The Free-Recipe Limit: Every Recipe Effect Measures Which Premise Broke

Anonymous authors
Paper under double-blind review

Abstract

Post-training pipelines must choose which skills to train together, whether to run one blocked stage per skill or interleave them, and in what order, and a large literature reports that these choices matter. Whether any of them can be decided depends on two quantities that are never reported together: how far apart the best and worst recipes can be, and how finely a re-execution of the whole protocol can tell two recipes apart. Where the second exceeds the first, no quantity of search converts into a decision, and the signature is not an absence of winners but winners that do not survive re-running. We measure both sides for supervised fine-tuning on competition mathematics, pricing the resolution from nested variance components over repeated executions of the protocol rather than from seed spread alone. Across six skill pairs the largest order contrast on any single corpus is 0.062 while one execution resolves only 0.081 – a ratio of **0.76** that is an upper bound, since its numerator is the maximum of a search – and re-running the winner reverses its sign. Most of that noise is cheap to remove: at four samples per problem, evaluation sampling rather than training stochasticity carries 0.955 of the variance of the order contrast itself, so quadrupling the samples costs minutes and beats a fourth training seed costing hours. One departure clears the wall: split a corpus so its halves answer under incompatible but equally correct conventions, and arrangement stops moving capability and starts moving which convention the model commits to, by two orders of magnitude over its same-convention control, across four scales in two pretraining families. Recipe ablations should therefore publish their re-execution resolution beside their effect; a public scorecard grades all 30 preregistered claims, of which 7 failed.

1 Introduction

Every post-training pipeline answers three questions it rarely asks out loud: which skills to train together, whether to run one blocked stage per skill or interleave them, and which skill goes first. The search space is $O(n!)$, each point costs a training run, and curriculum and interleaving studies report that order matters, does not matter, or matters in reverse (Bengio et al., 2009; Rohrer & Taylor, 2007; Kornell & Bjork, 2008; Zhang et al., 2025; Jia et al., 2025; Elgaar & Amiri, 2026). Instead of one more measurement, we start from the setting where the answer is exact.

The recipe search. Pin the corpus and the per-run compute, give a team unlimited patience, and let them train every order, every arrangement from fully blocked to fully interleaved, every composition, keeping the best. Two quantities decide what that search was worth, one a property of the system and one of the instrument, neither estimated from the other: search buys something only when the reachable set is wider than the resolution. Below that the ranking a team recovers is its own noise, and no further search repairs it — the search is not underpowered but unresolvable, with a signature not an absence of winners but winners that do not replicate. §3 defines both quantities and proves the signature; §5 measures both sides. Every number here comes from base (non-instruct) checkpoints, full-parameter

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.

SFT under AdamW, exact-match scoring at k=4, and competition-mathematics skills with single-valued answers; §6 states the scope this buys.

The idealisation. Consider a student given two textbooks under three idealisations: patience (re-read until nothing changes), coherence (the books never disagree), and room (never forget one thing to hold another). In that limit every recipe question dissolves — order, arrangement and composition are exactly free, and only the page count of the pile is not. Real fine-tuning is not that limit, so the reachable set $\mathcal{R}$ is a set, and the paper's first task is to measure its span against the resolution δ. Only one of the two is negotiable: Δ/σ_c decides Proposition 1 information-theoretically and no search strategy crosses it, whereas δ belongs to the decision procedure, so anything lowering σ_c moves the wall. §5 breaks each idealisation on purpose and asks which enlarges $\mathcal{R}$ past the wall. One does. A linear learner trained by gradient descent in the lazy regime (Jacot et al., 2018; Chizat et al., 2019) makes each idealisation a hypothesis and each "dissolves" a theorem (App. D): a measured recipe effect is never a primitive, it is a measurement of which premise of the free-recipe limit broke.

What we found, in the order it survived its controls. Disagreement inside one corpus moves a quantity accuracy cannot see. Across two skills, arrangement moves capability: revisiting the first skill beats not revisiting it by +0.0483 at t=8.06 in 17 of 18 cells, 78% of that being the first skill overwritten (forgetting, named as such); a replay arm at matched volume recovers all of that first-skill term with no interleaving at all (+0.0430 against +0.0377, 18/18), which is what licenses stating it that way rather than as a fact about alternation. Inside a single distribution the story changes: halve one pooled corpus at random and let the halves write answers under incompatible conventions, and capability stays inside a floor while the committed share moves two orders of magnitude above its own same-convention control, in a direction that control cannot produce; writing the convention into the input returns both to zero (§5.1–5.2). Patience breaks into a transient, not a signature. The order-carrying term grows one power of the budget faster than the symmetric one, so the sign must move, and it does, twice outside the noise band inside a single run; the grid's largest order contrast clears the resolution and then reverses on replication (§5.3). Room we could motivate and could not demonstrate: the predicted sign change appears along a capacity axis and its own volume-matched control removes it (§5.4).

Contributions.

1. A separability pre-check that costs no training (§5.2): train on convention A alone, require $\mathrm{acc}_A/\mathrm{acc}_B \geq 2$ before reading any mixed arm. Every published row here passes at 16.6× or better.
2. A noise budget showing which half of an SFT ablation's noise is cheap, and that the split must be measured on the quantity being resolved: borrowed from single arms it is 0.861, measured on the contrast it is 0.955 (Table 6, §5.3).
3. Base rates for convention disagreement in real corpora: two orders of magnitude rarer between corpus pairs we treated as independent than between two derivation pipelines of one curator, and every pipeline disagreement we found is exact form against decimal, invisible to exact match (§5.2).
4. A seven-item checklist assembled from the places this paper's own instrument failed and had to be repaired (§6).
5. Two standing recommendations: check a read-out's coordinate against the mechanism it is supposed to move, reporting the parts of any sum; and report allocation beside accuracy whenever two sources could label one input differently.
6. The wall itself, measured on both sides (Proposition 1, §5.3), the switch that clears it (§5.2), and a public scorecard grading all 30 registered claims, seven of which failed.

Four of the six are instruments for measuring post-training ablations rather than findings about arrangement, and the three premises did not break equally: patience is a replicated negative result, room is graded mixed and largely explained by a volume confound, and coherence is the one positive finding, on conflicts we constructed. The single takeaway we

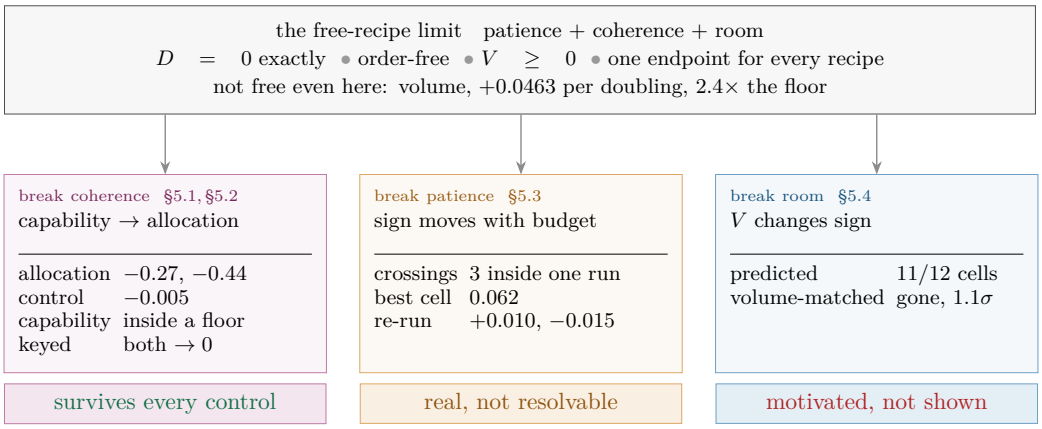


Figure 1. One limit, three ways out of it, three different fates. Each column breaks one premise on purpose; its table is the measurement, its last row the verdict its own controls returned. Allocation is quoted against its own same-convention control, capability and the floors in units of the 0.0191 contrast floor; 0.062 is the grid's largest order contrast, beside what the identical protocol returned on re-running.

would ask a reader to keep is item 2: most of the variance these ablations fight is bought down at the evaluation rate, and almost nobody is buying it there.

2 Related Work

Sequential training and forgetting. Our headline half, that blocking loses capability to the skill it trains first, is catastrophic interference (McCloskey & Cohen, 1989; French, 1999), forty years old, with mitigations (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017), an alternating-projection analysis that is our own mechanism (Evron et al., 2022), and a non-monotone dependence on task similarity (Lee et al., 2021). Reporting +0.0483 and stopping would be a rediscovery. A registered replay arm (§5.1) instead puts our positive result inside the rehearsal literature (Robins, 1995; Lopez-Paz & Ranzato, 2017; Rolnick et al., 2019; Chaudhry et al., 2019): Ibrahim et al. (2024) sweep replay from 0.5% to 50% of a fixed budget in continual pre-training and recommend 5%, matching compute by substitution as we do; our increment is comparing against an interleaved arm at equal budget, the arm that decides which of two names an arrangement effect deserves, and a per-skill accounting showing the recovery is a transfer between skills rather than the volume matching itself. Forgetting has no parameter for whether the two things disagree: it cannot predict that the capability term falls under resolution exactly when they do while a second term, which convention the model commits to, moves two orders of magnitude further, nor the key that returns both to zero. We tested the closest similarity-dependent account (Lee et al., 2021) directly: a domain-boundary row makes two sources maximally dissimilar with nothing to dispute and leaves arrangement inert at 0.5σ, while numeral-against-spelled makes them maximally similar and fires at 5.16 floors — disagreement is the axis, not dissimilarity.

