let θ evolve by gradient descent on the per-example losses. As $\eta \to 0$ with $L\eta \to 0$ and $T\eta$ fixed, the trajectory converges uniformly on compacts to the mean-field flow $\dot{\theta} = -\nabla[\alpha L_A + (1-\alpha) L_B]$, $\alpha = n_A/(n_A + n_B)$, which depends on π only through the source proportions. For cross-entropy losses and a contested input x carrying $p_A(\cdot \mid x)$ and $p_B(\cdot \mid x)$, the flow's minimiser at x is $q^\star = \alpha p_A + (1-\alpha) p_B$.

Two consequences are immediate. Any intervention leaving the period-average of the field unchanged leaves the endpoint unchanged in the limit, so batch composition at fixed block structure and any time-symmetric reordering of a block cannot work — both were pre-registered as mechanisms and both are dead (§5.2).

Remark 1 (Arrangement cannot escape underdetermination). Let I be what the model may condition on when answering a held-out query. If the applicable convention is not a function of I ($H(\mathrm{conv} \mid I) > 0$) then no path attains zero loss on the contested inputs, and the minimiser of expected loss is the mixture of Proposition 1 for every π. Any π-dependence of a finite run's endpoint must therefore be a failure to reach that optimum — an escape-hatch effect — and all three positive findings below sit in that hatch. This is not a capacity limit — a 7B model has room for both conventions — so the escape is not more addressable bits but putting the convention into what the query can condition on, which is what the key of §5.4 does.

Proposition 2 (The schedule sets the reach, the arrangement sets the period). Let θ follow $\dot{\theta} = -\eta(t) \nabla L_{\pi(t)}(\theta)$ on $[0, T]$, and $\bar{\theta}$ the same flow with $\nabla L_{\pi(t)}$ replaced by $\bar{g} = \nabla[\alpha L_A + (1-\alpha) L_B]$. Write $\Delta = \nabla L_B - \nabla L_A$ and let s_π be the square wave equal to $1-\alpha$ on A-blocks and $-\alpha$ on B-blocks, so $\bar{g} - \nabla L_{\pi(t)} = s_\pi(t) \Delta$ with zero mean over a period τ_L. Define the endpoint weight $w(t) = \Phi(T, t)\, \eta(t)\, \Delta(\bar{\theta}(t))$, with $\Phi(T, t)$ the state-transition operator of the mean flow linearised about $\bar{\theta}$. Then to first order, writing $S_\pi(t) = \int_0^t s_\pi$,

$$\theta(T) - \bar{\theta}(T) \;=\; \underbrace{S_\pi(T)\, \eta(T)\, \Delta(\bar{\theta}(T))}_{\text{terminal}} \;-\; \underbrace{\int_0^T S_\pi(t)\, \dot{w}(t)\, \mathrm{d}t}_{\text{interior}}, \tag{1}$$

and when the path runs to a whole number of periods the terminal term vanishes and

$$\big\|\theta(T) - \bar{\theta}(T)\big\| \;\le\; 2\,\alpha(1-\alpha) \;\cdot\; \underbrace{\tau_L}_{\text{arrangement}} \;\cdot\; \underbrace{\sup_t \|w(t)\|}_{\text{schedule}}. \tag{2}$$

Proof sketches are in Appendix C. Equation (2) separates the two factors, and the separation, not either factor's value, is the content. Since $\tau_L \propto L$ a ladder's interior should be monotone in L, which a limit theorem alone does not predict; at blocked the path is one period ($\tau_L = T$), the bound buys nothing, and the displacement is set by the whole profile of w rather than by any part a schedule can delete. And a decaying schedule has a strictly shorter

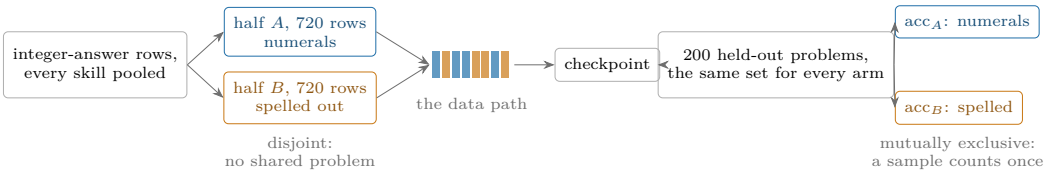


Figure 2. The instrument: one corpus that disagrees with itself. No single training row disagrees with itself; the disagreement is a property of the corpus. One held-out set is scored twice against mutually exclusive conventions, so a sample is credited to at most one — which is what makes Definition 2's sum a probability of solving and its ratio a policy.

reach than a constant one at the same peak rate: where η is large it still has the run left to contract through, and where $\Phi \approx I$ its η has decayed, whereas a constant schedule has both at $t = T$. This is the sense in which the schedule is the averaging operator. The bound does not predict the size of that gap: the direction is derived, every magnitude is measured.

Proposition 3 (The stopping phase). Treat π_L as a periodically driven system whose state oscillates about the mean-field point with amplitude increasing in L. Then (i) the blocked arm is π_L at maximal L, stopped at maximal displacement rather than a distinct mechanism; and (ii) reversing which source occupies the final block relocates the endpoint to the opposite side of the cycle, moving the two accuracies antisymmetrically while leaving C fixed to first order. We label this picture post hoc: it was formed after the block-length sweep and before the mirror test, and only the mirror test is evidence for it.

4 Experimental Setup

Corpora and arms. Two conflict constructions over competition-mathematics problems with integer answers (Hendrycks et al., 2021; Hugging Face, 2025), each split into halves of matched size with byte-identical prompts: synthetic (cf2: one half's boxed answers shifted +1, contradictory by construction, 864 rows per condition) and natural (nat: numeral against spelled-out answers, both correct, 720 rows per condition), each with a same-convention control; ratio corpora rebuild the natural conflict at 936 rows with weights 1:1 and 2:1. All are asserted by fixed-seed build scripts before any training. On cf2 the accuracies measure conformance to a stipulated convention rather than correctness against ground truth, so "neither is wrong" is exact on nat and is a statement about the scoring rule on cf2, which carries the flagship arms because its conventions are the ones the scorer separates cleanly. Per condition: only for each half, both blocked orders, the batch-mixed shuffle (shuf), pure batches in random order (purerand), and same-source blocks of $L \in \{1, 3, 6, 9, 18, 27, 54\}$ optimiser steps with trainer shuffling off so the file order is the path.

Training, observables, and registration. Full-parameter SFT on Qwen2.5-7B (Qwen Team, 2024) with AdamW, three epochs at lr $3{\times}10^{-5}$, three seeds on every ladder rung and eight on the only, shuf, blocked and keyed arms, run with ms-swift (Zhao et al., 2025) and evaluated with vLLM (Kwon et al., 2023). Evaluation is k=4 samples on 200 held-out problems, exact match on the boxed answer; the single-run noise floor is $\hat{\sigma} = 0.0234$ and the three-seed contrast floor 0.0191, and every magnitude below is quoted in contrast floors. The held-out 200 are a skill-sorted prefix of a larger pool, so every absolute level is measured on three skills of five while every between-arm contrast sees identical problems. Every number is written by a named analyzer to a verdict file, and each deciding test's read-out rule was frozen in the job script the scheduler executed before the run: ten of twelve registrations from under three hours to nearly three days ahead of the result they decided, and two that are not, each stated rather than averaged away (Appendix E).

The instrument's first failure. On its first build the positive control came out backwards — the arm that should move did not (0.10σ), the arm that should be inert did (2.41σ) — because the +1 convention had lost to the pretrained prior. Pooling integer-answer rows across skills and raising the epochs fixed it, so the 12.29σ headline is contingent on the conflict being learnable at the budget (Appendix F).

Table 1. The conservation decomposition (synthetic conflict, Qwen2.5-7B, three seeds). Capability acc_A+acc_B is flat across every arrangement; only the allocation moves. Source: the cf2b3lr3e5/ctl2b3lr3e5 result families.

arm	acc_A	acc_B	capability	share of B
A only	0.483	0.018	0.502	0.04
B only	0.064	0.423	0.487	0.87
shuf	0.310	0.218	0.528	0.41
purerand	0.303	0.224	0.527	0.43
$L = 1 \dots 54$	0.272–0.318	0.216–0.265	0.519–0.537	0.41–0.49
A→B (blocked)	0.067	0.445	0.513	0.87
coherent control (all arms)			≈ 1.01	0.500

5 Results

5.1 Allocation Moves, Capability Does Not

Table 1 is the decomposition across every arm of the synthetic conflict. The capability sum is constant to within 9.7% of its mean across all twelve arms — including only arms that never saw a row of the other convention — while the allocation share runs 0.04 to 0.87; the coherent control shows the same discipline at 1.01 with the share pinned at 0.500, and the natural conflict and both ratio corpora reproduce the pattern. The decomposition is nearly an identity, and the finding is that nobody measures it: the arrangement switch exact match reports at -0.2346, $12.29\times$ the noise floor against a same-convention control inert at 0.15σ, is a movement of the share at fixed sum: under a metric that accepts either convention it is exactly zero by construction. Any intervention evaluated by exact match on material admitting more than one correct form can post a large effect while changing nothing a user would call capability. Two agnostic rules differ here, and our zero is stated for the sum rule: crediting either convention makes the switch zero identically, crediting each problem's better convention does not.

The coin-flip fingerprint, and the same problems re-labelled. Taking for each problem the fraction of its k=4 samples landing on whichever convention it favours, committed arms (only, blocked) put **26**–**30**% of problems at 4/4, while every mixed arm (shuf, purerand, L=27) puts **7**–**9**% there and 64–68% in the middle bins — the same problem answered under different conventions across four samples — at the same unsolved mass, 25–30%. Conflict has not made the model unable; it has made the policy stochastic, which is the Bayes-optimal response to contradictory supervision (Figure 6 in Appendix G). And the movement falls on problems the model already solves: crossing from shuf to blocked over 600 instances the solved set barely changes (7.3% newly solved, 8.5% lost), while among the 386 instances solved in both arms 48.2% change which convention they favour, one-directionally — 179 flip $A \to B$ against 7 the other way, pointing at the source the blocked arm ends on. Blocking re-labels rather than teaches.

How exact the conservation is. It is an approximation: the capability sum's full range across fifteen synthetic arms is 9.97% of its mean and 15% across the natural ones, and the key of §5.4 recovers 74.9% of the gap to the additive ceiling, so "conserved" means to within an eighth, with a mechanism for the remainder (Appendix G).

5.2 The Ladder, and the Schedule Knob

Proposition 1 makes a strong claim about its own corner: below corpus scale, how the path is arranged is irrelevant. Under a cosine, four measurements agree — allocation is flat from L=1 to L=27 and across shuf and purerand (0.407 to 0.449 over a $27\times$ change of block length and of batch composition), capability spans 0.0183 inside the contrast floor, L=54 is the first rung off that floor at 0.492, and blocked sits at 0.869.

Two pre-registered mechanisms, both dead. Batch purity is not the variable: the deciding cell — pure batches in random order — lands at the shuffled arm, not the blocked one

Table 2. The same ladder under two learning-rate schedules. Ten arms, one corpus, one budget, the identical data path files, and the same three seeds (the intersection, computed rather than assumed). The families differ in lr_scheduler_type and in nothing else, checked against the training arguments and the corpus checksum, and both run to the same step. The cosine interior spans 2.20 contrast floors, below anything this design can resolve; the constant one spans **11.63**, with an intraclass correlation of 0.836 whose bootstrap interval $[0.315, 0.926]$ lies above zero. The corner does not move. Sources: constlr_ladder_table, constant_lr_verdict.

arm	cosine	constant	difference	floors
shuf (mixed within batch)	0.4129	0.3630	−0.0499	2.6
L=1	0.4491	0.4254	−0.0237	1.2
L=3	0.4126	0.4505	+0.0379	2.0
L=6	0.4363	0.4592	+0.0229	1.2
L=9	0.4140	0.4498	+0.0358	1.9
L=18	0.4257	0.4823	+0.0566	3.0
L=27	0.4071	0.5650	**+0.1579**	8.3
purerand (pure batches, random order)	0.4253	0.5851	**+0.1598**	8.4
L=54	0.4922	0.8164	**+0.3242**	17.0
blocked (one source, then the other)	0.8692	0.8501	−0.0191	1.0

(purerand 0.2242 against shuf 0.2179 and blocked 0.4454; control span +0.0037, noise), so purity carries 3% of the span. The optimiser's memory is not the variable either: AdamW's momentum is a 10-step low-pass filter (Sweeney, 2026b), the registered signature was a response turning on at $L^\star \approx 10$, and the sweep kills that and a competing basin-escape account together — the normalised response reads $+0.075, -0.007, +0.037, +0.002, +0.022, +0.002$ at $L = 1, 3, 6, 9, 18, 27$ against noise ≈ 0.06, with only L=54 rising, and the guard passed. Both deaths were derivable in advance: at small η averaging is associative, so the first prediction was impossible to leading order and the second's flat curve was forced. We keep both because the instinct they formalised — conflict destroys, and arrangement decides how much — is the field's instinct too.

Flatness is a null, so we test it as one. The interior span is below anything this design can resolve, at any evaluation budget: $\mathrm{MDE}_{80} = 3.72$ contrast floors against an interior spanning **2.20**, and 2.33 even at $k \to \infty$. What survives is one statistic, an intraclass correlation of −0.076 whose interval contains zero, and no power argument touches the corner — blocked sits 21.99 floors above the interior. The cluster count replicates where the registered criterion did not, and the fault is in the registration: the ten arms occupy two clusters on both pretraining families, while the pre-registered criterion — a span in floors — returns a miss on the second family and would have missed on the published first one too, checkable before the run and not checked, so the count is post hoc (Appendix G).

The schedule is the operator. Every arm above trains under a single cosine, so an interior flat because arrangement does nothing looks exactly like an interior flat because little happens after the midpoint. The same ladder at a constant rate separates them, and decides for the second (Table 2). The interior is not flat; it was flat to this instrument, under this schedule: it spans 11.63 floors against a detectable difference of 3.72, and its shape is what Eq. (2) predicts, allocation being monotone in τ_L up to 2 rank inversions out of 36 pairs. The averaging wall as a claim about the path is withdrawn, and what replaces it is stronger: Proposition 1 is not refuted but located, since its conclusion holds in the limit where the tail of training carries no weight and a decaying schedule is what puts a run in that limit. The cosine is not a neutral background against which the path was measured; it is the averaging operator, and the constant-rate family is the same corpus with the operator removed. The corner is schedule-free, which is why the central result does not move: blocked differs by 1.0 floor between the families against 17.0 at L=54, Proposition 2's corner case measured. One caveat: capability spans 0.0680 across the constant family against 0.0246 across the cosine one, so §5.1's conservation result is quoted from the cosine family throughout.