Order prediction and the variance literature. That the commutator is the right object for order is Sweeney (2026a)'s, who predict pairwise order at 98.1% from a one-step Lie bracket; our contribution is a measurable domain of validity, since a one-step score presumes its sign holds over the horizon it is used on, and §5.3 shows the sign turns over three times inside a single run. One consequence is a tool: a reported pairwise accuracy A lower-bounds the reproducibility q of the reporter's own sign labels ($A = pq + (1-p)(1-q)$ maximised at $p = 1$); we measure q directly at 66.7% (App. F). A line from Reimers & Gurevych (2017); Melis et al. (2018); Henderson et al. (2018); Dodge et al. (2019) through Dodge et al. (2020) and Bouthillier et al. (2021) decomposes benchmark variance into sources cheaper and more expensive to buy down; Hochlehnert et al. (2025) finds competition-mathematics evaluation itself moving with decoding parameters and seeds and answers with a reporting standard. That work sets the standard; §3 prices it, adding the comparison that a search's diameter and the noise sit on one scale. Dai et al. (2026)'s data-reorganisation gains act at sample

level across a heterogeneous mixture, evaluated at single endpoints; ours are at stage level within one coherent domain at matched volume, with every winner re-run.

Disagreement, mixing, and the family. Dsouza et al. (2025); He et al. (2026); Xue et al. (2026) treat disagreement as a defect to find, notice, or resolve; §5.2 instead holds it fixed and asks what arrangement of the same rows does. Closest in shape is the proportion-varying knowledge-conflict line (Gekhman et al., 2024); ours varies how much of a corpus contradicts itself and adds a learnability guard. Mixing work optimises proportions (Xie et al., 2023; Fan et al., 2024; Gu et al., 2025); at fixed proportions and volume the remaining recipe freedoms are inside noise unless the mixture disagrees with itself, and Proposition 1 prices the step a recipe search and an operator search both end with, certifying a winner (Sculley et al., 2018; Cooper et al., 2021; Bastani et al., 2026; Bui et al., 2025). The idealisation-first structure of this paper is the one introduced by the Capability Convergence Hypothesis (Anonymous, 2026), applied at the recipe level; Anonymous (2026) applies it at the path level. The conflict switch appears in both, and the split is stated so the overlap does not read as two confirmations of one thing: this paper owns its extent, four scales, two pretraining families, three convention constructions, a dose axis; Anonymous (2026) owns its mechanism, cited for provenance and never as evidence. App. A extends every grouping above, and App. I is the full owns-and-borrows accounting.

3 Method

Definition 1 (Recipe set, reachable set, resolution). Fix a corpus, a base model and a budget. A recipe r (order, arrangement, composition) is drawn from a finite set $\mathcal{S}$ of size m; training and evaluating gives $Y(r) = \mu(r) + \varepsilon_r$, ε_r independent mean-zero with s.d. σ. The reachable set is $\mathcal{R} = \{\mu(r)\}$, diameter $\Delta = \max_r \mu(r) - \min_r \mu(r)$. A team comparing two recipes at R runs each has contrast s.e. $\sigma_c = \sigma\sqrt{2/R}$ and resolution $\delta = 2.80\,\sigma_c$ (the 80%-power, two-sided-5% minimum detectable gap). R counts training runs; k, reserved for generations sampled per problem at evaluation, is bought at a rate that differs by orders of magnitude (§3).

Proposition 1 (The recipe-search wall). Let $\hat{r} = \arg\max_r Y(r)$, $\hat{\Delta} = Y(\hat{r}) - Y(\check{r})$ its margin over the observed worst, ε_r Gaussian. Then (i) $\mathbb{E}[\hat{\Delta}] \geq \max\{\Delta, d_2(m)\sigma\}$, $d_2(m)$ the expected range of m standard Gaussians, so a search over m recipes reports a margin bounded below by a noise term growing with m, even at $\Delta = 0$; and (ii) if $\Delta = 0$, an independent repetition re-selects $\hat{r}$ with probability exactly $1/m$ and preserves a fixed pair's sign with probability exactly $1/2$, both continuous in Δ/σ_c.

Proof in App. D. The signature is therefore not an absence of winners but winners that do not replicate: a search below the wall still returns a best recipe, with a margin that looks larger the harder one searched, and re-running returns a different one. At the corpus level (m=2 directions) $d_2(2) = 1.128$; the grid level (m=6 pairs) has $d_2(6) = 2.534$, and under $\Delta = 0$ at the independently measured run-mean dispersion the six pairs' observed span sits at $z = 1.10$ ($p = 0.281$): consistent with noise (App. J).

The noise model and resolution ratio. Writing $y_{prs} = \mu_p + u_{pr} + e_{prs}$ for pair p, execution r, seed s, with protocol-level $u \sim (0, \sigma^2_{\text{run}})$ and seed-level $e \sim (0, \sigma^2_{\text{seed}})$, the seed term splits further into a closed-form generation component and a training remainder, $\sigma^2_{\text{seed}}(k) = \sigma^2_{\text{eval}} k_0/k + \sigma^2_{\text{train}}$, giving a resolution schedule $\delta_{\text{run}}(k) = 2.8\sqrt{2}\sqrt{\sigma^2_{\text{seed}}(k)/n + \sigma^2_{\text{run}}}$ that is strictly decreasing in k with a floor set by σ_{run} (App. D states and proves this as Proposition 2). For an effect E at resolution δ, write $\rho = E/\delta$; a plug-in ρ using the largest of m observed contrasts is biased upward (Proposition 3), so a reported $\rho < 1$ understates how far below the wall a search sits. Comparing two departures' ratios ρ_1, ρ_2 over unrelated protocol components $\sigma_1 \in [0, a_1]$, $\sigma_2 \in [0, a_2]$ reduces to one corner: $\inf(\rho_2 - \rho_1) = \rho_2(a_2) - \rho_1(0)$ (Proposition 4), which §5.3 evaluates at 0.98 against 0.82.

Predictions at finite budget. Two extensions turn the limit into falsifiable signs. At finite budget T each projection becomes the spectral filter $M_X(T) = (I - \eta\Sigma_X)^T$, and the

Table 1. Instrument calibration: reliability of each observable (ICC(1), one-way ANOVA over training seeds; 187 triples for V/D, 105 for order; natural corpus sizes). Composition is the only recipe contrast whose structure reliably survives reseeding; arrangement carries no significant reliable variance. This is what the observables can measure, reported before what they did measure.

observable	ICC(1)	F	$F_{.05}$	$\hat{\sigma}_{\text{within}}$	$\hat{\sigma}_{\text{between}}$	signif.
V (co-training value)	0.803	14.63	2.00	0.0232	0.0469	yes
FWT $= V - D$	0.820	16.20	2.00	0.0207	0.0442	yes
ASYM (directional order)	0.524	4.62	2.53	0.0268	0.0281	yes
RECENCY(B)	0.481	4.04	2.53	0.0187	0.0180	yes
RECENCY(A)	0.295	2.37	2.53	0.0204	0.0132	no
D (arrangement penalty)	0.197	1.82	2.00	0.0202	0.0100	no

arrangement operator $\Delta(T) = M_B M_A - M_{A\cup B}^2$ expands (small η, equal sample sizes) into

$$\Delta(T) \;\approx\; \underbrace{-\tfrac{\eta^2 T}{4}\,(\Sigma_A - \Sigma_B)^2}_{\text{symmetric, } O(T)} \;+\; \underbrace{\tfrac{\eta^2 T^2}{2}\,[\Sigma_B, \Sigma_A]}_{\text{order-carrying, } O(T^2)} \quad , \tag{1}$$

so the order-carrying term grows one power of T faster than the symmetric one and an effect set by their competition must change sign along the budget axis. Second, if coherence fails ($\kappa_A \neq \kappa_B$), the interleaved arm converges to a sample-weighted compromise on the contested directions while the blocked arm's final stage takes them outright:

$$D_{\text{bulk}} = -\left\| \Sigma_B^{1/2} P_{\text{con}} \tfrac{n_A}{n_A+n_B} (\kappa_B - \kappa_A) \right\|^2 < 0, \tag{2}$$

so under disagreement, blocked training must win. Both are sign and existence predictions on a first-order lazy account; full SFT is measurably not lazy and the optimiser is not memoryless (App. D), so Eqs. (1)–(2) are formalised intuition with an exact domain, not quantitative models of the experiment.

4 Experimental Setup

Twelve base models from three families (Qwen3 1.7–14B, Qwen2.5 0.5–14B, Llama-3 1–8B), full-parameter SFT under AdamW on competition-mathematics skill slices from OpenR1-Math (Hugging Face, 2025; Hendrycks et al., 2021; Qwen Team, 2025; 2024; Grattafiori et al., 2024), evaluation k=4 on 120 held-out problems per skill (Kwon et al., 2023), exact match on the boxed answer, avg@4 and not pass@4. Skills confound order with volume unless matched: natural sizes (962 algebra rows against 622 calculus) reversed our order finding once; every headline contrast below is matched to 622 rows per stage. Reliability (ICC(1) over training seeds, 187 triples) ranks composition 0.803, order 0.524, arrangement 0.197 (F=1.82 against $F_{.05}$=2.00, no significant reliable variance) at natural sizes; under matching, order's reliability moves with the budget from 0.860 to 0.149, so no single ranking is a property of the instrument independent of corpus and budget (Table 1). The grid analysed is 761 fine-tuning runs and 2158 scored evaluations, drawn from a released base of 1233 runs and 3274 evaluations (App. F states the counting rule).

5 Results

5.1 Arrangement at fixed volume: forgetting, not free, and what recovers it

The limit predicts all three recipe decisions sit at zero within a coherent domain at fixed volume. Scored on the coordinate each contrast was defined on, that is what we see: composition 0.0101, arrangement 0.0140, order 0.0205 against a 0.0191 floor (Table 3), none resolvable. That reading does not survive changing the coordinate. Scored on both skills an arm was trained on — the coordinate a practitioner receives — interleaving beats blocking by +0.0483 at t=8.06 in 17 of 18 (pair, seed) cells (Table 2, top row), with 78% of it sitting on the skill trained first and then overwritten (t=11.45, 18/18) and only 22% on the skill D was defined to score (t=2.96, 12/18). This is catastrophic interference (McCloskey & Cohen,

Table 2. The load-bearing comparison never leaves one distribution. Row 1 trains two different skills and calibrates the ratios; rows 2–5 are random halves of one pooled corpus differing only in whether the halves disagree and whether a write-time key is present. Both contrasts are the shuffled arm minus the mean of the blocked arms; intervals are two-sided 95% t. Qwen2.5-7B, learning budget.

		capability		allocation	
corpus	n	contrast	95% CI, floors	ΔSHARE	× coherent
coherent — four maths skills, six pairs	18	**+0.0483**	[**+1.87**, **+3.19**]	+0.0032	1×
conflicted — numeral against spelled	8	+0.0016	[−0.61, +0.78]	**−0.2677**	**84**×
conflicted — synthetic +1 on half the answers	8	−0.0052	[−1.77, +1.22]	**−0.4431**	**139**×
the same conflict, keyed at write time	8	−0.0003	[−0.63, +0.60]	−0.0008	0.25×
control — both halves numerals, no conflict	8	−0.0099	[−1.40, +0.37]	−0.0051	1.6×

1989). And we can now say which variable it is. Every arm that revisits the first skill also interleaves, so a registered replay arm, volume-matched by substitution (stage two carries exactly 622 rows, 156 of skill one and 466 of skill two), separates them: replay recovers **+0.0369** of the gap pooled (76.5%, interval [0.27, 1.25] once the extra protocol-level offset of a separately trained comparator is carried through), but per skill it recovers all of the first-skill term (+0.0430 against +0.0377, 18/18, 114%) while costing the diluted second skill −0.0061 (−57% of its part). Interleaving was our name for the arm and not the variable: revisiting the skill trained first is what a coherent domain rewards, and the result and the abstract are renamed accordingly. A pooled sum over two skills moving in opposite directions is, by this section's own argument, the wrong coordinate to trust, which is the first of four coordinate errors this paper's own instrument made (§6 lists all four).

Allocation as an observable. What carries this paper's positive result is not an accuracy but SHARE $= \mathrm{acc}_B/(\mathrm{acc}_A + \mathrm{acc}_B)$, free once both conventions are scored and orthogonal to accuracy by construction. Held fixed, the pooled corpus, volume, budget and scoring, and changed only whether its two halves demand incompatible outputs, allocation moves −0.2677 and −0.4431 against a same-distribution control's −0.0051 (84× and 139×), while capability stays inside a floor throughout (0.09–0.52 floors, both conflicted intervals containing zero); writing the convention into the input returns allocation to 0.25× the coherent anchor and capability to 0.02 floors (Table 2). Forgetting has no term that switches on when two sources disagree and nothing to say about which convention a model commits to at fixed total accuracy; it also cannot be a confound here, since the tempting cross-distribution test is itself confounded (coherence and content change together), so the load-bearing comparison never leaves one distribution.

5.2 Breaking coherence: disagreement is a switch

Every corpus below splits one pool into two halves of matched size, holding rows and budget identical by construction; a separability precondition, $\mathrm{acc}_A/\mathrm{acc}_B \geq 2$ on a model trained on convention A alone, is required before reading any mixed arm, since a learnability guard is computed with the same scorer whose blindness is the risk. Every published row passes at 16.6× or better; every same-convention control fails it. On numeral against spelled (both correct) $D = -0.0985$ at 5.16 floors, all eight seeds negative, replicated twice more from independent split seeds (−0.1258, −0.0825); Roman against Arabic gives −0.1180 at 6.18 floors on five seeds and replicates on a disjoint held-out draw (−0.1494, five of five negative); the synthetic +1 ceiling gives −0.2346 at 12.3 floors; at 14B the effect grows to 10.67σ and the switch replicates cross-family on Qwen3-8B (−0.0875, 4.58 floors, five of five). Neither difficulty alone (both halves spelled, +0.0012) nor domain distance alone (algebra + physics, +0.0096) moves anything: disagreement is the axis, not difficulty or distance. Four discriminating controls rule out a bare recency reading: the effect is absent at every timescale below the stage (a 27-fold block-length sweep sits at the interleaved arm's own allocation until one full pass per epoch); the two difficulty/distance nulls share the design and not the result; the shuffled arm does not collapse to a coin flip; and a write-time key, changing no schedule, moves ΔSHARE from −0.2677 to −0.0008 while leaving the order itself untouched. Re-scored at the better-of-two-conventions per problem, blocking's win

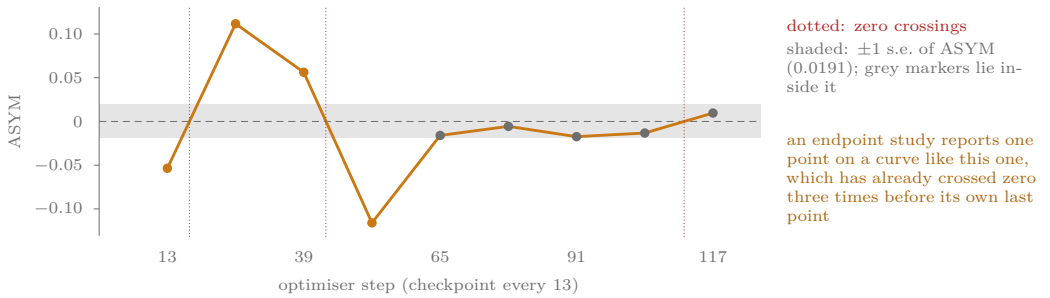


Figure 2. The order effect is a function of the budget, not a scalar. ASYM for algebra–combinatorics at every 13th optimiser step of both stage-2 directions (Qwen2.5-7B, 622 rows per stage, 3 epochs, lr 3×10^{-5}; the axis spans the entire stage-2 run), with zero crossings interpolated at steps 17, 43 and 112. This pair is an example, not the evidence base: all four pairs measured this way change sign inside a single run, and in every pair at least one crossing is made by each seed separately (Table 15), here those at 17 and 43, both 3/3, while the in-band crossing at 112 is 1/3.

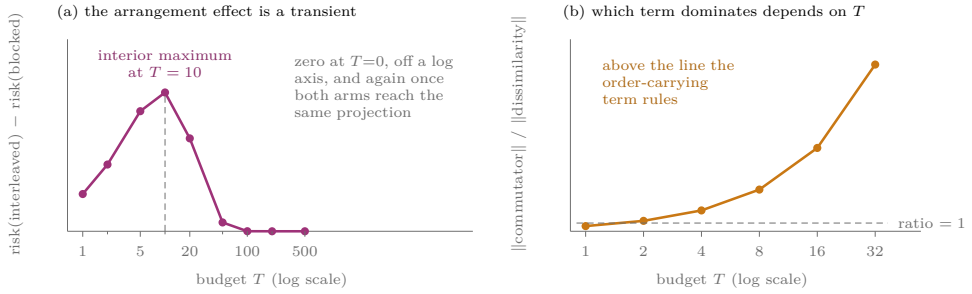


Figure 3. What breaking patience does, computed rather than assumed. Both panels run gradient descent step by step in the linear model. (a) The blocked-versus-interleaved gap vanishes at both ends of the budget axis and peaks in between: a mid-training phenomenon, which is why endpoint-only designs report it inconsistently. (b) The order-carrying commutator term (§3) grows one power of T faster than the symmetric term, so the balance between them, and with it the sign of the order effect, is a function of the budget.

at 7B vanishes (−0.0084, +0.0308, both inside two floors); it does not vanish at 14B or under the synthetic conflict, which App. E shows is per-problem sample-splitting under a metric that cannot see it, not solving fewer problems. What arrangement moves under conflict is therefore mostly which convention the model commits to: on the synthetic corpus $\text{acc}_A + \text{acc}_B$ spans only 2.7 floors across twelve arms while the B-share runs $0.04 \to 0.41 \to 0.87$.

Base rates. Every corpus above was made to disagree by us. Matching problems by text across 164,984 shared public-corpus pairs, proved convention conflict is 9 in 92,359 and 12 in 7,169 between pairs we treated as independently labelled, against 1,284 in 65,456 between two derivation pipelines of one curator (Table 10, App. A) — and every one of those 1,284 is exact form against decimal, invisible to exact match. A concurrent re-survey of the same pairs (Anonymous, 2026a) finds the larger "independent" pair is one lineage (one corpus's gold field is filtered through the other's canonicaliser) and widens the genuinely external pair's proved rate by an order of magnitude (26.7–29.2% form against 2.06% value), strengthening rather than undercutting the phenomenon this paper constructs its conflicts to study. An attempt to find disagreement in the wild without constructing it returned an instrument failure rather than a row: the guard passed at 20× the floor while 250 of 250 evaluation gold pairs were mathematically equal under our own scorer, which the separability precondition above now catches (App. G).

5.3 Breaking patience: the order effect is a transient

The two-term expansion of §3 predicts a sign that must move with budget. Measured directly (checkpoints every 13 steps through both directions of algebra–combinatorics, three seeds), ASYM crosses zero three times inside a single run, two of them evidential ($+0.112$, -0.116, about 6σ); all four measured pairs change sign within a run. At the endpoint, the grid's largest single-corpus contrast is $|\text{ASYM}| = 0.0619$ (algebra–combinatorics), clearing a three-seed resolution of 0.054 but not a re-run resolution of $\delta_{\text{run}} = 0.081$ (95% CI $[0.072, 0.119]$, fitted on 39 (pair, run, seed) cells); re-run twice more under an identical protocol, that cell reads $+0.0096$ and -0.0146 — a reversal, not merely a shrinkage. This is Proposition 1's signature exactly. The run component net of seed noise is $\hat{\sigma}_{\text{run}} = 0.0072$ on seven degrees of freedom, interval $[0, 0.025]$ containing zero (a first estimate on the one cell selected for failing to replicate, 0.0364, is withdrawn as biased and superseded, App. J). The evaluation share of ASYM's seed variance, measured directly rather than borrowed from single-arm cells, is 0.955 (not 0.861), so at the point estimate k=16 crosses $\operatorname{diam}\mathcal{R}$ for minutes of inference; at the interval's upper end no k, including $k \to \infty$, does. The split is not identified because the draws are not independent: no evaluation driver in this project passes an engine seed, so the shared stream inflates the estimated share above one in 232 of 433 cells (Table 6), a seventh failed registered claim, and the borrowed 0.861 measured on a different corpus reads 0.3213, a factor of 2.7 — the recommendation to raise k is a statement about this corpus and not a constant.