Two more levers, one allowed and one designed. The mixture ratio is the one lever the limit leaves free, and it behaves: the shuffled arm's share moves from 0.453 at parity to 0.312 at 2:1

against mean-field predictions 0.500 and 0.333, both inside the frozen 0.05 tolerance, while the blocked arm sits at 0.731/0.699 regardless of the ratio — a mixture optimum moves with the weights, a corner state does not. The designed schedule does not escape (pre-registered, dead): A then B is exactly a Lie–Trotter splitting (Trotter, 1959), whose canonical repair is palindromic (Strang, 1968) and second-order accurate where blocked is first-order (Blanes et al., 2009; Nguyen et al., 2026), yet in the linear model the palindrome is roughly four times worse than the blocked schedule it was built to beat, and a 16× refinement of the step moves either penalty in the fifth decimal — the penalty is not a discretisation error at all. A palindrome is still one period — it symmetrises the block without shortening τ_L — so Eq. (2) buys it nothing (Appendix G).

5.3 The Stopping Phase

Read $A^L B^L \cdots$ as a periodically driven system whose steady state is a limit cycle around the mean-field point, so an endpoint measures where the path stopped on the cycle. Two predictions ran. Amplitude: shares 0.407–0.449 for $L \leq 27$, 0.492 at L=54, and 0.869 for blocked, which on this reading is not a different species but the lowest-frequency alternation there is, half a period that never swings back. The mirror: our sweep always ends on a B-block, so rebuilding L=54 to end on A must relocate the same magnitude to the other side while moving capability not at all. It does, antisymmetrically — acc_A shifts +0.0850 while acc_B shifts −0.0875 at capability moving 0.0025 — and at L=27 both constructions sit at the floor and the mirror manufactures nothing, as it must. It repeats with the same sign and symmetry on nat, where neither convention is wrong (+0.0183 against −0.0229 at capability −0.0046), so the pattern is not an artefact of a corpus in which one convention is arguably wrong; and the mirror of blocked itself relocates the share from 0.8692 to 0.0343 across the mean-field point at capability moving 0.0020, though a comparability control fires on the small-accuracy regime, so we read only that the mirror does cross (Appendix G). Recency is not a confound of this system; it is the path's one non-average degree of freedom. And the exact-match reader should hold this against every blocked-beats-shuffled result, ours included: blocked wins the exact-match score precisely because it fails to optimise its own objective, never reaching the mixture that maximises the likelihood of its own corpus.

5.4 The Key: Training's Index Channel

Everything above concerns a learner with only the compressive channel: no signal in the prompt says which convention applies, so the state can carry only a mixture weight. The escape Remark 1 names is to make the convention queryable — rebuild the natural-conflict corpus with a convention marker in the prompt, and re-run the arms. The contribution is not the marker: prepended tags are established (Keskar et al., 2019; Korbak et al., 2023; Gao et al., 2025), and what that literature has not had is a setting where the tag disambiguates mutually contradictory supervision (Higuchi et al., 2025).

The deciding read-out had two components (Table 8 in Appendix G). Does the switch collapse? It does: the arrangement effect falls from $D = -0.0985$ (5.16 floors) unmarked to −0.0006 (0.03 floors, eight seeds) marked. The switch is convention-selection, and when selection moves from the path to the query, arrangement stops mattering — the point is not that a marker helps but that it removes the path's influence, which no schedule could. Does capability reach the union? Partially: the marked mean reaches 0.3259 against the additive ceiling 0.3727, 87.5% of the ceiling and 74.9% of the gap. So conservation is an approximation, not a law — and the deficit has a mechanism: marked B-only, trained on nothing but the spelled convention, scores 0.3767 on the numeral evaluation when the key asks for a numeral against 0.1283 unmarked, while marked A-only asked for a word scores 0.0025. The key selects among conventions the model already holds; it cannot recover one that was never written, so the 87.5/12.5 split is two walls and not one. Registered before running on Qwen3-8B-Base the collapse replicates (D from −0.0875 to −0.0125, 4.58 to 0.66 floors), and the second registered branch is the one worth the twenty trainings: that family writes four times less of the minority convention at the same corpus, budget, seeds and k, so if the residue is coverage it must leave a larger one. The threshold was fixed at two floors before the run and the residue grows from 0.0468 to 0.1025, +**2.92** floors. It was

registered as a necessary-condition test and cannot establish the coverage reading, so the difficulty-matched control stays a missing cell.

6 Threats to Validity

Scope, and what travelled. These are one model at one budget on two conflict constructions. Three results were asked to travel to a second pretraining family, each registered before its run, and the answers are not uniform: the key replicates with its residue moving as the coverage reading requires; the ladder's registered branch misses under a criterion we misspecified (§5.2); the mirror misses at a block length inside that family's flat region, with no amplitude to relocate. The guards pass well — the conflict installs better there — so none of this is an instrument at its floor. Only mutually exclusive conflicts are tested; style conflicts are open.

The schedule confound, entered as a residual and closed by measurement. blocked is two cosines where a ladder rung is one, a gap that could account for nineteen floors, so we registered a single-cosine $A^{162}B^{162}$ control before building it. It came back void: its validity check fires (0.4442 against blocked's 0.5125, 2.5 capability floors): one epoch over a tripled file is not equivalent in learning to three epochs per stage, and the registered consequence is taken in full. We then registered an arm changing the schedule while holding the path byte for byte — two cosines of 162 steps against one of 324, same rows, same order. Its guard passes and the allocation moves a great deal: at the pre-designated rung L=27 the share falls 0.4071 to 0.2912, 6.07 floors at $t = -10.1$. The two arms have identical time-averages, so the stopping phase is the path's only non-average term — but the path is not the endpoint's only input. The confound is closed in the direction it threatened: a schedule term explaining blocked would have to push allocation toward it and this pushes away on both rungs; and the block-length family moves the share 0.085 where one cosine restart moves it 0.116 — allocation is not rigid, the path is simply not the channel carrying it.

Borrowed context, and one byproduct. Three claims are borrowed. Scale: the conflict switch at 3B and 14B is measured in a concurrent submission, so this paper alone establishes nothing about how the switch scales — the largest gap in its external validity. Prevalence: the rate at which public corpora disagree about form (12.7% of shared problems, 27.1% against externally authored keys) is a survey's, and licenses the construction rather than any number here. Harness: the sampling stream is common-mode across seeds, so an external replication should expect a higher noise floor than $\hat{\sigma} = 0.0234$, widening every interval and shrinking no effect. One byproduct, flagged and not developed: purerand and blocked are the same multiset in different order and differ by 0.22 exact match, so an attacker controlling only data-loader order, touching no byte of the corpus, decides which convention a model commits to.

7 Discussion and Conclusion

One factorisation organises everything above: arrangement enters an endpoint only as a period and the schedule only as a reach (Proposition 2), so the two ends of the divided record on ordering are one knob at two settings and not two findings to reconcile. Under the field's default decaying schedule the path below corpus scale is compressed to its time-average (§5.2); that average installs the Bayes-optimal mixture, so conflict trains a stochastic policy at capability conserved (§5.1); the one term that is not an average is where the path stops, so blocked training is a stopping phase whose exact-match victory is its own objective's defeat (§5.3); and a query-time key is a second channel, not more bandwidth on the first, which is why it collapses an arrangement effect no schedule could (§5.4). What costs a reader nothing to adopt is the correction: conflicting supervision moves commitment, exact match reports it as capability, and Definition 2's two coordinates separate them from scores an evaluation already produces.

Reproducibility statement

Every number in this paper is written by a named analyzer to a verdict file, and every deciding test's read-out rule was frozen in the job script the scheduler executed before the run. Appendix E is the registration index: one row per registered claim, giving the commit that froze each threshold, the timestamp of the result it decided, and the interval between them, including the two rows where that interval is not favourable to us. Appendix D is the full ledger of registered claims and outcomes, including the three falsified mechanisms, the inverted instrument build and the void control. Appendix C gives the proof sketches for Propositions 1 and 2 and states the assumptions each needs. Appendix F gives the corpus constructions, row counts, the held-out set's skill composition and the three instrument builds. §4 states the model, optimiser, budget, seed counts and evaluation protocol; the corpora are built by fixed-seed scripts from public benchmarks (Hendrycks et al., 2021; Hugging Face, 2025), and training and evaluation use published open-source stacks (Zhao et al., 2025; Kwon et al., 2023). Corpus build scripts, the arranged path files, the analyzers that write each verdict, and the verdict files themselves are provided as supplementary material, together with the script that rebuilds Appendix E from the commit history and the result base. Checkpoints for most published arms were reclaimed after evaluation, which is stated wherever it limits a re-analysis.

Ethics statement

This work trains models on deliberately conflicting supervision and reports one finding with a direct misuse reading, which we state rather than bury: an actor controlling only data-loader order, touching no byte of a corpus, moves which of two correct conventions a model commits to by 0.22 in exact match (§6). We report it as a defensive result — content-level auditing of a training corpus cannot see an ordering — and develop no attack; the arms that demonstrate it are the same arms the paper's positive results are measured on. No human subjects were involved, and no personal data. The corpora are derived from public competition-mathematics benchmarks under their own licences, and the conflicts are synthesised by published scripts rather than harvested. The evaluation finding — that exact match on material admitting more than one correct form silently reports commitment as capability — is a correction to measurement practice and carries no group-differential harm we are aware of. We have read the ICLR Code of Ethics and adhere to it.

AI use statement

We report our use of generative AI tools against the categories of the ICLR 2027 AI Policy for Authors. Our use was confined to preparing the document.

Tasks in the required-disclosure list: none. Generative AI was not used to generate synthetic data sets, to develop the theoretical model or conceptual framework, to formulate the mathematical claims, to supply critical ingredients for the proofs or to write them, to propose or refine the hypotheses, to design the methodology or the experiments or to give feedback on them, to set or choose any pre-registered threshold, to implement methods, to clean or reformat any dataset, or to interpret the results. The corpora are emitted by fixed-seed builders, the training and evaluation runs of §4 were configured and launched by the authors, and every number reported here is computed by an analyzer the authors wrote, from the result files those runs produced; builders, analyzers, job scripts, result files and verdicts are all in the supplementary material. Translation and qualitative or thematic data analysis are not applicable to this work.

Tasks in the recommended-disclosure list. Generative AI was used on the manuscript itself. It condensed a longer version of this work into the present conference format, which involved drafting and rewriting substantial parts of the exposition; it edited for readability, formatted tables and references, and assembled and anonymised the supplementary archive. It was not used to discover the research topic, identify the gap, select the literature, or decide what the paper claims: the claims, the numbers they quote and the qualifications attached to them are the authors', carried over from the longer version.

Verification. Every quantitative statement in this version was checked against the verdict file it cites, and every registered outcome against the read-out rule frozen in the job script the scheduler executed, which Appendix E tabulates against its commit. The one transformation AI-written code applied to a released artifact — rewriting absolute paths and host names in the supplementary archive for anonymisation — was scoped to provenance strings and verified not to have altered any evaluation set, score vector, or reported number. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims and artifacts.

References

Niccolò Ajroldi, Antonio Orvieto, and Jonas Geiping. When, where and why to average weights? arXiv preprint, 2025. arXiv:2502.06761.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences, 116(32):15849–15854, 2019.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In International Conference on Machine Learning (ICML), 2009.

Dimitri P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. In Optimization for Machine Learning, pp. 85–119. MIT Press, 2011.

S. Blanes, F. Casas, J. A. Oteo, and J. Ros. The Magnus expansion and some of its applications. Physics Reports, 470(5–6):151–238, 2009.

Nikolay N. Bogoliubov and Yuri A. Mitropolsky. Asymptotic Methods in the Theory of Non-Linear Oscillations. Gordon and Breach, 1961.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

Henry C. Conklin, Tom Hosking, Tan Yi-Chern, Julian Gold, Jonathan D. Cohen, Thomas L. Griffiths, Max Bartolo, and Seraphina Goldfarb-Tarrant. Learning is forgetting: LLM training as lossy compression. arXiv preprint, 2026. arXiv:2604.07569.

Yalun Dai, Yangyu Huang, Tongshen Yang, Yonghan Wang, Xin Zhang, Wenshan Wu, Qihao Zhao, Hao Li, Yuanyuan Gao, Kim-Hui Yap, and Scarlett Li. Demystifying data organization for enhanced LLM training. In Proceedings of ACL, 2026. arXiv:2605.30334.

Natalia Díaz-Rodríguez, Vincenzo Lomonaco, David Filliat, and Davide Maltoni. Don't forget, there is more than forgetting: new metrics for continual learning. arXiv preprint, 2018. arXiv:1810.13166.

Mohamed Elgaar and Hadi Amiri. Curriculum learning for LLM pretraining: An analysis of learning dynamics. arXiv preprint, 2026. arXiv:2601.21698.

Konstantinos Emmanouilidis, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Rene Vidal. Shuffling the data, stretching the step-size: Sharper bias in constant step-size SGD. arXiv preprint, 2026. arXiv:2604.10373.

Itay Evron, Edward Moroshko, Rachel Ward, Nathan Srebro, and Daniel Soudry. How catastrophic can catastrophic forgetting be in linear regression? In Conference on Learning Theory (COLT), 2022.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. DoGE: Domain reweighting with generalization estimation. arXiv preprint, 2024. arXiv:2310.15393.

Robert M. French. Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences, 3(4):128–135, 1999.

Tianyu Gao, Alexander Wettig, Luxi He, Yihe Dong, Sadhika Malladi, and Danqi Chen. Metadata conditioning accelerates language model pre-training. arXiv preprint arXiv:2501.01956, 2025.

Xinran Gu, Kaifeng Lyu, Jiazheng Li, and Jingzhao Zhang. Data mixing can induce phase transitions in knowledge acquisition. arXiv preprint arXiv:2505.18091, 2025.

Mert Gürbüzbalaban, Asuman Ozdaglar, and Pablo A. Parrilo. Why random reshuffling beats stochastic gradient descent. Mathematical Programming, 186:49–84, 2021.

Gustav Olaf Yunus Laitinen-Fredriksson Lundstrom-Imanov. Mechanistic analysis of catastrophic forgetting in large language models during continual fine-tuning. arXiv preprint, 2026. arXiv:2601.18699.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations. Springer, 2nd edition, 2006.

Jeff Z. HaoChen and Suvrit Sra. Random shuffling beats SGD after finite epochs. In International Conference on Machine Learning (ICML), 2019.

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

Rei Higuchi, Ryotaro Kawata, Naoki Nishikawa, Kazusato Oko, Shoichiro Yamaguchi, Sosuke Kobayashi, Seiya Tokui, Kohei Hayashi, Daisuke Okanohara, and Taiji Suzuki. When does metadata conditioning (NOT) work for language model pre-training? a study with context-free grammars. arXiv preprint arXiv:2504.17562, 2025.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In Proceedings of EMNLP, 2021.

Hugging Face. OpenR1-Math-220k, 2025. https://huggingface.co/datasets/open-r1/OpenR1-Math-220k.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In Advances in Neural Information Processing Systems (NeurIPS), 2018.

João Maria Janeiro, Mathurin Videau, Andrea Caciolai, Benjamin Piwowarski, Patrick Gallinari, and Loic Barrault. Are we evaluating knowledge or phrasing? mitigating MCQA sensitivity with ParaEval. arXiv preprint arXiv:2606.10657, 2026.

Yaning Jia, Chunhui Zhang, Xingjian Diao, Xiangchi Yuan, Zhongyu Ouyang, Chiyu Ma, and Soroush Vosoughi. What makes a good curriculum? disentangling the effects of data ordering on LLM mathematical reasoning. arXiv preprint, 2025. arXiv:2510.19099.

Yiming Ju, Ziyi Ni, Xingrun Xing, Zhixiong Zeng, Hanyu Zhao, Siqi Fan, and Zheng Zhang. Mitigating training imbalance in LLM fine-tuning via selective parameter merging. arXiv preprint arXiv:2410.03743, 2024.

Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. Quagmires in SFT-RL post-training: When high SFT scores mislead and what to use instead. arXiv preprint, 2025. arXiv:2510.01624.

Pyotr L. Kapitza. Dynamic stability of a pendulum with an oscillating point of suspension. Journal of Experimental and Theoretical Physics, 21:588–597, 1951.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. CTRL: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858, 2019.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 2017.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining language models with human preferences. In Proceedings of the 40th International Conference on Machine Learning (ICML), 2023.

Nate Kornell and Robert A. Bjork. Learning concepts and categories: Is spacing the "enemy of induction"? Psychological Science, 19(6):585–592, 2008.

Dmitrii Krasheninnikov, Richard E. Turner, and David Krueger. Fresh in memory: Training-order recency is linearly encoded in language model activations. arXiv preprint, 2025. arXiv:2509.14223.

Konstantin Krestnikov. Truth as a compression artifact in language model training. arXiv preprint arXiv:2603.11749, 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, et al. Efficient memory management for large language model serving with PagedAttention. In Symposium on Operating Systems Principles (SOSP), 2023.

Jordan LeDoux. The order is the message. arXiv preprint arXiv:2603.25047, 2026.

Jin Hwa Lee, Matthew Smith, Maxwell Adam, and Jesse Hoogland. Influence dynamics and stagewise data attribution. arXiv preprint, 2025. arXiv:2510.12071.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In Advances in Neural Information Processing Systems (NeurIPS), 2021.

Emmy Liu, Kaiser Sun, Millicent Li, Isabelle Lee, Lindia Tjuatja, Jen-tse Huang, and Graham Neubig. What do language models learn and when? the implicit curriculum hypothesis. arXiv preprint, 2026. arXiv:2604.08510.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In Advances in Neural Information Processing Systems (NeurIPS), 2017. arXiv:1706.08840.

Yucheng Lu, Wentao Guo, and Christopher De Sa. GraB: Finding provably better data permutations than random reshuffling. In Advances in Neural Information Processing Systems (NeurIPS), 2022. arXiv:2205.10733.

Kairong Luo, Zhenbo Sun, Haodong Wen, Xinyu Shi, Jiarui Cui, Chenyi Dang, Kaifeng Lyu, and Wenguang Chen. How learning rate decay wastes your best data in curriculum-based LLM pretraining. arXiv preprint, 2025. arXiv:2511.18903.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation, 24:109–165, 1989.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In Advances in Neural Information Processing Systems (NeurIPS), 2020. arXiv:2006.05988.

Kaixiang Mo, Yuxin Shi, Weiwei Weng, Zhiqiang Zhou, Shuman Liu, Haibo Zhang, and Anxiang Zeng. Mid-training of large language models: A survey. arXiv preprint, 2025. arXiv:2510.06826.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In International Conference on Learning Representations (ICLR), 2020.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In International Conference on Machine Learning (ICML), 2022.

Angelia Nedić and Dimitri P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. SIAM Journal on Optimization, 12(1):109–138, 2001.

Lam M. Nguyen, Dzung T. Phan, and Jayant Kalagnanam. Learning to shuffle: Block reshuffling and reversal schemes for stochastic optimization. arXiv preprint, 2026. arXiv:2604.00260.

Irina Piontkovskaia and Sergey Nikolenko. First-order predictable but pairwise fragile: Local task adaptation in trained transformers. arXiv preprint, 2026. arXiv:2607.16821.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022. arXiv:2211.02570.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. arXiv preprint arXiv:2201.02177, 2022.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

Qwen Team. Qwen2.5 technical report. arXiv preprint, 2024. arXiv:2412.15115.

Jun Rao, Xuebo Liu, Lian Lian, Shengjun Cheng, Yunjie Liao, and Min Zhang. CommonIT: Commonality-aware instruction tuning for large language models via data partitions. In Proceedings of EMNLP, 2024. arXiv:2410.03077.

Doug Rohrer and Kelli Taylor. The shuffling of mathematics problems improves learning. Instructional Science, 35(6):481–498, 2007.

Hanne Say, Suzan Ece Ada, Emre Ugur, Minoru Asada, and Erhan Oztop. Interleaved multitask learning with energy modulated learning progress. arXiv preprint, 2025. arXiv:2504.00707.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In Advances in Neural Information Processing Systems (NeurIPS), 2023.

Gilbert Strang. On the construction and comparison of difference schemes. SIAM Journal on Numerical Analysis, 5(3):506–517, 1968.

John Sweeney. The geometry of sequential learning: Lie-bracket prediction of transfer order. In International Conference on Machine Learning (ICML), 2026a. arXiv:2606.24993.

John Sweeney. Optimizer memory makes shuffle order a first-order source of fine-tuning noise. arXiv preprint, 2026b. arXiv:2606.29554.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. IEEE Information Theory Workshop (ITW), 2015.

H. F. Trotter. On the product of semi-groups of operators. Proceedings of the American Mathematical Society, 10(4):545–551, 1959.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. arXiv preprint, 2024. arXiv:2402.01364.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In Advances in Neural Information Processing Systems (NeurIPS), 2023. arXiv:2305.10429.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024. arXiv:2403.08319.

et al. Xue. Why supervised fine-tuning fails to learn: A systematic study of incomplete learning in large language models. arXiv preprint, 2026. arXiv:2604.10079.

Jewon Yeom, Jaewon Sok, Heejun Kim, Seonghyeon Park, Jeongjae Park, and Taesup Kim. Hallucination as commitment failure: Larger LLMs misfire despite knowing the answer. arXiv preprint, 2026. arXiv:2605.22007.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

Yang Zhang, Amr Mohamed, Hadi Abdine, Guokan Shang, and Michalis Vazirgiannis. Beyond random sampling: Efficient language model pretraining via curriculum learning. arXiv preprint, 2025. arXiv:2506.11300.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, and Others. Swift: A scalable lightweight infrastructure for fine-tuning. In Proceedings of the AAAI Conference on Artificial Intelligence, 2025.

Supplementary Material

This appendix follows the references and is supplementary. The main text is self-contained: every claim it makes is stated with the measurement that decides it, and the sections below hold proofs, full tables, the registration record and the cells we did not buy. Reviewers are not required to read it.

A Scope of the Claims

Six statements a reader could reasonably extract from the main text are not supported by what is measured here, and separating them from the ones that are is worth stating in full. Not "order does not matter under decay": the cosine interior spans 2.20 floors against a minimum detectable difference this design never reaches, so the null is carried by a pooled intraclass correlation (-0.076, $[-0.409, +0.166]$) and the two-state count, both weaker than "no effect". Not a general law of LLM training: these are one budget on Qwen2.5-7B, and Qwen3-8B-Base replicates the switch and the key but not the ladder's registered criterion or the mirror. Not independence of the two coordinates: we claim conservation of C, while $\mathrm{Cov}(p, s) \neq 0$ is measured and fails at 5.17 floors. Not that conflict is harmless: an exact-match benchmark reporting one convention sees a real loss, and what is conserved is the union. Not a quantitative theory: the direction is derived and every magnitude is measured. Not a statement about every coordinate: everything above is measured along a one-parameter family of paths and on one behavioural coordinate, so another coordinate could carry order information we did not look at, and the checkpoints that would settle it were reclaimed after evaluation. Two controls searched off that family and found nothing outside the same interval — batch purity moves 3% of the span and the mirror relocates within it — and adding a coordinate can only raise the number of occupied states, so two is a floor.

B Extended Related Work

A conflict is an error — the premise almost everyone shares. The knowledge-conflict literature names our object exactly (Xu et al., 2024 call it intra-memory conflict) and every remedy it surveys is a repair: refine, regulate, reweight or filter the offending rows (Xue, 2026). A repair presupposes that one of the conflicting forms is wrong; our natural corpus is built so that neither is, and under that construction Proposition 1 makes the mixture the loss-minimising policy rather than a malfunction. Krestnikov (2026) come nearest to dropping the premise, finding that a coherent alternative rule system destroys the preference for the true answer while a second competing rule restores most of it — our construction reached from the truth-tracking side. Their corpora make one form wrong and ours make neither, so their restored accuracy and our conserved C are different quantities that happen to move together. The same premise, inverted, organises the evaluation side (Plank, 2022): perspectivist repairs fix evaluation by matching the distribution of human labels, which also needs a target unavailable here, since with two correct conventions every allocation is equally correct. An undecomposed accuracy does not merely undercount capability in the familiar way that exact match penalises three against 3 (Schaeffer et al., 2023; Kang et al., 2025); it cannot express the coordinate along which our arms move. Schaeffer et al. (2023) is the closest precedent and the closest warning — emergence turned out to be a property of the metric rather than of the model — and §5.1 performs the same move on a different axis when it reports its own 12.29σ arrangement switch as exactly zero under a convention-agnostic score. Holtzman et al. (2021) names the mechanism by which several surface forms of one correct answer compete for probability mass, so a scorer reading only the highest-probability string reports the competition rather than the knowledge; that competition is our S, and Definition 2 adds only that it can be divided out of C. Yeom et al. (2026) measure the same split at inference time, finding 16–47% of instruct-model hallucinations occur with substantial mass already on the correct concept, the distinguishing factor being whether that mass concentrates on one surface form or disperses. Janeiro et al. (2026) price what it costs an evaluation: on a 1–8B testbed, models trained on identical knowledge post false gaps above two points from answer phrasing alone. Their remedy and ours point in opposite directions on purpose — ParaEval averages the surface-form term away, which is right when the phrasing is nuisance, whereas we keep it as a coordinate, because in a conflicted corpus the phrasing is exactly what arrangement moves.

Where Proposition 1 comes from, and what sits inside the wall. In the deterministic cyclic case the proposition is the central dichotomy of the incremental-gradient literature (Bertsekas, 2011; Nedić & Bertsekas, 2001): with a step size decaying to zero the iterates converge to a minimiser of the summed objective and the order does not survive, while at a constant step they enter a limit cycle whose position depends on the order. Proposition 2 interpolates between those regimes and Proposition 3 is that limit cycle. The without-replacement line prices the orderings against each other (Gürbüzbalaban et al., 2021; HaoChen & Sra, 2019; Mishchenko et al., 2020; Lu et al., 2022), and the constant-step-size bias our constant-rate family reads is under current study in its own right (Emmanouilidis et al., 2026). What we add is not the theorem but its transport: that literature states its results for a fixed objective, and the question here is what the same dichotomy does to a policy when the summed objective's minimiser is a mixture rather than a point. The transport supplies an allocation coordinate that moves while the loss does not, and the observation that the field's default schedule puts nearly all of published fine-tuning practice at one end of the dichotomy without saying so. The contradiction that seeded our own first mechanism — purify the batch (Rao et al., 2024), mix it (Dai et al., 2026), mix it with a theorem (Lu et al., 2022) — is then resolved sideways rather than adjudicated: at conflict, purity carries 3% of the span, and the axis those three papers dispute lies strictly inside the wall, which is why it can persist without either side being wrong about its own measurements. Sweeney (2026b) shows that optimiser state makes shuffle order a first-order noise source, while our block sweep shows the momentum window moves the conflict signal not at all; the two are consistent under Proposition 1, buffers adding variance about a mean-field point the time-average sets. Sweeney (2026a) proposes the sharpest positive claim we could find, that the Lie bracket of two tasks' update operators predicts which order transfers better, and Appendix G.7 measures a one-shot commutator score built in its spirit and finds it inverted

at Spearman -0.543 and -0.600. That is a range boundary and not a refutation, because the two experiments do not meet: their tournament reports 98% pairwise accuracy at block length k=1 falling to 73% at k=20, whereas our budgets are 162–324 steps per block, and Proposition 2 says why a score computed once at θ_0 must decay with τ_L. Per-step gradient-conflict methods (Yu et al., 2020; Liu et al., 2021; Navon et al., 2022) and ordered-shuffle schemes (Lu et al., 2022) likewise operate inside the wall: they change optimisation and can change variance, but Remark 1 says they cannot change the installed allocation. From the other side, methods that reweight proportions (Xie et al., 2023; Fan et al., 2024; Gu et al., 2025; Mo et al., 2025) act on exactly the quantity Proposition 1 leaves free, and our ratio experiment is the controlled version of their premise. Catastrophic forgetting (McCloskey & Cohen, 1989; French, 1999; Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Wu et al., 2024; Díaz-Rodríguez et al., 2018) concerns capability lost when a second task overwrites a first; under Definition 2 that is a movement of C, whereas the conflict switch is a movement of S at C held to within 9.7%. Evron et al. (2022) treat blocked linear regression as alternating projections, and the stopping phase is the nonlinear-policy face of the same recency; Krasheninnikov et al. (2025) fine-tune sequentially on six datasets and find training-order recency linearly encoded in the activations, with a linear probe separating early- from late-learned entities at about 90% — their read-out on the representation and ours on the policy, the same statement at different depths. Conklin et al. (2026) and Gustav Olaf Yunus Laitinen-Fredriksson Lundstrom-Imanov (2026) characterise forgetting mechanistically, and neither predicts a term that switches on when two sources disagree while capability holds; Xue (2026) isolates internal SFT-data inconsistency per sample, and our decomposition says what that inconsistency does — it moves the model from a deterministic to a stochastic policy at fixed capability.