The switch was independently re-executed under the same nested model: $\delta^{\text{switch}}_{\text{run}} = 0.080$ ($\hat{\sigma}_{\text{run}} = 0.0152$ on two degrees of freedom, interval $[0, 0.0228]$), giving the paper's two ratios, $\Delta/\delta_{\text{run}} = 0.0619/0.081 = \mathbf{0.76}$ and $|\bar{D}|/\delta^{\text{switch}}_{\text{run}} = 0.102/0.080 = \mathbf{1.28}$. The recipe ratio's supremum over every possible $\sigma_{\text{run}} \geq 0$ is 0.816, so the recipe search is below its own resolution at every value of the unidentified component, not merely across an interval (Corollary 1); the switch ratio runs $1.96 \to 1.28 \to 0.98$ across its own interval, so at the top of its own interval the switch does not clear its resolution either, which we report against our own positive result. Proposition 4's corner check gives 0.98 against 0.82: the switch is better resolved than the recipe axes at every pair of values the two unrelated components can take. Carrying $\hat{\sigma}_{\text{seed}}$'s own interval and undoing the numerator's winner's-curse selection bias, a full bootstrap gives the recipe ratio 95% CI $[0.25, 1.21]$ with $\Pr(\rho{>}1) = 0.091$ (falling to 0.028 once selection is undone), and the ordering of the two ratios holds in 95.5% of resamples — weaker than "settled" and the honest form (App. J).

5.4 Breaking room: motivated, not demonstrated

Varying LoRA rank $r \in \{4, \ldots, 128\}$ at fixed data, the predicted sign change of the composition value $V(r)$ appears, monotone, in 11/12 (pair, rank) cells across three seeds — and does not survive its own controls. Two thirds of the change is the single-skill baseline collapsing (overfitting, not capacity); a rank×learning-rate grid shows the learning-rate axis moves V about as much as rank; and a volume-matched re-run (both arms 622 rows at every rank) removes the crossing entirely for the pair that carried it, leaving the other pair one cell at 1.1σ, below the design's own MDE_{80} gate. The theory's predicted crossing location was excluded with the direction reversed (ratio 0.382, 95% CI $[0.231, 0.583]$ against a predicted 1.0–1.27). Existence of the rank-axis sign change is supported; its attribution to capacity is mixed. No scalar property of the base model (parameters, competence, encoded density) orders the sign of V consistently across Qwen2.5 and Qwen3 (Simpson reversals throughout), which retires per-model scalar reasoning about co-training sign in this range.

6 Analysis and Limitations

What the results do not license. The geometry motivating composition > order > arrangement (matching the measured reliabilities 0.803>0.524>0.197) fails on the model in all 12 ordered skill pairs by 3.9–13.5×, at every span dimension and probe depth measured; the suppression premise is unreachable for any pair of subspaces derived from one base model, so the hierarchy is an empirical regularity the geometry motivates and does not license (App. D). Six registered claims failed — five location/magnitude claims from the theory

(existence and sign survived; location did not, in every case tested) and this paper's own name for the coherent-domain result, refuted by the replay arm built to test it and registered with that consequence in advance. The public scorecard grades all 30 claims: 19 supported, 7 failed, 2 untested, 1 mixed, 1 void (App. J). The two conflicted capability terms both read zero at eight seeds and cost us a published claim on extension (App. J); the coherent row is not a control for the conflicted ones, since moving between them changes coherence and content simultaneously, so a capability term at the floor is evidence for neither account and what survives, the write-time key inside one distribution, is why we believe the switch rather than merely observe the cross-regime contrast. Ten pre-registration criteria were frozen before the data that judge them existed; eight are read out and every branch that fired is reported, including the ones that cost a claim (App. J). The paper reports 145 quantitative read-outs and applies no family-wise correction; the discovery claims sit at 3.9–12.3σ with nineteen of nineteen conflict seeds negative across four scales and two families, and survive Bonferroni over all 145 at $\alpha = 0.05$ ($z \approx 3.58$; smallest firing row 3.86σ).

Scope and missing cells. Every wall measurement is at 622 rows per stage against production post-training's 10^4–10^7, and our own volume result argues against us: doubling a skill's rows buys +0.0463 (2.4× the floor, Table 3), so a small $\operatorname{diam}\mathcal{R}$ at these scales is consistent with either arrangement not mattering or both arms being far from saturation, and nothing here separates them. Six missing cells remain, headed by pinning σ_{run} (two further executions, no new corpus, taking both sides' verdicts from seven degrees of freedom to nineteen) and a second 120-problem draw for $\operatorname{diam}\mathcal{R}$ specifically (36 two-stage trainings; the switch side's own second-draw replicate already returned, $D = -0.1494$, and cost no training). A seventh check, re-scoring one family with an independent engine seed per run, prices the instrument itself (App. H lists all seven with their specifications).

7 Conclusion

Recipe search is bounded by the reachable set's span against the resolution of the instrument that would rank it, a ratio of 0.76 whose denominator turns on a run component this design bounds and does not identify; the failure is not that effects are absent but that they are unresolvable, with a specific signature, the grid's best cell clearing an optimistic threshold and then reversing sign on identical re-execution. That is the story for order and not for arrangement, where the difference is a coordinate: scored on the skill a contrast targets, arrangement sits at 0.0140 against the floor; scored on both skills it moves +0.0483 at t=8.06, still roughly half of what doubling one skill's data buys. What arrangement has that volume does not is a switch: internal disagreement, not difficulty and not domain distance, changes which quantity arrangement moves at all, from capability to allocation, by two orders of magnitude, and writing the convention into the input turns it off. The widest-reach single result is the noise budget: at k=4, most of the variance an SFT ablation fights is evaluation sampling (0.861 on single arms, 0.955 on the order contrast itself), raising k from 4 to 16 costs minutes against a training seed's hours, and the one condition that travels with the recommendation, an independently seeded evaluation engine, is this paper's own seventh failed claim.

The screening question. Would the two corpora ever label the same input differently? If no, spend the budget on volume, not arrangement. If yes and the conflict is resolvable at write time, resolve it there — keying the convention into the input or cleaning the disagreement returned both terms to zero here and no schedule matched it. If yes and it cannot be resolved, arrangement is the operative lever, and it acts on allocation rather than capability. Our own base rates argue against expecting much of this in practice within one curator's lineage, and for expecting more of it once a stronger grader is used on genuinely external pairs (§5.2); the question is validated on constructed conflicts in mathematics SFT and offered as a procedure worth testing elsewhere rather than as advice that transfers. Every checklist item in §6's seven-item list is a repair to this paper's own instrument, so none of it is advice we have not paid for.

Reproducibility statement

Every number in this paper is written by a named analyzer to a verdict file; every registered read-out's rule was frozen in a job script before the run it decided, and Appendix J tabulates every criterion against when it fired. Appendix J is the public scorecard, one row per registered claim, generated from the result files by a ledger script. Appendix J states the counting rule behind every quantitative read-out in the paper and Appendix J lists every reading printed and withdrawn across revisions, so a reader meeting an earlier number elsewhere can identify it. Pre-registrations, analysis code, every per-run result file with its per-problem score vectors, and every derived verdict are included in the anonymised supplementary material; MANIFEST.md there maps every artefact in this paper to the verdict and analyzer that produced it, so any number can be recomputed with numpy and scipy and no GPU. Model checkpoints (the set is 9.4 TB) are not included; no number in this paper depends on them.

Ethics statement

All corpora are derived from public competition-mathematics benchmarks (Hendrycks et al., 2021; Hugging Face, 2025); no new human data was collected, and every "conflict" corpus is a deterministic, disclosed relabelling of existing problems (numeral/spelled, Arabic/Roman, or a synthetic +1 shift), not a claim about any external corpus's correctness. The public-corpus base-rate survey (§5.2) compares existing gold answers by matched problem text and reports aggregate counts only; it does not identify or evaluate any individual annotator, curator, or organisation beyond citing the corpora by name. All models trained are openly released base checkpoints; we introduce no new capability, only a measurement of how training order and arrangement interact with pre-existing disagreement in supervision. We are not aware of a Code of Ethics concern raised by this work.

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 J tabulates. The one transformation AI-written code applied to a released artefact — 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 artefacts.

References

Anonymous. The capability convergence hypothesis: Capability from access structure, not scale. arXiv preprint, 2026. author list and repository withheld for double-blind review; a concurrent manuscript of the same research programme.

Anonymous. The infinite-division limit: Below corpus scale, training writes only the time-average of its data path. Manuscript under concurrent review; cited for provenance only, 2026.

Anonymous. Which corpus supplies the gold: A free variable in exact-match evaluation. concurrent manuscript; author list and repository withheld for double-blind review, 2026a.

Anonymous. The underdetermined stream: What a single-gold score measures, and how to recover capability from it. concurrent manuscript; author list and repository withheld for double-blind review, 2026b.

Hamsa Bastani, Osbert Bastani, and Bryce McLaughlin. Winner's curse drives false promises in data-driven decisions: A case study in refugee matching. arXiv preprint, 2026. arXiv:2602.08892.

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

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In Proceedings of Machine Learning and Systems (MLSys), 2021.

Matthias Brunmair and Tobias Richter. Similarity matters: A meta-analysis of interleaved learning and its moderators. Psychological Bulletin, 145(11):1029–1052, 2019.