Ordering, its nulls, and the sharpest counter-result. The record on ordering is divided (Bengio et al., 2009; Rohrer & Taylor, 2007; Kornell & Bjork, 2008; Say et al., 2025; Jia et al., 2025; Zhang et al., 2025; Elgaar & Amiri, 2026; Liu et al., 2026; Luo et al., 2025), and the division is what the averaging wall predicts: these are paths below corpus scale, where Proposition 1 says the endpoint is the same and only transients and stopping phases differ. Luo et al. (2025) find curriculum advantages over random shuffling that hold at a constant rate and diminish under standard decay, in pretraining at 1.5B parameters over 30B tokens. We reproduce that moderator under control, at 7B in supervised fine-tuning, over ten arrangements of one conflicted corpus and two families differing only in lr_scheduler_type. Their reading is that decay wastes the curriculum; ours is that decay is the averaging operator of Proposition 1, so "order matters" and "order does not matter" are the two ends of one schedule knob rather than two findings to be reconciled. The divided record should then sort by how much step size survives to the end of training, and Eq. (2) makes that a consequence rather than a caution, since the two factors multiply. One case can be checked without new runs: Elgaar & Amiri (2026) hold Pythia's configuration fixed across orderings, decaying a cosine to a tenth of peak at every size, and report ordering effects on stability largely gone by 410M, which is the sign Proposition 1 requires; Piontkovskaia & Nikolenko (2026) report pairwise order predictions degrading with block length, which the same proposition explains. Against all of this, LeDoux (2026) report the opposite, and report it cleanly: training small networks on modular arithmetic from scratch, two fixed orderings reach 99.5% test accuracy from a training set covering 0.3% of the input space where random ordering does not, and the learned Fourier representation's fundamental frequency is the mathematical dual of the ordering's own structure — wide enough that the paper names it a covert information channel able to bypass content-level auditing. We measure ten orderings of one corpus into two distinguishable endpoint states, on a range its own resolution would divide into about ten. Both results are right, and what the path can write depends on how much the prior has already fixed: from scratch the parameters carry no structure the order must compete with, whereas on a pretrained 7B prior the same channel writes into parameters that already encode the answer format, the arithmetic and the convention preference, and it moves the one coordinate pretraining left underdetermined. Order is a wide channel into an empty state and a narrow one into a full state, a resource trade neither paper states alone, and it predicts that the channel narrows monotonically with pretraining scale. The two safety readings converge from opposite regimes: LeDoux (2026) name a covert channel that content auditing misses, and §6 reports the pretrained-model version, where an at-

tacker controlling only data-loader order moves the allocation 0.22 in exact match. At LLM fine-tuning scale, Ju et al. (2024) is the closest published claim to a positive result in our own setting: data order produces training imbalance in LLM SFT, with the proposed fix being to merge models fine-tuned under different orderings. Read against Proposition 1 the remedy is the informative half, since averaging endpoints across orders explicitly constructs the quantity the infinitely divided limit installs implicitly. Their measurements are on heterogeneous instruction data where our coherence premise does not hold, so we report this as consistent rather than as replication. Finally, the lazy and kernel regimes (Jacot et al., 2018; Chizat et al., 2019) give the setting in which the averaging argument is exact, and full-parameter SFT is measurably not fully lazy, which is why Proposition 1 is reported as governing signs and orderings rather than magnitudes; averaging itself is classical (Bogoliubov & Mitropolsky, 1961; Kapitza, 1951), and its modern empirical form is Ajroldi et al. (2025), who benchmark weight averaging across seven workloads, ask whether it can replace learning-rate decay, and conclude that the two are best combined rather than exchanged — which is what leaves lr_scheduler_type free to be varied on its own. Phenomena living in the transient rather than the endpoint, grokking (Power et al., 2022) and double descent (Belkin et al., 2019; Nakkiran et al., 2020), are outside our budget regime but share the moral that an endpoint measurement can be a statement about where the path stopped. Attribution methods (Pruthi et al., 2020; Lee et al., 2025) track which examples moved the parameters; Definition 3's write-channel rate is the complementary statistic, asking how much of the path's order survives at all. Reading training as a channel with a rate has an ancestor in the information bottleneck (Tishby & Zaslavsky, 2015), and the difference is not one of degree: the bottleneck compresses the input while preserving the label, whereas the quantity compressed here is the order of a fixed multiset and the receiver is the endpoint's behaviour.

C Proof Sketches

Proposition 1 (the infinite-division limit). The first statement is Bogoliubov–Krylov averaging (Bogoliubov & Mitropolsky, 1961) applied to a piecewise-constant vector field whose period $2L\eta$ tends to zero: the trajectory tracks the period-average of the field, which is the proportion-weighted gradient. The second is the first-order condition for $\min_q \alpha\,\mathrm{CE}(p_A, q) + (1-\alpha)\,\mathrm{CE}(p_B, q)$ with $\alpha = n_A/(n_A+n_B)$, whose solution is $\alpha p_A + (1-\alpha)p_B$. That second statement is about the unconstrained minimiser over conditional distributions; identifying it with the flow's stationary point in parameter space additionally requires the family to be rich enough to represent $q^\star$ at the contested inputs, which Remark 1 argues is comfortably met at 7B and which we state here rather than leave to prose.

Proposition 2 (reach and period). Subtract the two flows and linearise: $u = \theta - \bar\theta$ obeys $\dot u = -\eta(\bar H - s_\pi \nabla\Delta)u + \eta s_\pi \Delta + O(\|u\|^2)$ with $\bar H = \nabla^2 \bar L(\bar\theta)$. The term in $\nabla\Delta$ is the instantaneous Hessian's departure from the averaged one; it is not smaller than the term we keep, so it has to be disposed of rather than dropped. At leading order u oscillates as $S_\pi\,\eta\Delta$, and $\langle s_\pi S_\pi\rangle = \tau_L^{-1}\int \frac{\mathrm{d}}{\mathrm{d}t}(S_\pi^2/2) = 0$ over a whole period because S_π is periodic, so it first contributes at $O(\tau_L^2)$. Variation of constants on what remains gives $u(T) = \int_0^T s_\pi w$; integrating by parts with $S_\pi(0) = 0$ gives Eq. (1). For the bound, S_π is the triangle wave of s_π, so $\sup_t |S_\pi| = \alpha(1-\alpha)\tau_L$, and w rises and falls at most once under a monotone or single-peaked schedule, so $\int \|\dot w\| \le 2\sup_t \|w\|$.

What the bound is used for, and what it is not. The reach depends on Φ as well as on η, the two schedule families do not share a Φ, and a bound is not an estimate: a different functional of the same profile — its total variation — orders the two schedules the other way when the run contracts little. So the direction is derived and the magnitude is measured. The proposition is a first-order statement about a linearisation, and we use it for signs, orderings and which factor a knob enters, which is the standing it has in the averaging literature it comes from (Bogoliubov & Mitropolsky, 1961; Hairer et al., 2006). One prediction it makes that this paper has not tested: the terminal term in Eq. (1) is the stopping phase of Proposition 3, multiplied by $\eta(T)$ with no contraction discount, so a decaying schedule

should suppress the mirror in the same way it suppresses the ladder — and our mirror arms are all cosine (§G).

D The Ledger of Registered Claims

The main text says three mechanisms were pre-registered and falsified. Table 3 is the whole ledger rather than those three, because a scorecard listing only the informative failures is a selection, and because two further entries cost us something a reader should not have to reconstruct: the positive control inverted on its first build, and the control registered to discharge the learning-rate confound came back void.

Table 3. The ledger: every claim this paper registered, and what came back. Read the verdict column as a distribution rather than a score: three mechanisms died, four predictions held, one control was void and one instrument build inverted. The three deaths were derivable in advance from Proposition 1, which is why they are kept rather than deleted. Rules are quoted from the artefact that froze them; Appendix E gives the commit and the interval for each.

registered claim	the rule, frozen first	what came back	verdict
v1 batch composition carries the switch	purerand near blocked ⇒ purity; near shuf ⇒ order	purerand 0.2242 against shuf 0.2179, blocked 0.4454: purity is 3% of the span	dead
v2 optimiser memory cancels contested directions	a response turning on at $L^\star \approx 10$, or at $\tau_{\text{write}} = 13.6$	$M(L)$ flat through both onsets, $\sigma(M) \approx 0.06$; only L=54 rises	dead
Strang a palindromic schedule beats a blocked one	S1 median $\lvert D_{\text{str}}\rvert / \lvert D_{\text{blk}}\rvert < 0.5$; S2 slopes differ by ≥ 0.5	S1 3.65; S2 both slopes 0; S3 passes	dead
T1 allocation tracks the mixture weights	share within 0.05 of mean-field 0.500 and 0.333	0.453 and 0.312: misses 0.047 and 0.021, both toward the pretrained prior	held
Ladder every block length lands at the interleaved allocation	L=1 … 54 within noise of shuf's own share	0.407–0.449 against shuf's 0.413	held
M1 the mirror relocates antisymmetrically at fixed C	ending on A moves the same magnitude to the other side	+0.0850 against −0.0875, capability moves 0.0025	held
C the key collapses the switch	$\lvert D_{\text{keyed}}\rvert$ within one floor of zero	$-0.0985 \to -0.0006$ (5.16 → 0.03 floors)	held
L the key lifts capability to the union	marked mean $\geq 0.90\times$ the additive ceiling 0.3727	0.3259: 87.5% of the ceiling, short of the bar by 0.50 floors	partial
S-C a single-cosine arm discharges the schedule confound	validity check S-C4 read before any allocation comparison	total accuracy 0.4442 against 0.5125, a gap of 2.5 capability floors	void
Control conflict fires and its twin stays inert	control within noise, conflict beyond it	v1 inverted (0.10σ against 2.41σ); v3 reads 0.15σ against 12.29σ	v1 failed
Q3-key the key collapses on a second pretraining family	branches C, L and V frozen before the first result file	D $-0.0875 \to -0.0125$; unlock 75.5% against 87.5%; residue grows 2.92 floors	K1/L4/V1
Q3-wall the averaging wall replicates	span of the flat region ≤ 2 floors, gate read first	gate passes at 15.0 floors; span 2.85 floors, and the published row is 2.18, so the bar was mis-specified	W2

registered claim	the rule, frozen first	what came back	verdict
Q3-mirror the stopping phase replicates	antisymmetric shift at L=54, capability within a floor	+0.0204 against +0.0012; L=54 is inside the flat region on this family	M2
high-k the borrowed evaluation share transfers, so σ falls as $\sqrt{0.136 + 0.864 \cdot 4/k}$	ratio(16) $\in [0.475, 0.713]$ around a predicted 0.594; all eleven arms required at each of k=16 and k=32	k=16 complete at 33/33; ratio(16) = 1.079, outside the band and above one, which the additive model can carry only as a negative component. §5.2's shared engine stream predicts one	HK-4

The high-k row's bookkeeping, since its verdict turns on it rather than on a measurement: the job ran nine arms at k=32 where the guard asks for eleven, so the cell is void on bookkeeping and not on measurement; the six missing evaluations were submitted and then withdrawn unstarted when the project's compute closed. They would only have moved the label, because §5.2's correction comes from a direct measurement that decides no branch.

E The Registration Index

Registering a read-out inside the job script the scheduler executes is stronger than a note in one way and weaker in another. Stronger, because the file carrying the threshold is the file that ran. Weaker, because a job script is committed when it is written, which is usually but not always before the run. Table 4 carries the gap, measured, rather than the assurance, and Figure 3 puts every interval on one axis. Registered is the first commit in which the artefact carries the threshold, found with git log -S on a distinctive fragment of the rule, not the first commit of the file; result is the modification time of the verdict file on the result base. Both are machine-checkable, and a script released with the supplementary material rebuilds the whole table from the history and the result base without reading it.

Two rows are evidenced differently and are marked so. The CM and NI registrations were frozen on the shared volume the scheduler reads from, and timestamped there, before the jobs that used them were submitted; they reached a commit only afterwards. A file timestamp on a volume we can write to is weaker evidence than a commit, and the rows say which they have rather than borrowing the strength of the rows above them.

One correction this index made to our own record. An earlier version named the wrong registering artefact for the ladder, and the correction was applied to half the row: the artefact column was changed and the result column was not, so the row printed an interval of +23h 41m where the true one is +10h 30m. Both are positive and neither changes a verdict, but the number was wrong and it was wrong in the direction that flattered us. It was found by the rebuild script, not by us, which is why the result column now names its verdict file rather than a bare timestamp. An index nobody audits is a claim rather than a check.

Table 4. Seventeen registrations, what each returned, and the interval each one bought. Seven confirmed at their primary branch, two decided at another, two came back partial, three void on a guard, and three returned no branch at all. Fifteen were frozen from about an hour to nearly three days ahead of the result they decided. Two are not, and both are stated rather than averaged away: the palindromic check is a linear-model computation of a few seconds whose thresholds and whose answer entered the repository in the same commit, so it can assert an ordering it cannot demonstrate; and the amplitude-onset registration was withdrawn before any training because the rungs it named cannot be built at this budget, which is a fact about the ladder rather than about either family. The second-family replications carry positive intervals and are nonetheless printed here together with their answers rather than announced in an earlier version, which is the weaker of the two guarantees a registration can offer. The last two rows' intervals are measured against a file timestamp on the shared volume rather than against a commit, which is weaker again and is why those cells say no commit instead of a hash. Of the two only NI is still recomputable from this tree: CM's registration gained an outcome section after its verdict, and writing that section overwrote the timestamp its interval was measured from. NI's was left byte-frozen for that reason and its outcome is recorded here and in its verdict file instead.

claim	outcome	artefact carrying the threshold	registered	result (verdict file, mtime)	gap
instrument, the positive control	confirmed	slurm_ natconflict. sbatch	65acb7a 08-01 12:52	natconflict 08-03 08:23	+1d 19h
v1 batch composition	confirmed	slurm_ batchcomp. sbatch, decision rule	2e5f5a9 08-02 12:42	batchcomp 08-02 15:32	+2h 50m
v2 and the ladder	no branch fired	slurm_pathspec. sbatch and PLAN.md: predicts $L^\star \approx 10$	cd53c09 08-02 16:02	pathspec 08-03 02:33	+10h 30m
M1 the mirror	M1 confirmed	slurm_dpd_ mirror.sbatch: an asymmetry falsifies the limit-cycle picture	a85c247 08-03 15:32	dpd_mirror 08-06 07:44	+2d 16h
C, L the key	C1, L4 confirmed	slurm_disambig. sbatch, C on the switch and L on the level	4830e37 08-04 06:46	disambig 08-04 10:49	+4h 03m
T1 the mixture ratio	T1 confirmed	slurm_ratio2. sbatch, tolerance 0.05 frozen with the prediction	ff94a95 08-05 06:22	ratio2 08-05 16:20	+9h 58m
S-C the single-cosine control	S-C4 void	prereg/single_ cosine_control. md, four branches, S-C4 first	39e106d 08-08 08:41	singlecos 08-08 16:53	+8h 12m
S1–S3 the palindromic schedule	asserted, not demon-strated	verify_splitting. py docstring	e1eadef 08-02 11:07	same commit	none

claim	outcome	artefact carrying the threshold	registered	result (verdict file, mtime)	gap
Q3-onset where the amplitude turns on	withdrawn	prereg/q3_amplitude_onset.md, withdrawn before any training: the rungs it names lie across the epoch boundary, and the only route there is the tripled file the S-C control measured as not learning-matched (§5.3)	79b05eb 08-11 01:56	none possible	—
Q3-key the key on a second family	K1, L4, V1 confirmed	prereg/q3_key_replication.md with readout_q3unlock.py	66eab04 08-09 14:25	q3unlock 08-11 01:37	+1d 11h
Q3-wall the ladder on a second family	W2, M2 partial	prereg/q3_ladder_replication.md with readout_q3ladder.py	a679a1b 08-09 15:41	q3ladder 08-10 05:04	+13h 23m
SC the schedule control	SC-3 decided	prereg/schedule_control.md with readout_schedctl.py, guard G-SC first, primary rung fixed at L=27	a8bcd80 08-13 00:19	schedctl 08-14 02:03	+1d 1h
BM the mirror of blocked	BM-2 partial	prereg/blocked_mirror.md with readout_blocked_mirror.py, comparability control read before the branch	3fcc1a0 08-16 12:16	blocked_mirror 08-16 15:06	+2h 50m
CL the ladder at a constant rate	CL-2 decided	prereg/constant_lr_ladder.md, four branches on the interior span against the cosine ladder's	a67e3a1 08-17 06:05	constlr_ladder 08-17 14:25	+8h 20m
HK resolution at higher k	HK-4 void	prereg/highk_resolution.md with readout_highk.py, guards G-HK1–3 read before HK-1–3	1e1daa0 08-17 14:37	highk 08-18 13:06	+22h 29m

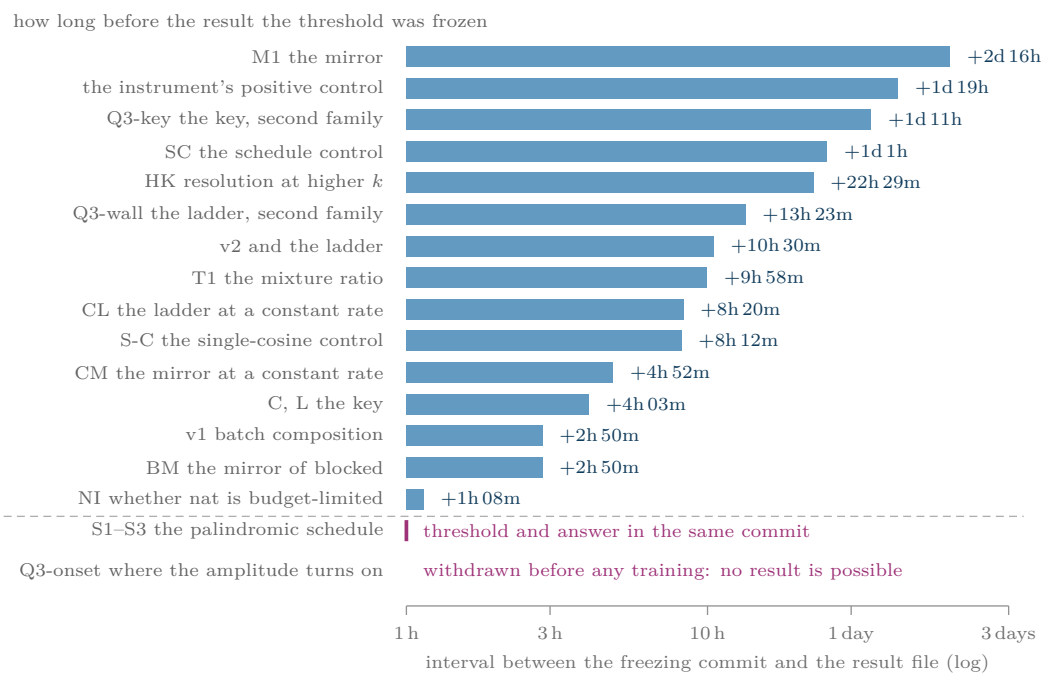


Figure 3. Seventeen registrations, and the interval each one bought. Table 4 on one axis: the distance between the commit that froze a threshold and the file that answered it. Fifteen run from about an hour to nearly three days ahead of their result. The two that are not intervals are drawn on the same axis rather than dropped from the count: the palindromic check put its thresholds and its answer in one commit, so it can assert an ordering it cannot demonstrate, and the amplitude-onset registration was withdrawn before any training because the rungs it named cannot be built at this budget.

claim	outcome	artefact carrying the threshold	registered	result (verdict file, mtime)	gap
CM the mirror at a constant rate	CM-4 void	prereg/constant_rate_mirror.md with readout_constant_mirror.py, guard CM-4 read first, magnitude registered as exploratory and turning no branch	no commit 08-26 05:43	constant_mirror 08-26 10:35	+4h 52m
NI whether nat's weakness is a budget artefact	NI-1 confirmed	prereg/nat_budget_gate.md with readout_nat_budget.py, guards G-1–G-3 and a capability bar computed from the family's own dispersion before the run	no commit 08-26 15:28	nat_budget 08-26 16:37	+1h 08m

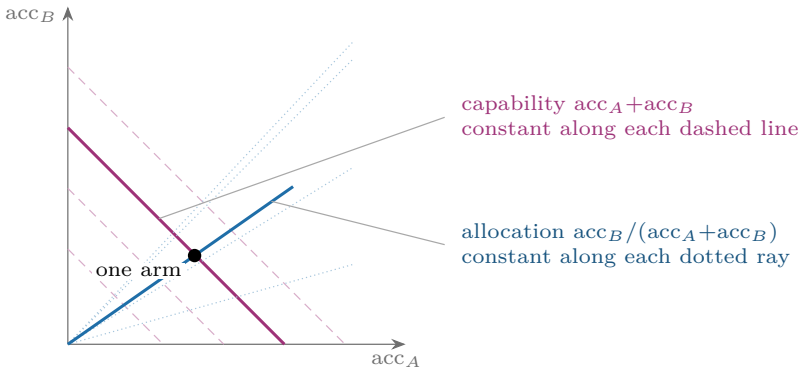


Figure 4. The read-out as a change of coordinates. Capability is constant along the dashed anti-diagonals, allocation along the dotted rays. A movement along a ray changes capability at fixed allocation and a movement along a diagonal does the reverse, and an undecomposed accuracy cannot tell the two apart although their meanings are opposite.

F Corpora, the Instrument, and Its Builds

Which corpus carries which claim. Table 5 prints the pair. The philosophical claim of this paper is a claim about nat, and nat is the weaker instrument. What travels between them is every qualitative statement the paper defends: capability conserved, allocation moved, interior flat, mirror antisymmetric at fixed C. What does not travel is size, and the direction is against us, so the headline figures in the abstract are cf2 figures. On the standard install guard — the minority convention's accuracy in the arm trained on nothing else — cf2 reads 0.4225 and nat 0.2091. That gap is mostly a budget gap, and we registered the cheap half of the separation before running it: the published pair is cf2 at lr 3×10^{-5} against nat at 1×10^{-5}, and rerunning the nat B-only arm at cf2's learning rate with nothing else changed raises the minority install from 0.2091 to 0.3738, 8.62 contrast floors at $t = 15.06$, closing 77.2% of the distance. A corroborating number was already on the result base and unread: at the published lr 1×10^{-5}, cf2's own install reads 0.1988 against nat's 0.2091, no separation. So "less than half" is a statement about two learning rates before it is a statement about two corpora, and the discount a reader rebuilding this on a both-correct conflict should expect is about seven eighths rather than about half. That is one arm of four and says nothing about the arrangement effect. One asymmetry the construction itself creates survives the correction: on nat the A-only arm scores exactly 0.0000 on the spelled convention across eight seeds, so the spelled form is never produced unless trained, whereas cf2's shifted form leaks at 0.0169.

The held-out set is a prefix, and the prefix is not skill-balanced. The held-out file carries 292 problems on nat and 342 on cf2, sorted by skill, and the evaluation takes the first 200. Every arm sees identical problems, which is what the contrasts require, but the set is not the pooled corpus: algebra and geometry are complete, combinatorics is truncated ($97 \to 48$ on nat, $129 \to 33$ on cf2), and calculus and physics are absent entirely. Every absolute capability level in this paper is therefore measured on three skills rather than five, and the difficulty coupling of Definition 2 on the same restricted set. Nothing here affects a between-arm comparison and everything here affects a level. It is reported rather than repaired because the checkpoints for most of these arms were reclaimed after evaluation, so the pool cannot be re-scored.

Prevalence. The rate at which public corpora disagree about answer form (12.7% of shared problems on the pair with the most multiplicity, rising to 27.1% against externally authored answer keys) is measured in a concurrent submission and is not re-derived here. What it licenses is the choice of construction rather than any number in this paper: it says the contested case is common enough to be worth an instrument, and every arm we report is built rather than found.

Table 5. Two corpora, and which one carries which claim. The last block is the arms nat does not have; they are not claimed for it. Sources: review_statistics, nat_replication.

	cf2 (flagship)	nat	status
what the two conventions are			
construction	boxed answer vs. that +1	numeral vs. spelled out	—
is either wrong?	yes, by construction	neither	claim is about nat
rows per condition	864	720	—
learning rate	3×10^{-5}	1×10^{-5}	nat ran colder
how strong an instrument each is			
minority install (B-only)	0.4225	0.2091	unmatched rates
at lr 3×10^{-5}	0.4225	**0.3738**	NI-1: 7/8
capability, interleaved arm	0.5202	0.3727	two thirds, unmatched
minority form leaked untrained	0.0169	**0.0000**	different regimes
what each corpus is asked to carry			
capability conserved	yes	yes	agrees
allocation moves	yes	yes	agrees
ladder interior flat	2.20 floors (10 arms)	1.92 floors (4 rungs)	agrees
mirror antisymmetric	+0.0850/−0.0875	+0.0183/−0.0229	sign agrees
mirror leaves C fixed	0.0025	0.0046	agrees
what only cf2 carries			
the 12.29σ switch	✓	not measured	cf2 only
schedule contrast, 11.63 floors	✓	not measured	cf2 only
the key (87.5% of ceiling)	✓	not measured	cf2 only

Table 6. The instrument, across its three builds. D is the arrangement contrast on the last-seen convention, signed: $D = \text{acc(last seen)}_{\text{blocked}} - \text{acc(last seen)}_{\text{shuf}}$ with the sources in the published order, so the switch is negative. Elsewhere the same quantity is quoted as a span between two arms and is therefore positive; −0.2346 here and +0.2275 in §5.2 are the same measurement at eight and at three seeds. The control is the same construction with both halves under one convention, so it must stay inert; in v1 it did not. v1 and v2 are three-seed builds; v3 and its control carry eight. Sources: the conflict_switch, conflict_switch2 and conflict_switch2b3lr3e5 verdicts.

build		control D	conflict D	conflict σ	verdict
v1	algebra pool, 1 epoch	+0.0461 (2.41σ)	+0.0020	0.10	positive control fails
v2	pooled skills, 3 epochs	+0.0017 (0.09σ)	−0.1425	7.46	control inert, switch fires
v3	the learning budget, lr 3×10^{-5}	−0.0028 (0.15σ)	**−0.2346**	**12.29**	control inert, switch fires

G Supporting Measurements

The power analysis behind §5.2's null, in full. Two arms of three seeds give four degrees of freedom, so the smallest difference detectable at 80% power is 3.72 contrast floors against an interior spanning 2.20, and even $k \to \infty$ leaves an MDE_{80} of 2.33 floors at the evaluation share measured in place. The one statistic that survives is the between-arm intraclass correlation of the allocation share, −0.076, bootstrap interval $[-0.409, +0.166]$, permutation $p = 0.61$: between-arm dispersion does not exceed within-arm seed noise, which is weaker than saying arm identity explains none of the variance.

The cluster count, and the registration it replaces. At $2\hat{\sigma}$ the ten arms occupy two clusters on both pretraining families, with inside-to-between gap ratios 8.7 and 20.7. The pre-registered second-family criterion was a span in floors and returns a miss: that family spans 2.85 floors against a bar of two, and the published first family spans 2.18 and misses the same bar — checkable before the run and not checked. The fault is in the registration, and the cluster count is therefore reported as post hoc.

G.1 The ladder in full, and every axis on one scale

G.2 Capability and allocation move on separate clocks

Figure 7 is the third bound on conservation, and its shape says something the range does not. Sampled every four optimiser steps through the first 54 steps of blocked's B stage, resuming from the completed A-only checkpoint, capability falls from 0.4900 to 0.3463 between steps 0 and 4 — 97.4% of its entire excursion — then climbs back to 0.4191 and stays. Allocation

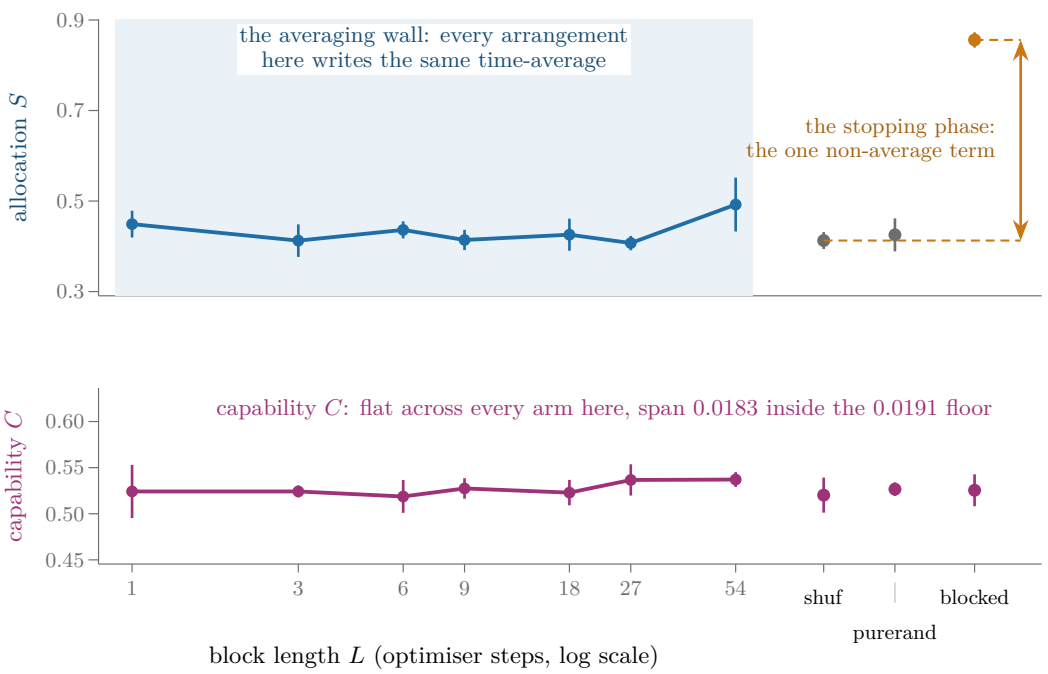


Figure 5. The averaging wall, measured. Same corpus, volume, budget and seeds; only the data path differs (Definition 1). (a) Allocation is flat from L=1 to L=27 and across shuf and purerand: 0.407 to 0.449 over a 27× change in block length and a change of batch composition. L=54 is the first rung off that floor at 0.492 and blocked sits at 0.856. (b) Capability over the identical arms spans 0.0183, inside the contrast floor: the path relocates commitment without changing what can be solved. Bars are the seed standard deviation, three seeds per point (shuf and blocked carry eight). Source: cf2b3lr3e5.