Nghia Tuan Bui, Guergana K. Savova, and Lijing Wang. Assessing the macro and micro effects of random seeds on fine-tuning large language models. In Proceedings of IJCNLP (short), 2025. arXiv:2503.07329.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. arXiv preprint, 2019. arXiv:1902.10486.

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

A. Feder Cooper, Yucheng Lu, Jessica Zosa Forde, and Christopher De Sa. Hyperparameter optimization is deceiving us, and how to stop it. In Advances in Neural Information Processing Systems (NeurIPS), 2021.

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.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In Proceedings of EMNLP-IJCNLP, 2019.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305, 2020.

R. Dsouza et al. Sources of disagreement in data for LLM instruction tuning. In Proceedings of the CoMeDi Workshop at ACL, 2025. ACL Anthology 2025.comedi-1.3.

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

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.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In Proceedings of EMNLP, 2024. arXiv:2405.05904.

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

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

Xingwei He, Qianru Zhang, Pengfei Chen, Guanhua Chen, Linlin Yu, Yuan Yuan, and Siu-Ming Yiu. ConInstruct: Evaluating large language models on conflict detection and resolution in instructions. In Proceedings of AAAI, 2026. arXiv:2511.14342.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In AAAI Conference on Artificial Intelligence, 2018.

Dan Hendrycks, Collin Burns, Saurav Kadavath, et al. Measuring mathematical problem solving with the MATH dataset. NeurIPS Datasets and Benchmarks, 2021. arXiv:2103.03874.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. In Conference on Language Modeling (COLM), 2025. arXiv:2504.07086.

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

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. Transactions on Machine Learning Research (TMLR), 2024. arXiv:2403.08763.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In International Conference on Learning Representations (ICLR), 2023.

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.

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.

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

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

Suhas Kotha and Percy Liang. Replaying pre-training data improves fine-tuning. arXiv preprint, 2026. arXiv:2603.04964.

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.

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

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In International Conference on Machine Learning (ICML), 2021.

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.

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.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In International Conference on Learning Representations (ICLR), 2018.

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.

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

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.

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

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In Proceedings of EMNLP, 2017.

Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. Connection Science, 7(2):123–146, 1995.

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

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

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.

D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? On pace, progress, and empirical rigor. In ICLR Workshop Track, 2018.

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.

Shanyong Wang, Shuhang Lin, Yining Zhao, Xi Zhu, and Yongfeng Zhang. Meet dynamic individual preferences: Resolving conflicting human value with paired fine-tuning. arXiv preprint, 2026. arXiv:2604.12479.

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.

Zhenyu Xue et al. Why supervised fine-tuning fails to learn: A systematic study of incomplete learning in large language models. In Proceedings of ACL, 2026. arXiv:2604.10079; ACL Anthology 2026.acl-long.1393.

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.

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, the base-rate survey, the per-skill order potential, and the wild-conflict episode. Reviewers are not required to read it.

A Extended Related Work

Rehearsal, in full. A registered replay arm (§5.1) puts the coherent-domain result inside the rehearsal literature: named by Robins (1995), carried into gradient-episodic and generative-replay methods (Lopez-Paz & Ranzato, 2017; Rolnick et al., 2019), shown to work from very small memories (Chaudhry et al., 2019), and measured for LLMs by Ibrahim et al. (2024), who sweep replay from 0.5% to 50% of a fixed compute budget in continual pre-training and recommend 5%. Their comparison is sequential-with-replay against no replay and against joint training on the union; ours is against an interleaved arm at equal budget, the arm that decides which of two names an arrangement effect deserves and which that line does not run. Concurrently Kotha & Liang (2026) find replay of generic pre-training data improving the target task rather than only protecting the generic one, by up to 1.87× in data efficiency; that is the same direction our replay arm reports inside one domain, reached from the pre-training side and with the cross-domain variant we do not run.

Curriculum and interleaving, in full. The divided empirical record (Bengio et al., 2009; Rohrer & Taylor, 2007; Kornell & Bjork, 2008; Zhang et al., 2025; Jia et al., 2025; Elgaar & Amiri, 2026; Say et al., 2025; Liu et al., 2026; Luo et al., 2025) is, on this account, what sampling a transient at one endpoint per study must produce; the missing covariate is the budget, not a better difficulty metric. A meta-analysis of human interleaved practice finds it generally beats blocking, with between-category similarity as the moderator (Brunmair & Richter, 2019); our coherent grid reproduces that direction, and our conflict rows show what the moderator becomes when similarity is pushed to contestation of the same inputs: the sign flips and the moved quantity changes.

Order prediction, in full. The projection algebra behind Theorem 1 is classical (Evron et al., 2022); new here is the comparison across arms, where the blocked-versus-interleaved difference is an identity rather than a bound and the conflict term (§3) is a switch with a testable sign. The same author's optimiser-memory result (Sweeney, 2026b) supplies at once the sharpest alternative reading of the order transient and the validation of the memoryless mechanism, engaged as a confound in App. F and §5.2's block-length sweep. Concurrent perturbation-window work independently finds order sensitivity that exists but resists localisation (Piontkovskaia & Nikolenko, 2026), and influence itself is reported to flip sign during training (Pruthi et al., 2020; Lee et al., 2025).

The variance literature, in full. A line running from Reimers & Gurevych (2017) and Melis et al. (2018) through Henderson et al. (2018) and Dodge et al. (2019) established that single-number reporting in NLP and RL hides run-to-run spread large enough to reverse published rankings, and Dodge et al. (2020) showed the specific case this paper works in, that fine-tuning outcomes move with weight initialisation and data order alone. Closest of all is Bouthillier et al. (2021), which decomposes benchmark variance into its sources and prices what each buys: Table 6 is that exercise on a post-training grid, and the two conclusions agree that most controllable variance is cheaper to buy down at evaluation than at training. Closest in domain is Hochlehnert et al. (2025), which finds competition-mathematics evaluation to move with decoding parameters, random seeds, prompt formatting and even hardware, and answers with a standardised protocol and a reporting standard. That work sets the standard; this one prices it. A standard says run several seeds and report the spread; it does not say which of the two draws a reader is spending on, and Table 6 finds at k=4 that they are not close. What the wall adds is not the observation that variance is underreported but a comparison: the diameter of the reachable set is measured on the same scale as the noise, so the question stops being how large the error bars are and becomes whether anything is far enough apart to rank.

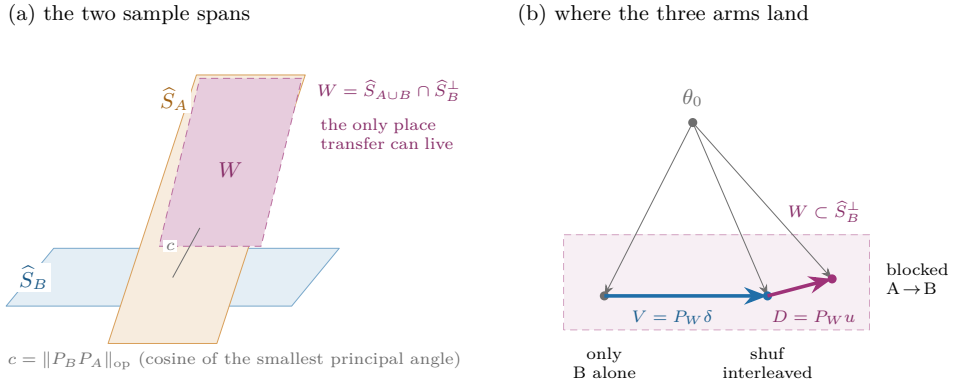
Model merging and task arithmetic (Ilharco et al., 2023) sit outside the recipe set entirely and are the natural alternative when arrangement turns out not to be a lever: they compose skills after training rather than sequencing them during it, so they are not bounded by $\operatorname{diam}\mathcal{R}$ as defined here. Dai et al. (2026) report that reorganising existing training data improves stability and performance, which reads as the opposite of our arrangement null and is not: their interventions act at sample level across a large heterogeneous mixture and are evaluated as single endpoints, ours at stage level within one coherent domain at matched per-source volume, with each winner re-run.

Disagreement, in full. Dsouza et al. (2025) catalogue where instruction-tuning labels disagree; He et al. (2026) ask whether a model notices conflicting constraints at inference time; Xue et al. (2026) and Wang et al. (2026) take conflict as something to be resolved. All three treat disagreement as a defect to be found, flagged or repaired. §5.2 asks the question none of them does: holding the disagreement fixed and repairing nothing, what does the arrangement of the same rows do? Because the answer is that it changes which quantity arrangement moves, we read allocation and capability separately rather than reporting one accuracy. Per-step gradient-conflict methods (Yu et al., 2020; Liu et al., 2021; Navon et al., 2022) resolve contested directions inside the optimiser step; our conflict lives at the stage timescale, and the separating cell of §5.2 shows none of the stage-level effect is batch-level.

The family, in full. The idealisation-first structure of this paper is the one introduced by the Capability Convergence Hypothesis (Anonymous, 2026), applied at the recipe level; Anonymous (2026) applies it at the path level. The conflict switch appears in both, and the split is stated so the overlap does not read as two confirmations of one thing. This paper owns its extent, measured here and nowhere else: four scales in two pretraining families, three convention constructions, a dose axis, matched controls at every point, and the nulls showing the trigger is disagreement rather than heterogeneity. Anonymous (2026) owns its mechanism, which we cite for provenance and never as evidence: no claim, control or number in this paper depends on it, and if its averaging limit were wrong the rows in Table 9 would stand and lose only their explanation.

B Supporting Measurements Referenced from the Main Text

The floats below carry the full record behind contrasts the main text states in prose. Nothing here changes a number quoted in §5; each caption is the one attached to the same table or figure in the full manuscript.