axis swept	on the path?	share, one end to the other	contrast floors
stopping phase, shuf against blocked	yes	0.4129 → 0.8692	23.89
model scale, 3B to 14B	no	0.1322 → 0.4732	17.85
mixture ratio, 1:1 to 2:1	no	0.4531 → 0.3116	7.41
cosine restart, byte-identical path	no	0.4071 → 0.2912	6.07
block length, L=1 to L=54	yes	0.4071 → 0.4922	4.46
batch purity at fixed block structure	yes	0.4129 → 0.4253	0.65

Table 7. Every axis this project has swept, on one scale, sorted by how far it moves the allocation. All six are measured on the allocation share of a mixed arm at one model and one budget, so the column is commensurable. The stopping phase is the largest term measured anywhere in the project, which is why it gets a section of the main text. The block-length ladder ranks fifth of six, and all three non-path axes move the allocation further than the whole ladder does. The path is a narrow channel and this paper's claims are about the path, but it is neither the narrowest thing measured here nor the only input to an endpoint. Source: axes.

over the same four steps goes 0.029 to 0.061, 4.5% of the range it eventually covers; the sweep from 0.03 to 0.72 happens between steps 12 and 32, by which time capability has returned to within a floor and a half of where it ends. Two consequences, the first against us. The dip is not the reassignment: whatever costs capability at the start of conflicted training is a different process from the commitment shift this paper is about. Second, the dip is not a scoring artefact of the mixture — at k=4 a model randomising between two mutually exclusive golds leaves the sum untouched, so a drop in the sum is answers matching neither. What we cannot say is which: a trajectory recording only accuracy cannot separate knowledge briefly lost from output the scorer briefly could not parse. That needs a run that keeps generations, and it is listed as a missing cell rather than argued.

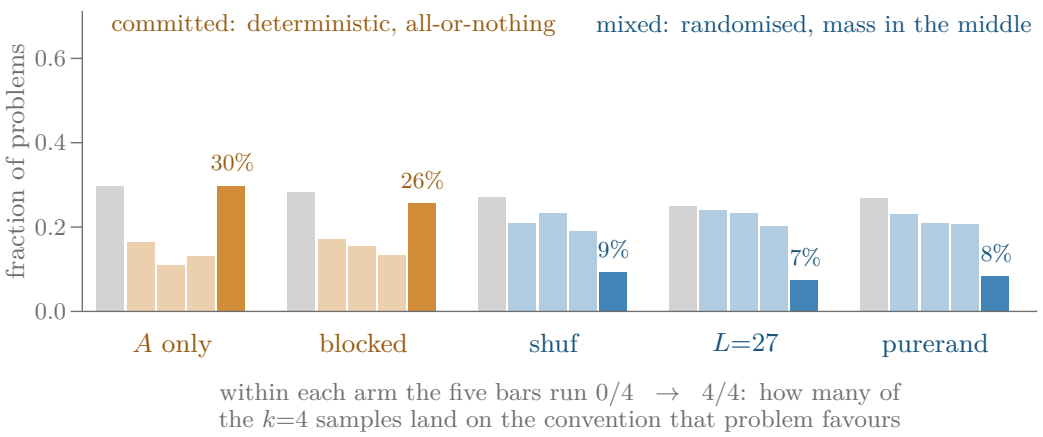


Figure 6. Conflict makes the policy stochastic, not the model weak. Per-problem commitment at k=4 samples on the synthetic conflict, three seeds, 200 problems each. Bars within an arm run 0/4 to 4/4. The committed arms are bimodal — the model either answers a problem the same way every time or cannot answer it — while the mixed arms move that mass into the middle without changing the unsolved fraction (grey, leftmost bar in each group). Same corpus, same volume, same budget; only the data path differs. One binning throughout: counting each problem's best convention gives a factor of about three and a half and counting each (problem, convention) cell separately about three, and the figures in §5.1 use the first. Source: cf2b3lr3e5 per-problem arrays.

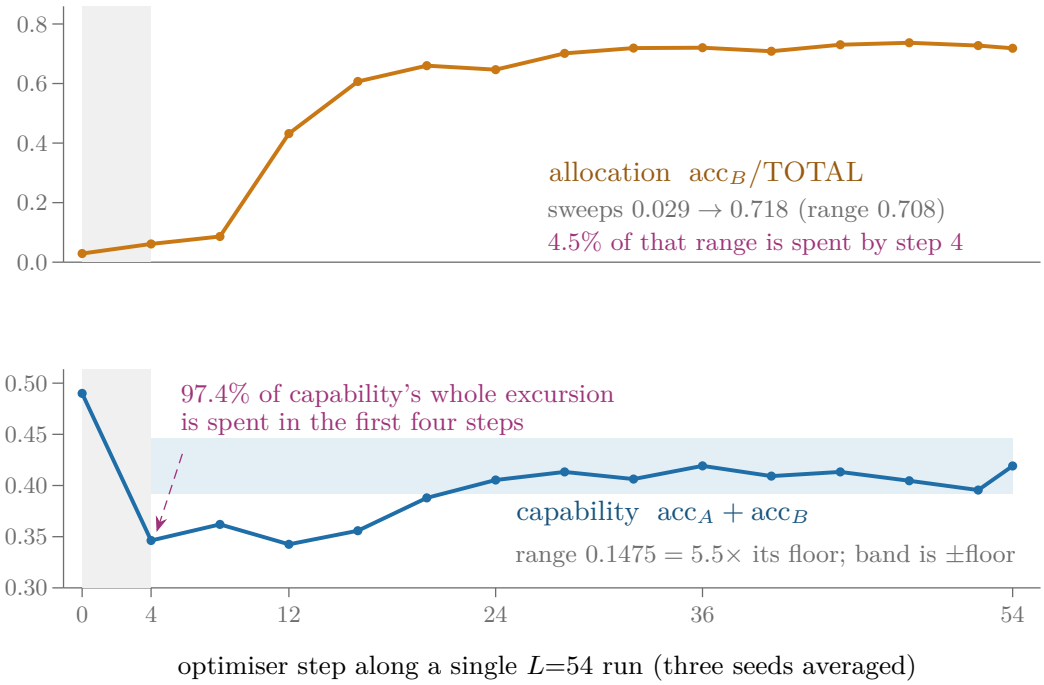


Figure 7. Capability and allocation move on separate clocks, which the range alone hides. The blocked construction's second stage, checkpointed every four optimiser steps over its first 54, three seeds averaged. (top) allocation acc$_B$/TOTAL; (bottom) capability with a band of ± its floor 0.0270 around the endpoint. The shaded column is the first four steps.

G.3 The stopping phase: the mirrors in full

The mirror of blocked, and what its control permits. Training all of B then all of A against the published A then B relocates the share from 0.8692 to 0.0343, crossing the mean-field point, with capability moving 0.0020, well inside the 0.0270 capability floor. The registered branch is that the mirror goes to the other side but not by the same magnitude, $|-0.4657|$ against $|+0.3692|$, a gap of 5.05 contrast floors where the branch allowed two. The

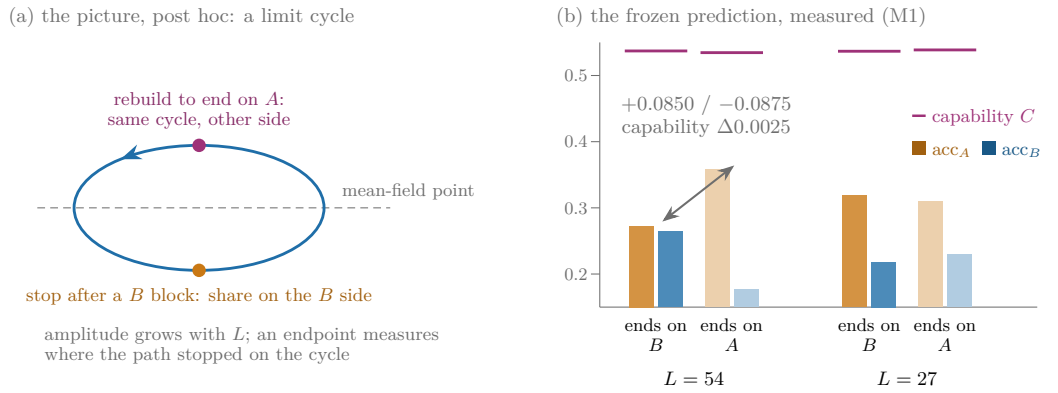


Figure 8. The stopping phase is a position on a cycle, and the mirror moves it to the other side. (a) The reading of Proposition 3, labelled post hoc: a driven system oscillates about the mean-field point with amplitude growing in L, and an endpoint measures where the path stopped. (b) The measurement: at L=54 the shift is antisymmetric, +0.0850 against −0.0875, with capability moving 0.0025. At L=27 both constructions sit at the floor and the mirror manufactures nothing, as it must.

comparability control fires, and it separates the two halves of that result. The published checkpoints were reclaimed, so the mirror had to be scored on a node whose kernel cache was cold; before reading it we re-evaluated a published checkpoint that does survive, the B-only arm the mirror resumes from. It reproduces the majority accuracy (0.4250 against 0.4238, 0.06 floors) and does not reproduce the minority one (0.0900 against 0.0575, 1.70 floors). Small accuracies moved and large ones did not — which is exactly the regime the mirror's share lives in, since at $S = 0.0343$ the numerator is $\mathrm{acc}_B \approx 0.0175$; propagating the drift puts S anywhere in $[0.000, 0.092]$. So: the crossing is robust — across the control's entire drift the share stays far below the mean-field point, so blocked is a position on a cycle rather than a distinct mechanism — and the symmetry is not readable, since the magnitude gap is 0.0964 and the control shows the small-accuracy regime can move 0.0325, 34% of it. We report the branch label and decline to interpret it.

The mirror at a constant rate, and a guard that could not have passed. Every mirror above trains under a cosine, so the branch of Proposition 2 that speaks most directly to a mirror — the terminal term carrying $\eta(T)$ with no contraction discount — is the one branch with no schedule contrast. The cell was registered on a sign and a symmetry and explicitly not on a magnitude. The registered outcome is void: capability moves +0.0358 between the two arms, 1.33 capability floors, past the one-floor bar the registration set as the condition for reading a share. The guard could not have passed, which is a defect in the registration rather than a property of the result: post hoc, the three seeds give −0.0238, +0.0612 and +0.0700, a sign change with a standard deviation of 0.0518 that is itself 1.92 capability floors, and §5.2 had already measured that capability spans 2.5 capability floors across the constant family. A bar of one capability floor set for a family reported at two and a half is not stringency. What the run printed, and what we decline to read: $\Delta\mathrm{acc}_A = +0.4150$ against $\Delta\mathrm{acc}_B = -0.3792$, +21.73 and −19.85 contrast floors, consistent across the three seeds. We print the numbers so that no reader need wonder what we saw, and we do not read them, because the arms whose difference they are are not matched in what they can do. Closing the cell needs 17 seeds at the observed dispersion.

The onset region cannot be refined at this budget. On the second pretraining family the L=54 mirror returns a miss, and taken with the rung it was built on that is what the picture requires: on that family L=54 sits inside the flat region rather than above it, so there is nothing at that block length to relocate. The test that would decide it is a mirror at a rung where the amplitude has appeared, and the ladder cannot supply one. The rungs are quantised by the epoch: the builder writes a one-epoch ordering which the trainer passes over three times, so a block must divide the 54 chunks each source contributes per epoch — divisors $1, 2, 3, 6, 9, 18, 27, 54$, exactly the rungs already run. blocked is not on that ladder at

Table 8. What the key opens, and what it cannot. The read-out was frozen in the job script before the runs (C on the switch, L on the level). The residue has a mechanism rather than an excuse: a marked B-only model, trained on nothing but the spelled convention, answers in numerals when the key asks for numerals, but a marked A-only model asked for a word cannot produce one it never saw. Source: unlock_verdict.

read-out	quantity	unkeyed	keyed	verdict
C	arrangement switch D	−0.0985 (5.16 floors)	**−0.0006** (0.03 floors)	C1: it collapses
L	mean accuracy	0.1863	**0.3259**	L4: partial unlock
	against ceiling 0.3727	50.0%	87.5%	74.9% of the gap
	against L1 target 0.3354	—	short by 0.50 floors	
the residue, and why it is coverage rather than difficulty				
	marked B-only on the numeral evaluation	0.1283	**0.3767**	the key selects
	marked A-only on the spelled evaluation	—	**0.0025**	it cannot create

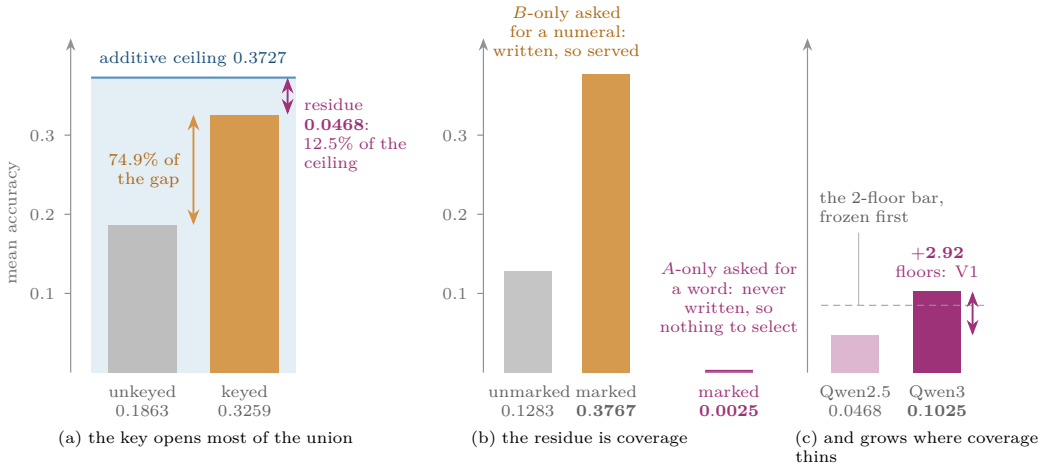


Figure 9. What the key opens, and why what it leaves is a different obstruction. (a) The unkeyed model sits at exactly half the additive ceiling, which is what a coin flip between two mutually exclusive golds scores; the key recovers 74.9% of the gap and leaves 0.0468. (b) That residue is coverage: marked B-only, trained on nothing but the spelled convention, answers in numerals at 0.3767 because numerals are the pretrained default and are therefore written; marked A-only asked for a word scores 0.0025, because there is nothing to select. (c) The residue grows on a family that writes four times less of the minority convention, +2.92 floors against a bar of two frozen before the run. It is a necessary-condition test. Sources: unlock_verdict, q3unlock_verdict.

all (L=162), and the interval $54 < L < 162$ is not a finer division of an epoch but a coarser one spanning several. Such a rung is writable; what forbids it is that writing the path across epochs means training one pass over a tripled file, which is precisely the construction the single-cosine control measured as not learning-matched (0.4442 against 0.5125, 2.5 capability floors). At a fixed multiset and a fixed budget, the path space cannot be refined in the region where the stopping phase turns on, because every route into that region changes what is learned. That is a property of the instrument in the same sense that a diffraction limit is a property of a lens. The registration for that region was withdrawn before any training, having spent nothing.

G.4 The key, and the results asked to travel

One instrument correction, because it fired first. The key's read-out audits the instrument before printing any branch, and on first execution it refused: eighteen of twenty keyed arms contained problems scored correct under both conventions, which the audit called impossible. The audit was wrong. Exclusivity holds for the unmarked corpus, where one prompt carries two mutually exclusive golds; the marked evaluations are two different prompts, so a model that serves both scores on both — which is what this section claims the key buys. The check was corrected against the published run: the keyed family carries 17–19 such problems per arm while the unmarked families carry exactly 0. The count is now recorded rather than

Table 9. The key, and the ladder, asked to travel. Both were registered before their runs with the branches and thresholds frozen, and both gates were read first. The key replicates and its residue moves as the coverage reading requires. The ladder's registered branch is a miss, and the row below it is why that verdict is reported together with the threshold it was read against: the published first-family ladder does not meet the same bar, which was checkable before the run and was not checked. Sources: q3unlock_verdict, q3ladder_verdict.

	quantity	Qwen2.5-7B	Qwen3-8B-Base	read-out
guards	minority convention written, $\mathrm{acc}_B(B$ only)	0.2091	0.0520	—
	on the synthetic conflict	0.4225	**0.4813**	G1 passes
	switch present, $\lvert D\rvert$ on that corpus	11.9 fl	**15.0** fl	G2 passes
the key	D unkeyed	−0.0985 (5.16 fl)	−0.0875 (4.58 fl)	
	D keyed	−0.0006 (0.03 fl)	−0.0125 (0.66 fl)	K1 collapses
	marked mean against its ceiling	87.5%	75.5%	L4 partial
	residue = ceiling − marked	0.0468 (2.45 fl)	**0.1025** (5.37 fl)	V1 +2.92 fl
the ladder	flat region span, L=1–27, purerand, shuf	0.0416 (2.18 fl)	0.0544 (2.85 fl)	W2 at a 2 fl bar
	which the published row also misses	differ by 0.67 floors		post hoc
	span as a fraction of the blocked excursion	9%	12%	post hoc
	resolvable endpoint states (Def. 3)	**2**	**2**	R_{real}=1 bit
	within-cluster / between-cluster gap	8.7×	20.7×	post hoc
	L=54	0.4922, off the floor	0.4547, inside it	
	blocked	0.8692	0.8907	
	mirror at L=54: $\Delta\mathrm{acc}_A$, $\Delta\mathrm{acc}_B$	+0.0850, −0.0875	+0.0204, +0.0012	M2

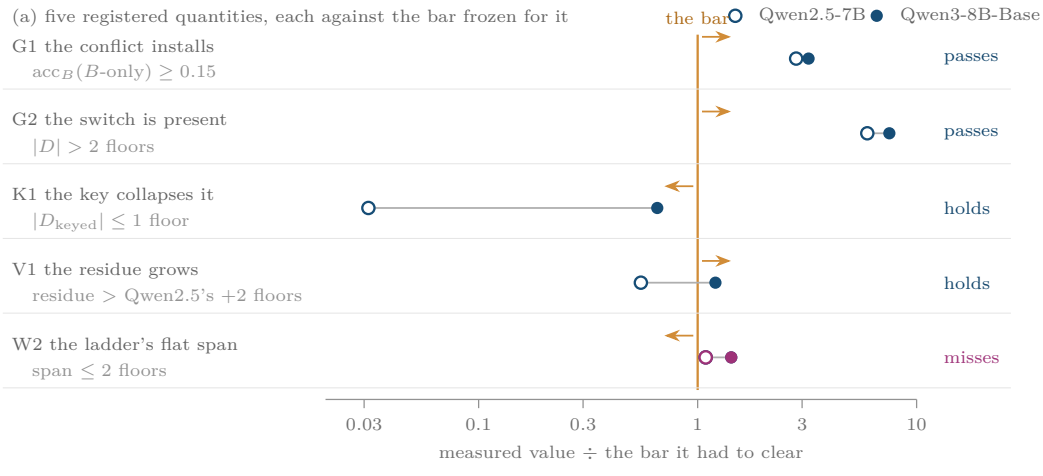
Table 10. The share tracks the mixture weights; the blocked arm does not (natural-conflict ratio corpora, 936 rows, three seeds; read-out rules frozen in the job script the scheduler executed). The 3:1 form was excluded by its own learnability guard and redesigned at 2:1; the redesign is labelled, not laundered.

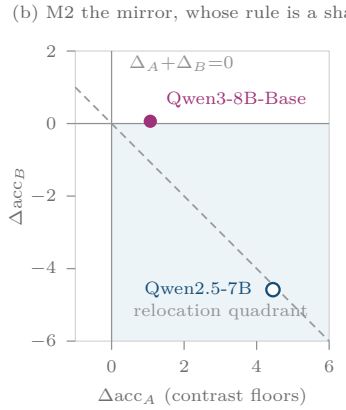
	1:1 (rr11)		2:1 (rr21)	
	share	capability	share	capability
mean-field prediction	0.500	—	0.333	—
shuf (measured)	0.453	0.372	0.312	0.377
blocked A→B	0.731	0.380	0.699	0.381
B only (guard: $\mathrm{acc}_B = 0.263/0.187 >$ floor)	0.713	0.368	0.595	0.315

treated as a failure, and is a coverage signal of its own: 2 on the second family against 17–19 on the first.

G.5 The mixture-ratio test, and what a better instrument would have been

The mean-field predictions were tested as point values, and a directional test would have been the better instrument. The ratio test compares an installed share against 0.500 and 0.333 with a frozen tolerance of 0.05, and both cells miss on the same side, toward the pretrained prior. Two misses in the same direction are evidence for a model this test cannot express: mean field plus a prior-bias term of unknown size. A point test with a tolerance cannot separate that from a failure of mean field, whereas a directional test can, and the design is a third ratio: a 1:1, 2:1, 4:1 sweep testing monotone decrease and the slope rather than three point predictions. We did not run it, and we flag the current test as the weaker instrument rather than reporting its two misses as though the alternative model were not on the table.





The mirror rebuilds L=54 to end on A instead of B and asks the endpoint to move the same magnitude to the other side. On Qwen2.5 it does: +4.45 against −4.58 floors, on the dashed line of pure relocation, with capability moving 0.09 of a capability floor. On Qwen3 both skills move up, +1.07 and +0.06, which is a capability change of 0.80 capability floors and not a relocation at all. The two families disagree about the shape rather than about a threshold, which is why this row is drawn and not scored; L=54 is inside the flat region on Qwen3 (panel (a), W2), so there is no amplitude there for a mirror to move.

Figure 10. Three results asked to travel, and the answers are not uniform. (a) Every quantity whose threshold was frozen before the second family ran, divided by that threshold, so the bar sits at 1 for all of them and the arrow says which side had to be cleared. Hollow is Qwen2.5-7B, filled is Qwen3-8B-Base. The guards pass well: the conflict installs better on the new family, so nothing below is an instrument at its floor. The ladder's span misses and so does the published row at 1.09, which is a fault in the registration rather than in the second family. Rows are ordered by what they decide, not by outcome, and the post-hoc statistic is ruled off. (b) The mirror is the one registration with no scalar bar: its rule is a shape, so it is drawn as one.

Table 11. The one lever the averaging limit leaves free. Proposition 1's mean field is weighted by source proportions, so changing the mixture ratio must move the allocation and changing the path must not. Both halves are measured on the same corpus family. Tolerance 0.05 was frozen with the prediction; both misses fall on the side of the pretrained prior.

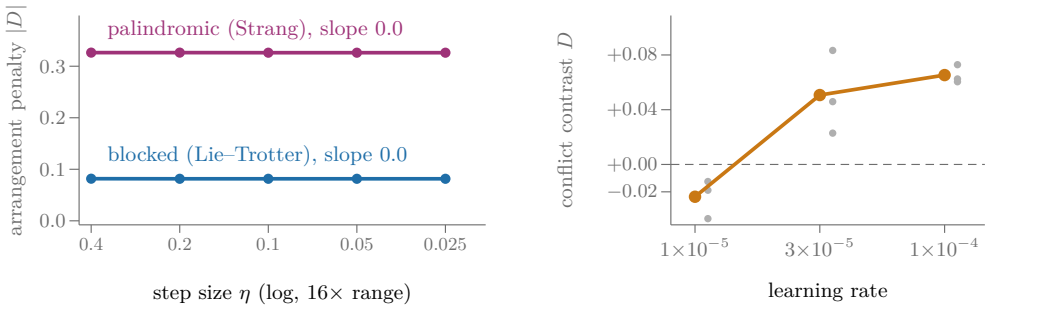
	allocation share of B			
intervention	mean-field prediction	measured	miss	moves the share?
mixture ratio 1:1 → 2:1	0.500 → 0.333	0.453 → 0.312	0.047, 0.021	yes, as predicted
block length L=1 → 54	no change predicted		—	no (0.407–0.492)
batch purity, fixed blocks	no change predicted		—	no (3% of the span)
palindromic reordering	no change predicted		—	no (and 4× worse)

G.6 The palindromic schedule, in full

One qualification about that registration. The other registrations in this paper are frozen in a job script committed hours to days before the run it launched. This check is a linear-model computation of a few seconds whose thresholds and answer entered the repository in the

Table 12. The construction splitting theory recommends, and what it did. Registered in the linear model, where every arm is computable exactly. S3 is an implementation guard, not a claim: a palindromic schedule is its own reverse, so its order effect must vanish, and it does. S1 and S2 are the claim and both fail: S1 by a factor of seven against its own bar, and S2 degenerately, since refining the step by 16× moves either penalty in the fifth decimal. Source: splitting_verdict.

check	registered bar	measured	
S3 order effect of a palindrome	ASYM $\equiv 0$	0 to machine precision	passes
S1 median $\lvert D_{\text{strang}}\rvert/\lvert D_{\text{blocked}}\rvert$	< 0.5	**3.65** over 300 draws (IQR 1.88–8.09); 3.7% of draws below the bar	fails
S2 log–log slopes in η differ	by ≥ 0.5	both slopes 0.0 to four figures across $\eta = 0.4 \ldots 0.025$	fails
the penalty across a 16× refinement of the step, which is what "no order to improve" means			
$\eta = 0.4, 0.2, 0.1, 0.05, 0.025$			
blocked		0.081646, 0.081638, 0.081634, 0.081632, 0.081631	
palindromic		0.32637, 0.32634, 0.32631, 0.32630, 0.32629	



(a) S1 and S2 both fail: the designed schedule is 4× worse and neither slope moves

(b) leaving the lazy regime moves D $-0.0236 \to +0.0653$; $|D|$ grows at every step

Figure 11. Two ways of leaving the averaging wall, one designed and one accidental. (a) The palindromic schedule against the blocked one it was built to beat, in the linear model where the joint arm is available in closed form. It is roughly four times worse, and refining the step by 16× moves either penalty in the fifth decimal: there is no order to improve because the penalty is not a discretisation error. (b) The one intervention that does move the contrast is leaving the lazy regime. Raising the learning rate grows $|D|$ monotonically, $0.0236 \to 0.0653$, through a change of sign; the highest rate also costs capability, 0.268 against 0.340, so it is the system leaving the regime the averaging argument is tight in.

same commit, so no earlier artefact carries the prediction. S1–S3 are stated as supported-if conditions rather than as a description of what happened, and the misses are wide enough (3.65 against 0.5; both slopes zero) that no threshold choice rescues them. But the ordering the other tests demonstrate, this one can only assert.

G.7 Budget, learning rate, and the commutator proxy

The stopping term in-model: real, and not where the theory says. If Proposition 3 makes the stopping phase the only π-dependent term of the path, a linear model in the lazy regime should reproduce the sign reversal we measure along the budget axis — and it does, without reproducing its location. Sampling random non-commuting task pairs inside the lazy regime and sweeping the budget T: the order effect changes sign in **72**% of draws and the contrast $D(T)$ in **70**%, stably across step sizes, so budget-dependent reversal is generic and not an artefact of our corpora. But the leading Magnus/commutator proxy $T^\star$ — what a practitioner would compute once at initialisation to predict where the crossing sits — correlates with the measured crossing budget at Spearman **0.031** over 278 pairs. That is, not at all. Median rotations at the crossing are small (11.2°), so the reversal is a near-cancellation, which is exactly why its location is hard.

Table 13. The headline pair, swept over the resolution constant it is defined at. R_{real} counts the endpoint states the ten arms occupy under single linkage; R_{cap} is the logarithm of the range in units of the resolution and therefore gains exactly one bit per halving, which R_{real} does not. Two clusters is what both families give from $2\hat{\sigma}$ upward, and the second family gives it from $1\hat{\sigma}$ upward. Below $2\hat{\sigma}$ the first family resolves a third state, so the paper's constant sits at the lower edge of the interval where the two families agree rather than in its middle. Source: rate_sensitivity.

resolution	$0.5\hat{\sigma}$	$1\hat{\sigma}$	$1.5\hat{\sigma}$	$\mathbf{2}\hat{\sigma}$	$3\hat{\sigma}$	$4\hat{\sigma}$
Qwen2.5-7B, clusters	4	3	3	**2**	2	2
R_{real} (bits)	2.00	1.58	1.58	**1.00**	1.00	1.00
R_{cap} (bits)	5.30	4.30	3.72	**3.30**	2.72	2.30
Qwen3-8B-Base, clusters	4	2	2	**2**	2	2
R_{real} (bits)	2.00	1.00	1.00	**1.00**	1.00	1.00

The commutator proxy on the real model. Measured on Qwen2.5-7B activations at layer fraction 0.75 across the matched pairs, it ranks the order effect at Spearman $-\mathbf{0.543}$ at one epoch and $-\mathbf{0.600}$ at three: the wrong sign, consistently, on both budgets. Four of the six pairs flip sign between budgets, so there is no fixed ordering for a static proxy to predict. Proposition 2 says why a score computed once at θ_0 must decay with τ_L: it is the leading term of the interior integral in Eq. (1), whose neglected remainder grows with block length. This is a range boundary rather than a refutation of Sweeney (2026a), whose tournament reports 98% pairwise accuracy at block length k=1 falling to 73% at k=20, against our 162–324 steps per block.