Both contrasts live in the same subspace W and are gated by the same unexplored mass ε_B, but D pays two factors V does not: it sees only A's unlearned residual $u = (I - P_A)\delta$, and that residual is orthogonal to $\widehat{S}_A$, so re-entering W costs a further factor c. Lengths are schematic: at matched volume the measured $|D|/|V|$ is 1.39, i.e. the suppression this panel illustrates is not what the data shows (§5.1, §6).

Figure 4. Why arrangement is doubly suppressed in the limit. (a) Each skill's gradient updates are trapped in the span of its own training inputs, so the only directions through which A can change B's population risk are those A supplies outside B's span, $W = \widehat{S}_{A\cup B} \cap \widehat{S}_B^{\perp}$ (dashed). (b) Composition value V rides on the full leakage $P_W\delta$, while the arrangement contrast D rides on $P_W(I - P_A)\delta$, the same subspace but only what stage 1 failed to learn, projected back in at cost c: two multiplicative penalties, neither of which goes away with more data. At matched volume $|D|/|V| = 1.39$, so the geometric ordering this picture suggests is not what §6 measures.

Table 1 (main text, §4) gives the same instrument calibration in full.

Table 3. Every recipe decision is inside single-run noise once volume is fixed; volume is not. The right-hand block carries the claim: three contrasts at the learning budget (3 epochs, lr $3{\times}10^{-5}$, Qwen2.5-7B, 622 rows per stage) against the 0.0191 floor, none resolvable. The last row is not a recipe decision and is the only quantity clearing the floor. The left-hand block is context and not a controlled comparison: the natural-size family is also a one-epoch budget (step 39 against the matched family's 117), so size and budget move together, and each row averages over a different set of pairs.

	contrast	natural sizes, 1 ep		matched 622, 3 ep		/floor
		mean $\lvert\cdot\rvert$	pairs	mean $\lvert\cdot\rvert$	pairs	
composition	shuf against only$_B$	0.0913	2	0.0101	2	0.53
order	A→B against B→A	0.0347	6	0.0205	4	1.07
arrangement	shuf against blocked	0.0396	2	0.0140	6	0.73
volume	622 rows against 311, one skill	—		0.0463	4 skills	2.43

Table 4. Doubling one skill's own data, with nothing else varying (Qwen2.5-7B, three seeds; per-skill contrast floor 0.0191; pre-registered read-out V1). The return is strongly heterogeneous: algebra and calculus ≈ 0.07, geometry 0.017, so data buys most where the model is already learning. Nothing here prices a second doubling.

skill	311 rows	622 rows	gain	/floor
algebra	0.4875	0.5556	+0.0681	3.56
calculus	0.4438	0.5146	+0.0708	3.71
geometry	0.2833	0.3007	+0.0174	0.91
combinatorics	0.2646	0.2937	+0.0291	1.52
mean			+0.0463	2.43

Table 5. Every dispersion this paper quotes, in one place, because there are fifteen of them and three pairs of them share a name. The three blocks are three different random quantities, not comparable across blocks: one arm's accuracy; ASYM, a difference of four accuracies; and D, a two-accuracy contrast under a protocol whose re-execution also rebuilds the corpus. Degrees of freedom are given because several rest on very few; every estimate carries its own interval.

symbol	value	what it is the dispersion of	df	estimated on
$\hat{\sigma}$	0.0234	one arm's accuracy across training seeds, frozen before the grid ran and never refit	—	the pilot grid
floor	0.0191	a two-arm contrast at three seeds, built from $\hat{\sigma}$	—	derived
δ_{seed}	0.054	the MDE of a three-seed design, 2.8× the floor$\sqrt{2}$	—	derived
one arm's accuracy, decomposed (Table 6)				
pooled	0.0142	the same quantity $\hat{\sigma}$ freezes, measured	918	379 (family, arm, target) cells
evaluation	0.0132	its k=4 generation-sampling part	—	analytic, from per_problem
training	0.0053	what remains when $k \to \infty$	—	residual
share	0.861	the evaluation fraction of the variance	918	the same 379 cells
ASYM, the order side (§5.3). A different quantity: a difference of four accuracies				
$\hat{\sigma}_{\text{seed}}$	0.0332	ASYM across seeds within one execution; 95% CI $[0.0261, 0.0455]$	26	39 (pair, run, seed) cells
$\hat{\sigma}_{\text{run}}$	0.0072	ASYM across executions, net of seed noise; 95% CI $[0, 0.025]$	7	13 (pair, run) groups
share	0.955	the evaluation fraction of $\hat{\sigma}^2_{\text{seed}}$; 95% CI $[0.51, 1.00]$	26	the same 39 cells
run mean	0.0205	a three-seed mean of ASYM, composing the two above	—	derived
δ_{run}	0.081	what one run per recipe resolves; 95% CI $[0.072, 0.119]$	—	derived
the switch side (§5.3), fitted on D and not on SHARE. Its re-execution rebuilds the corpus, which ASYM's does not				
$\hat{\sigma}_{\text{seed}}$	0.0228	D across seeds; 95% CI $[0.0162, 0.0387]$	11	14 (replicate, seed) cells
$\hat{\sigma}_{\text{run}}$	0.0152	D across executions; 95% CI $[0, 0.0228]$	2	3 replicates
$\delta^{\text{switch}}_{\text{run}}$	0.080	what one execution of the conflict protocol resolves	—	derived
withdrawn, and kept here so the digits can be recognised				
0.0364	—	the run-mean spread of the selected cell, once read as a run-level excess	2	3 runs of one pair
0.144	—	the δ that spread implies; a statement about selection, not about cells	2	the same cell

Table 6. Most of the noise inside one protocol is evaluation sampling, and it is the cheap half. Single-run s.d. is the pooled cross-seed value; the contrast floor is that s.d. carried through $\sqrt{2}$ for a two-arm contrast and $\sqrt{3}$ for three seeds, the paper's convention. Only the generation term falls with k; the training term, 0.0053, is the $k \to \infty$ limit and is what more evaluation cannot buy. This k axis is not the k axis of §5.3's resolution table: here the quantity is one arm's accuracy on one skill, there the run mean of ASYM, whose evaluation share is measured separately at 0.955 rather than taken from this table's 0.861.

samples per problem	single-run s.d.	contrast floor	MDE at 80%
$k = 4$ (as run)	**0.0142**	**0.0116**	**0.0326**
$k = 8$	0.0107	0.0088	0.0246
$k = 16$	0.0085	0.0069	0.0194
$k = 32$	0.0071	0.0058	0.0162
$k \to \infty$ (training only)	0.0053	0.0043	0.0122

Table 7. The directional order effect at two budgets, on the same skills, the same 622 rows per stage and the same seeds. Standard errors use the project floor, 0.0191. The only pair significant at the learning budget is significant in the opposite direction from the published (one-epoch) budget.

skill pair	1 ep, lr 10^{-5}		3 ep, lr $3{\times}10^{-5}$		outcome
	ASYM	$\lvert\cdot\rvert$/s.e.	ASYM	$\lvert\cdot\rvert$/s.e.	
algebra–calculus	+0.0743	3.9	+0.0020	0.1	falls into noise
algebra–geometry	+0.0320	1.7	−0.0333	1.7	sign reversal
algebra–combinatorics	+0.0779	4.1	−0.0619	3.2	sign reversal
calculus–geometry	−0.0396	2.1	+0.0041	0.2	sign reversal
calculus–combinatorics	+0.0104	0.5	−0.0125	0.7	sign reversal
geometry–combinatorics	+0.0506	2.7	+0.0091	0.5	falls into noise

Table 8. Run-to-run variation on all six pairs, at the learning budget. Every independent execution of the identical protocol that exists; δ converts each spread by the project's MDE convention ($2.8\, s\sqrt{2}$). The pair in bold is the one δ_{run} was originally estimated on; it was re-run because it had failed to replicate, and it has the largest spread of the six. Read the column of δ values as six noisy estimates of one number, not as six numbers, five of them resting on one degree of freedom (App. F). The pooled estimate the text quotes, $\delta_{\text{run}} = 0.081$ with 95% CI $[0.072, 0.119]$, comes from the nested model over all 39 (pair, run, seed) cells rather than from any row here.

skill pair	runs	ASYM per run	s.d.	δ	sign
calculus–geometry	2	+0.0041, +0.0035	0.0004	0.002	stable
algebra–calculus	2	+0.0020, +0.0055	0.0025	0.010	stable
calculus–combinatorics	2	−0.0125, −0.0014	0.0078	0.031	stable
algebra–geometry	2	−0.0333, −0.0208	0.0088	0.035	stable
geometry–combinatorics	2	+0.0091, −0.0083	0.0123	0.049	flips
algebra–combinatorics	3	−0.0619, +0.0096, −0.0146	**0.0364**	**0.144**	flips

Table 9. Arrangement is switched on by disagreement, not by difficulty, domain distance, one model or one convention family. $D = \mathrm{acc}_B(\mathrm{shuf}) - \mathrm{acc}_B(\mathrm{A}\to\mathrm{B})$ against the 0.0191 floor; guard is $\mathrm{acc}_B(\mathrm{only}_B)$, which must clear the floor for a row to be interpretable, and one dose point failed it and is excluded rather than read as zero. Qwen2.5-7B, three seeds, learning budget, except where a scale or seed count is named. The synthetic +1 rows are the rebuilt instrument. Every row is a single execution of the protocol; the headline row was executed twice more from the same pool under different split seeds, giving −0.1258 and −0.0825 beside its −0.0985 (§5.3).

premise broken?	the two halves	D	/floor	guard
no, 8 seeds	both halves numerals (control)	−0.0086	0.45	0.3339
no	both halves spelled: harder, no dispute	+0.0012	0.06	0.1983
no	both halves Arabic (control)	−0.0037	0.20	0.3121
no	algebra + physics: domain boundary, no dispute	+0.0096	0.50	0.1642
yes, 8 seeds	numeral against spelled, both correct	**−0.0985**	**5.16**	0.2091
yes	Arabic against Roman, both correct	**−0.1180**	**6.18**	0.1532
yes	synthetic +1 on half the answers	**−0.2346**	**12.3**	—
yes, 3B	numeral against spelled	−0.0263	1.38	0.0459
no, 3B	both halves numerals	−0.0117	0.61	0.2725
yes, 14B	numeral against spelled	**−0.2037**	**10.67**	0.2579
no, 14B	both halves numerals	−0.0071	0.37	0.3983
yes, dose 0.75	three quarters of one half disagree	**−0.0804**	**4.21**	0.1554
yes, dose 0.50	half of one half disagrees	**−0.0425**	**2.23**	0.0771
— , dose 0.25	excluded: minority convention below its learnability edge	—	—	0.0000

Table 10. How often two public corpora disagree about the same problem, and in which of the two ways. Problems are matched on normalised text; convention means the two golds denote the same value written differently, value means they do not, and undecided is what the canonicaliser declined to rule on, reported rather than folded into either column. The convention column is what the narrow canonicaliser proves and is a lower bound by construction: a companion survey's graded instruments on the same pairs widen proved form by an order of magnitude and re-classify the largest pair's lineage (§5.2).

corpora	relation	shared	same	convention	value	undec.
NuminaMath-CoT vs OpenR1-220k	two independent labellings	92,359	90,599	**9**	51	1,700
NuminaMath-CoT vs GSM8K	two independent labellings	7,169	4,928	**12**	148	2,081
OpenR1-220k vs NuminaMath-TIR	two pipelines, two organisations	14,362	13,641	**31**	70	620
NuminaMath-CoT vs NuminaMath-TIR	two pipelines, one curator	65,456	52,626	**1,284**	396	11,150
NuminaMath-CoT with itself	duplicate problems	42,535	33,812	**161**	3,153	5,409

C Full Statements: the Resolution Schedule and the Ratio Propositions

§3 states these narratively; here are the full statements and proofs.

Proposition 2 (Evaluation cannot buy through the run component). $\delta_{\mathrm{run}}(k)$ is strictly decreasing in k with $\lim_{k\to\infty}\delta_{\mathrm{run}}(k) = 2.8\sqrt{2}\,\sqrt{\sigma^2_{\mathrm{train}}/n + \sigma^2_{\mathrm{run}}} > 0$ whenever $\sigma^2_{\mathrm{train}} + n\,\sigma^2_{\mathrm{run}} > 0$. Hence a target separation Δ is reachable by buying samples alone if and only if $\Delta > 2.8\sqrt{2}\sqrt{\sigma^2_{\mathrm{train}}/n + \sigma^2_{\mathrm{run}}}$, a condition that does not involve k.

Proof. Immediate from the resolution schedule of §3: the only k-dependent term is $\sigma^2_{\mathrm{eval}}k_0/k$, which is positive and strictly decreasing with infimum 0. □

Remark 1 (A form of Proposition 1(i) we printed and withdraw). Earlier versions stated Proposition 1(i) as $\mathbb{E}[\hat{\Delta}] \geq \Delta + 2\sigma\sqrt{2\log m}\,(1 - o(1))$, and both halves of that are wrong at this paper's m. The constant $2\sqrt{2\log m}$ is the asymptotic expected range, not a finite-m bound: at $m = 6$ it reads 3.786 against a true $d_2(6) = 2.534$, an overstatement of 49%. And the additive form fails for every $\Delta > 0$; at $m = 2$ with $\mu = (0, \Delta)$ the range is $|\Delta + Z|$, $Z \sim N(0, 2\sigma^2)$, whose expectation tends to Δ rather than to $\Delta + \mathbb{E}|Z|$. The conclusion the paper draws from part (i) is unchanged, and the corrected constant is smaller, so the bound now bites less hard and can be checked against a table. Source: review_variance.

Proposition 3 (The plug-in ratio is an upper bound). Let the numerator be estimated by the largest observed contrast over m corpora, $\hat{\Delta} = \mathrm{range}\{Y(r)\}$. By Proposition 1(i),

$\mathbb{E}[\hat{\Delta}] \geq \max\{\Delta, d_2(m)\sigma\} \geq \Delta$, so $\mathbb{E}[\hat{\Delta}]/\delta \geq \Delta/\delta$: the plug-in estimate of ρ is biased upward, and a reported $\rho < 1$ therefore understates how far below the wall the search sits.

The recipe side's numerator is exactly such a maximum, so its ρ is an upper bound and the conclusion drawn from it (§5.3) is conservative. The switch's numerator is a single pre-registered row rather than a selected maximum, so Proposition 3 does not apply to it.

Proposition 4 (The ordering of two ratios is decided at one corner). Let two departures have effects $E_1, E_2 > 0$ measured on the same scale, seed dispersions $s_1, s_2 > 0$, n seeds each, and unrelated protocol-level components σ_1, σ_2 known only to lie in $[0, a_1]$ and $[0, a_2]$. Each $\rho_i(\sigma_i) = E_i/\big(2.8\sqrt{2}\,\sqrt{s_i^2/n + \sigma_i^2}\big)$ is strictly decreasing in its own σ_i and free of the other's, so

$$\inf_{\sigma_1 \in [0,a_1],\ \sigma_2 \in [0,a_2]} \big(\rho_2 - \rho_1\big) = \rho_2(a_2) - \rho_1(0).$$

Hence $\rho_2 > \rho_1$ everywhere on the rectangle if and only if it holds at the single corner $(\sigma_1, \sigma_2) = (0, a_2)$. No relation between the two components is assumed.

Proof. $\partial\rho_i/\partial\sigma_i < 0$ and $\partial\rho_i/\partial\sigma_j = 0$ for $j \neq i$, so $\rho_2 - \rho_1$ is decreasing in σ_2 and increasing in σ_1; a monotone function on a product of intervals attains its infimum at the corresponding vertex. □

With $a_1 = 0.025$ and $a_2 = 0.0228$, the two measured intervals, the corner reads $\rho_2 = 0.98$ against $\rho_1 = 0.82$: the switch is better resolved than the recipe axes at every pair of values the two components can take, even though neither side's absolute verdict is settled anywhere on the rectangle.

Proposition 5 (A σ-free bound when the component is shared). If instead $\sigma_1 = \sigma_2 = \sigma$, then

$$\frac{\rho_2(\sigma)}{\rho_1(\sigma)} = \frac{E_2}{E_1}\sqrt{\frac{s_1^2/n + \sigma^2}{s_2^2/n + \sigma^2}},$$

which for $s_1 > s_2$ is strictly decreasing in σ from $(E_2/E_1)(s_1/s_2)$ at $\sigma = 0$ to E_2/E_1 as $\sigma \to \infty$. Hence $\rho_2 > \rho_1$ for every $\sigma \geq 0$ if and only if $E_2 \geq E_1$, a condition on two measured effects with no variance component in it.

Proof. Substitute the resolution schedule of §3 into $\rho_i = E_i/\delta_i$; the constant $2.8\sqrt{2}$ cancels. The bracket is a ratio of two increasing functions of σ^2 with the larger constant on top, so it exceeds 1 and decreases to 1. The infimum of the product is therefore E_2/E_1. □

With the measured values, $E_2/E_1 = 0.102/0.0619 = 1.65$ and $s_1/s_2 = 0.0332/0.0228 = 1.46$, so Proposition 5 bounds the ratio of ratios in $[1.65, 2.40]$ for every shared σ — but its premise is one this paper's own measurements argue against: re-executing the conflict protocol rebuilds the corpus and re-executing the order protocol does not, so the two carry different protocol-level components, estimated at 0.0152 and 0.0072 (§5.3). Proposition 4 needs no such premise and is the load-bearing one; Proposition 5 is reported because it is exact and because it shows the ordering is not an artefact of where the two intervals happen to end.

Corollary 1 (A bound on ρ that needs no interval). $\rho(\sigma_{\text{run}}) = E/\big(2.8\sqrt{2}\sqrt{s^2/n + \sigma_{\text{run}}^2}\big)$ is strictly decreasing, so

$$\sup_{\sigma_{\text{run}} \geq 0} \rho \;=\; \frac{E}{2.8\sqrt{2}\,s/\sqrt{n}},$$

attained at $\sigma_{\text{run}} = 0$. A departure whose supremum is below one is unresolvable at every value of the component, and no interval on σ_{run} is needed to say so. §5.3 measures $\sup_{\sigma_{\text{run}}} \rho_{\text{recipe}} = 0.816$; this is stronger than reading a bootstrap interval's lower end, and it is where the paper's negative result lives.

D The Limit as Theorems

The wall of §3 says what a search can and cannot certify; it does not say which recipe decisions should move at all. That comes from an idealised learner whose three premises this paper breaks one at a time. This appendix states it, proves it, and grades it; the body carries only its record and the two finite-budget extensions of §3 that generate falsifiable signs.

Status of the idealisation, stated before it is used. The theorems below are a first-order, lazy-regime account, displayed because they generate the paper's predictions and say which experiment is worth running, not because they are load-bearing on their own.

Their record is the honest way to introduce them. On signs and existences, meaning which quantity a departure should move, in which direction, under which broken premise: five of five. On locations and magnitudes: zero of six. Every location-level prediction died (§6), and the one magnitude that turns out to be derivable at all is graded below and misses too. Beyond that, the geometric ordering is contradicted by the data ($|D|/|V| = 1.39$ where it needs < 1), its sufficient condition is unreachable for any pair of subspaces from one base model, and the excuse we had been offering for the first of those, that full SFT is not lazy enough, is measured below and returns untested. Treat what follows as an instrument for generating falsifiable directions and not a line further. We keep it because the five-for-five record on signs is what told us which experiments were worth running, and because a theory whose quantitative half is dead in six places is more useful stated with its scorecard than trimmed until it looks successful.

(For the registered ledger, which is a different count: five of the six failed registered claims came from the theory and the sixth came from our own name for a result; §6 keeps the two separate.)