Learning rate moves the conflict contrast monotonically. $|D|$ grows monotonically, $0.0236 \to 0.0653$, as the step size grows across three learning rates at three epochs on the natural conflict — the direction the suppression argument requires, since raising the learning rate is the one intervention that moves the system furthest from the lazy regime in which arrangement is doubly suppressed. The highest rate also costs capability (0.268 against 0.340), so the growth in $|D|$ is not a free improvement: it is the system leaving the regime in which the averaging argument is tight. Separately, on an algebra–combinatorics pair the published order effect reverses rather than attenuates when the budget is raised ($+0.0779$ to -0.0619, three seeds, s.e. $= 0.0191$), and the probe says why the two budgets are different regimes: at one epoch and lr 3×10^{-5}, fine-tuning damages three of five skills relative to zero-shot, so an order effect measured there is a statement about that regime.

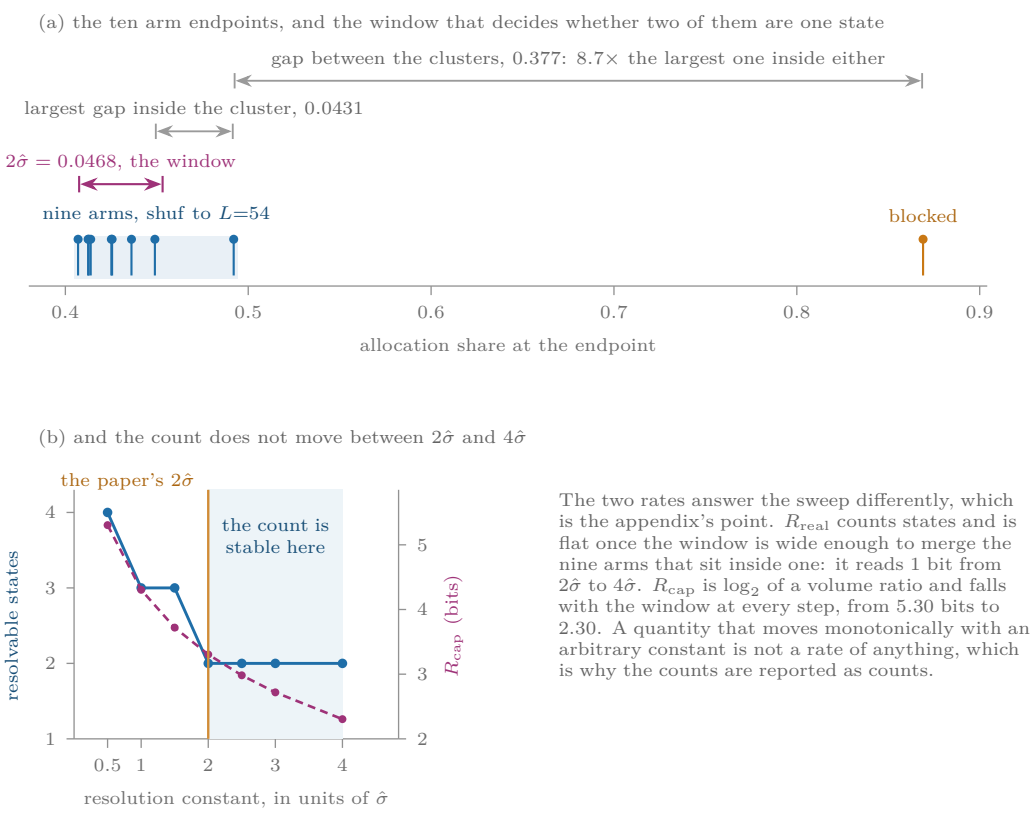
G.8 The clustering rule, and why it is reported as a count

Definition 3 (The write channel, its capacity, and its realised rate). The path carries $\log_2 \binom{n_A+n_B}{n_A}$ bits of source-order information. Let $\Phi(\pi) \in \mathbb{R}^d$ be a behavioural read-out of the endpoint and $\hat{\sigma}$ the seed dispersion. Two endpoints count as distinguishable when they are separated by $2\hat{\sigma}$. The factor of two is a convention and the headline depends on it, so Table 13 sweeps it and Figure 12 draws the sweep. The write channel is the induced map $\pi \mapsto \Phi(\pi)$, and it has two rates that must be kept apart:

$$\text{capacity} \quad R_{\text{cap}} = \log_2\big(\text{range}\,\Phi/2\hat{\sigma}\big), \qquad \text{realised rate} \quad R_{\text{real}} = \log_2\big|\Phi(\Pi)/{\sim_{2\hat{\sigma}}}\big|,$$

where Π is the family of paths actually run and $\sim_{2\hat{\sigma}}$ is single-linkage at the same resolution. R_{cap} counts the distinguishable values the range supports — the rate an encoder would achieve if it could place an endpoint anywhere in that range; R_{real} counts the values the path family occupies, and upper-bounds the information any encoder built from this family can transmit, however many paths it uses. The two coincide only when the image is spread; Proposition 1 says it is not, so the paper's central measurement is the gap between them, and quoting R_{cap} as though it were R_{real} is the error the gap corrects.

Three things Definition 3 is not. First, $\log_2 \binom{n_A+n_B}{n_A}$ is the entropy of the design space under a uniform measure over all orderings. We ran ten. An ensemble of ten codewords carries at



The two rates answer the sweep differently, which is the appendix's point. R_{real} counts states and is flat once the window is wide enough to merge the nine arms that sit inside one: it reads 1 bit from $2\hat{\sigma}$ to $4\hat{\sigma}$. R_{cap} is $\log_2$ of a volume ratio and falls with the window at every step, from 5.30 bits to 2.30. A quantity that moves monotonically with an arbitrary constant is not a rate of anything, which is why the counts are reported as counts.

Figure 12. The count, and the two reasons it is a count rather than a rate. (a) The ten arm endpoints of the cosine family on the share axis, with the $2\hat{\sigma}$ window that decides whether two of them are one state drawn at the size it actually is. Nine arms fall inside a span whose largest internal gap is 0.0431; the tenth sits 0.377 away, 8.7 times that gap. (b) The constant swept. The number of resolvable states is flat from $2\hat{\sigma}$ to $4\hat{\sigma}$, while R_{cap}, which is $\log_2$ of a volume ratio rather than a count, falls at every step. Sources: rate_verdict, rate_sensitivity.

most $\log_2 10 = 3.32$ bits, so no measurement in this paper can exhibit a compression larger than 3.32:1 — and that is the one it does exhibit: ten paths, two states. Second, neither rate is a mutual information: $I(\pi; \Phi)$ would require an explicit prior on paths and a noise model for seed dispersion, and we estimate neither; R_{real} is a count of resolvable clusters wearing a logarithm, and the logarithm is a naming convention rather than a result, which is why the body reports counts. Third, Φ here is one scalar, and a degenerate image in one coordinate is not a degenerate image; adding a coordinate can only raise the number of occupied states, so two is a floor. The statement the paper defends needs none of this vocabulary: ten paths spanning 0.4621 in allocation, at a resolution of $2\hat{\sigma} = 0.0468$ that their own span would divide into about ten distinguishable values, occupy two. The channel is not narrow because it cannot resolve; it is narrow because the encoder is degenerate, which is what Proposition 1 predicts.

G.9 The schedule control, in full

What we cannot say is why the schedule pushes toward A. The registered branch reports the movement as unexplained rather than folding it into the branch that would have closed the question, and we take that in full. Two accounts we had ready are excluded by the data rather than by argument. The registration predicted that if a mid-run stopping phase were the mechanism the two rungs would move in opposite directions, because the boundary falls on an A block at L=54 and a B block at L=27; they move the same way. And a bias from which source occupies the restarted cosine's peak fails for the same reason, since L=54's second stage begins on B.

Table 14. The schedule control: identical path, one cosine against two. The treatment consumes the control's own stream in the same order and differs only in that the cosine restarts and the optimiser state resets at the midpoint, which is blocked's structure. The guard that voided the single-cosine control is read first and passes on both rungs. L=27 was fixed as primary before either arm ran, because its control dispersion is four times smaller. Its movement is 6.07 contrast floors at $t = -10.1$. L=54 moves further and does not clear its own noise: its treatment seeds span 0.40, wider than the effect, and the registration recorded in advance that at this dispersion its silence is not evidence of absence. Both rungs move the same way, which the registration fixed in advance as the signature of the schedule rather than of a mid-run stopping phase. Source: schedctl_verdict.

	share				
rung	control (one cosine)	treatment (two)	Δ	t (crit 2.776)	guard $\lvert\Delta C\rvert$
L=27 primary	0.4071	**0.2912**	**−0.116**	**−10.10**	0.59 fl
	0.413, 0.419, 0.389	0.294, 0.302, 0.278	6.07 fl	SC-3	passes
L=54 secondary	0.4922	0.2053	−0.287	−2.35	0.26 fl
	0.424, 0.526, 0.527	0.024, 0.424, 0.168	15.02 fl	SC-2	passes
for scale: the entire block-length ladder, L=1 through 54, moves the share 0.085					

Table 15. Per-problem transitions from shuf to blocked (synthetic conflict, three seeds × 200 problems). A problem's state is the convention receiving more of its k=4 samples (A, B, tie) or unsolved (U). The solved set is nearly fixed while nearly half of it changes label, 26× more often toward the convention trained last. The tie row carries the 3 transitions $U \to$tie and the 7 tie$\to U$, which is why the newly-solved and lost counts (44 and 51) exceed what the four U rows sum to (41 and 44). Source: cf2b3lr3e5 per-problem arrays.

transition	count	frac	transition	count	frac
$A \to B$ (re-labelled)	179	0.298	$U \to B$ (newly solved)	26	0.043
$U \to U$ (never solved)	119	0.198	$A \to U$ (lost)	28	0.047
$B \to B$ (stable)	104	0.173	$U \to A$ (newly solved)	15	0.025
$A \to A$ (stable)	28	0.047	$B \to U$ (lost)	16	0.027
$B \to A$ (re-labelled)	7	0.012	ties, either side	78	0.130

Why the void control cannot be repaired cheaply. Under a single cosine the terminal B block trains in the decayed half of the schedule, so B is learned less and A is forgotten less: at one seed, acc_A rises from 0.078 to 0.215 while acc_B falls from 0.450 to 0.224. That is the recency-by-learning-rate interaction of §G.7, and it is exactly why the registration said in advance that this control measures two-stage against one-stage rather than the cosine alone. The share the void forbids us to read is 0.5089, which sits 0.87 floors above the ladder's top rung; we print the number so that no reader need wonder what we saw, and we do not read it, because the arms whose difference it is are not matched in what they learned.

G.10 Per-problem transitions

H Open Cells and Cross-Predictions

None of the following is run to a readable answer, and each is listed at the size we think it is. (i) Separating the two things a constant rate leaves open. Proposition 2 attributes the constant-rate family's 11.63 floors to the terminal weight $w(T)$; Sweeney (2026b) attributes order sensitivity instead to AdamW's buffers advancing on step count rather than on $\tau = \eta k$. Both predict a resolved ladder at a constant rate and the present design cannot tell them apart. One arm decides it: the same ladder without bias correction, or under SGD with momentum, where the fixed clock is absent and only the terminal weight remains. (ii) Eight seeds on the main rung, since every pairwise statement in §5.2 is limited by three seeds rather than by the effect. (iii) A conflict outside mathematics: every corpus here is competition mathematics with integer answers, so the convention is surface form. (iv) A run that keeps generations, to separate the 0.1475 capability excursion into knowledge briefly lost and scorer briefly unable to parse. (v) The mirror at a constant rate at enough seeds to read it — this cell has run and returned void, and it needs 17 seeds rather than more

arms. (vi) A difficulty-unweighted allocation: S is a p-weighted share and $\mathrm{Cov}(p, s) \neq 0$ is measured rather than assumed away, and the unweighted $\hat{\mathbb{E}}[s]$ over a fixed problem set needs no new training. (vii) The switch itself at matched budget, about 56 GPU-hours. (viii) A difficulty-matched control for the 12.5% residue, without which the two-wall decomposition of §5.4 is an interpretation with a plausible alternative.

The sharpest open test of Remark 1, and what happened when it ran. If the mixture is the optimum of an underdetermined query rather than a limit on what can be stored, then a conflict of the same size whose convention is a function of the input should be resolved with no key at all, at the same k, volume, budget and architecture. That cell has since been run by a concurrent submission and decided nothing: a corpus keying the convention on a single-character feature of the problem statement ran twice at this scale, the pooled read-out was voided by its own symmetry check, the paired within-problem contrast registered afterwards returned a miss ($\Delta = +0.0018$, $t = 0.23$), and the audit of the void found the keying feature nearly collinear with per-item scoring asymmetry. The reading of Remark 1 is therefore pending on that corpus, not supported and not refuted, and any rebuild must first pass an orthogonality certificate. Two things about a rebuild have to be said now rather than rediscovered. First, the read-out cannot be Definition 2's: setting $H(\mathrm{conv} \mid I) = 0$ is exactly the statement that each held-out problem has one applicable convention, so there is no allocation left to split, and the capability–allocation coordinates this paper is built on do not survive the construction that tests them. Second, the keying feature must be uncorrelated with difficulty: assigning the convention by skill, which is the obvious construction, fails that, because skills differ in how often they are solved at all. The feature has to ride on the problem rather than partition the problems.

Two cross-predictions, stated as missing cells. X1, the mixture-capacity wall: with N mutually exclusive conventions rather than two, the installed mixture's calibration against the mixture weights should degrade with N at fixed data per convention; a clean break would be training's capacity wall as distinct from the underdetermination wall of Remark 1. X2, replay as deferral: mixing a fraction ρ of first-stage data into the final stage should defer the commitment flip of §5.3 by a coefficient $1/\rho$, connecting this result to replay in continual learning (Lopez-Paz & Ranzato, 2017) with a quantitative prediction rather than a metaphor. Neither is run; both are cheap; either failing would cut the account at the joint we have named.

Where the correspondence stops. A third position exists and is deliberately not claimed here: a system that writes its own conclusions into a store it later retrieves from. There the bounded state is neither the parameters nor the context window but an accumulating record; capability can be conserved by the store's construction rather than by an averaging limit; and what fixes the allocation is a retrieval rule rather than a stopping phase. Definition 2 is written for parameters trained once over a fixed corpus, and its conservation is approximate and measured. Whether the same two coordinates describe an accumulating store is a question about a different object, and nothing in this paper is evidence either way.