A skill X is a finite sample set whose inputs span $\widehat{S}_X$, with orthogonal projection P_X; targets decompose as $w^\star_X = \bar{w} + \kappa_X$ with a shared core $\bar{w}$ and a skill-specific convention κ_X. Coherence is $\kappa_A = \kappa_B$. Gradient descent on skill X moves parameters only inside $\widehat{S}_X$, and at convergence acts as $\theta \mapsto (I - P_X)\theta + P_X w^\star_X$. Write $\delta := \theta_0 - \bar{w}$, $W := \widehat{S}_{A\cup B} \cap \widehat{S}_B^{\perp}$ (the directions A supplies outside B's span), and let e_S, e_J be the error vectors of the blocked (A→B) and interleaved arms on B.

The limit is three statements, and Figure 4 draws the subspace picture behind the first two. Theorem 1 says that under patience and coherence: arrangement carries only what stage 1 failed to learn, re-projected at a cost, so it is doubly suppressed and vanishes for orthogonal skills; order is carried by the commutator $[P_B, P_A]\delta$, which is zero for orthogonal and for nested skills; and composition without a capacity constraint can only add directions, so $V \geq 0$ and an observed $V < 0$ at matched budget is itself evidence that a premise broke. The proofs are four lines each, compositions of the convergence map. They sit here rather than in the body because the account's quantitative half is dead in six places, and two pages of derivation beside that verdict reads as reluctance to part with it. Every statement, with the two finite-budget extensions of §3, is verified numerically by the 25 machine checks of App. J.

We have offered that laziness gap as the reason the geometric ordering fails, and we can now say we do not know it is. The ordering predicts $|D| < |V|$ and full SFT at matched volume gives 1.39. If the reason is that full SFT is not lazy, the ordering should do better where the regime is more nearly lazy, and the LoRA rank axis is exactly that: already on disk at six ranks over two skill pairs and three seeds, with the arrangement contrast clean of the volume confound that afflicts the composition contrast, since shuffled and blocked arms both train the whole pooled corpus. Fitted over all 36 cells, $|D|$ against $\log_2 r$ has a slope of $+0.0022$ per doubling: the sign the disclaimer predicts, on an interval $[-0.0013, +0.0057]$ containing zero. And the design is too weak to make that a null, since it could only have caught a change in $|D|$ larger than 1.31 contrast floors across the whole five-doubling axis. So the laziness gap is untested, not measured, and the one axis this project can test it on

returns nothing either way, which costs us the most convenient account of our own theory's failure. Source: theory_domain.

An earlier version ended this paragraph with "nothing in this paper derives a magnitude", and that was one magnitude too many. Eq. (2) puts the interleaved arm at the sample weight $n_A/(n_A + n_B)$ on the contested directions, and the conflicted corpora are random halves of one pool, so the weight is 0.5 and the interleaved arm's allocation is measured. We had never graded it. Graded now, it misses: the interleaved arm sits at 0.413 on eight seeds, a deviation of -0.087 or -17.4%, which exceeds the resolution of the design that measured it at every point of the run component's interval, by $3.9\times$ at the optimistic end and still by $1.3\times$ at the pessimistic one. That resolution is a one-sample MDE, $2.8\,s$ with $s = \sqrt{\hat{\sigma}^2_{\text{seed}}/n + \hat{\sigma}^2_{\text{run}}}$, and not the $2.8\,s\sqrt{2}$ this paper uses everywhere else, because the comparison is one arm against the fixed constant 0.5 rather than two arms against each other; a reader carrying the contrast convention across would find a spurious factor of $\sqrt{2}$ and conclude we had erred. Continuing: the arm's whole 27-fold block-length sweep, 0.407–0.449, lies below 0.5. What the miss cannot say is which step failed. Eq. (2) is a statement in parameter space about the contested directions of κ, the observable is an accuracy share, and nothing here derives the map between them, so "the magnitude is wrong" and "the map is not the identity" are not separated. The honest record is that the theory's one derivable location was tested and did not survive, which takes its location score from 0 of 5 to 0 of 6. This test is post hoc and is not a registered claim, so the scorecard of §6 stays at 30 claims and 7 failures; the two ledgers are counted separately on purpose. Source: theory_magnitude.

Theorem 1 (The limit). Under patience (each stage to convergence) and coherence:

(a) Arrangement: $e_S - e_J = P_W(I - P_A)\,\delta$. This vanishes identically on B's training set, vanishes exactly for orthogonal skills, and is otherwise doubly suppressed: it carries only what stage 1 failed to learn, re-projected into W at cost $c = \|P_B P_A\|_{\text{op}}$.

(b) Order: the directional contrast is carried by the commutator $[P_B, P_A]\,\delta$, whose norm is $\max_i \cos\theta_i \sin\theta_i \le \frac{1}{2}$ over principal angles: zero for orthogonal skills and for nested ones. Under coherent targets the recency terms it compares are both $O(\varepsilon)$ in the unexplored mass, so order is free in the limit.

(c) Composition: without a capacity constraint, $e_O - e_J = P_W\delta$: co-training can only add directions, so its value $V \ge 0$. Observed $V < 0$ at matched budget is itself evidence that some premise broke.

E Supporting Measurements on the Three Axes

The room axis: the sufficient condition, measured on the model

§6 states that $c\rho < \gamma$ fails in every one of the 12 ordered skill pairs; here is the measurement. The count is over cells and the unit is worth naming: four skills give 12 ordered pairs, each measured at three span dimensions, so the failing set is 36 of 36 (pair, dimension) cells at one probe depth and 144 of 144 across the four depths. ρ and γ are ratios of projections of the fine-tuning displacement (Frobenius norms of ΔW at the probed layer's input space), c the largest principal-angle cosine between skill activation spans; k is the span dimension and the one free parameter, so the measurement is reported across a range. The projection algebra is verified against the definitions to 10^{-7}. Table 11 reports it.

The margin shrinks with k roughly as $k^{-0.45}$, and the tempting extrapolation to ratio ≈ 1 at full rank is meaningless: at $k = d$ both ρ and γ are identically zero and the ratio is $0/0$. The premise question is only posed where a span carries information, and at every such cut it fails by $3.9\times$ to $13.5\times$.

The axis and its control, plotted

The sign change Theorem 1(c) requires, and the volume-matched control that removes it for the pair that carried it. Kept because the sequence of effect, decomposition and control is

Table 11. The premise fails at every span dimension and every depth we can measure (Qwen2.5-7B, learning budget, averaged over the twelve ordered skill pairs). Left: span dimension k at layer fraction 0.75. Right: probe depth, with c, ρ and γ read at k=1024 and the failure count taken over all three span dimensions, so its denominator is 12×3 cells and not twelve pairs. The ratio $c\rho/\gamma$ must fall below 1 for the theorem to bite; it never falls below 3.9, and $c \geq 0.999$ everywhere, the four skills' representation spans being all but identical, which is the coherent-domain scope condition written in subspace terms.

k	c	ρ	γ	$c\rho/\gamma$
64	0.9994	0.9958	0.0738	13.5
256	0.9998	0.9876	0.1337	7.4
1024	0.9999	0.9616	0.2467	3.9

layer	c	ρ	γ	fails
7	1.0000	0.9568	0.2504	36/36
14	0.9991	0.9607	0.2462	36/36
20	0.9999	0.9616	0.2467	36/36
24	0.9999	0.9584	0.2493	36/36

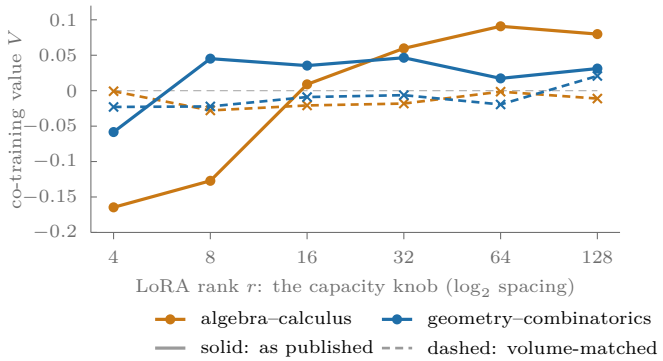


Figure 5. A sign change that its own control removes. Co-training value V against the capacity knob, for the two skill pairs measured on both versions of the axis. Points are three-seed means with problem-level standard errors. Solid, as published, with arms differing by a factor of two in rows: V crosses zero exactly as Theorem 1(c) predicts when capacity binds, monotonically and with no interior peak, 11 of 12 (pair, rank) cells agreeing in sign across three seeds. Dashed, the same grid re-run with both arms at 622 rows: the crossing does not survive for algebra–calculus at all, and geometry–combinatorics keeps one by a single cell. The distance between a solid curve and its dashed counterpart is what the volume confound was worth, and it is most of the published effect. Existence is confirmed and location falsified: part (c) predicts the higher-demand pair (geometry–comb., $\hat{y}$=2.65 against 2.08) crosses later, a ratio $r^{\star}_{\text{geo,comb}}/r^{\star}_{\text{alg,calc}}$ of 1.271 under load accounting and 1.0 if pair-independent; bootstrapping the zero-crossings over 4000 draws that resample both problems and seeds measures 0.382, 95% CI $[0.231, 0.583]$, excluding both and reversing the direction. Same data as Table 12.

the reusable part, and filed here rather than in §5.4 because a full-width figure is the wrong amount of emphasis for a result that section withdraws.

The conflict rows under convention-agnostic scoring

The objection. In every conflicted corpus the blocked arm trains convention A first and B last, and the committed read-out $D = \text{acc}_B(\text{shuf}) - \text{acc}_B(A{\to}B)$ scores exactly the convention the blocked arm wrote last. Blocking's win could therefore be the evaluation siding with the last stage.

The instrument. Both conventions are evaluated on the same problems in the same order (the analyzer re-asserts this on the eval files before reading any number), so a scoring rule with no favourite convention exists without new inference: count each problem at the rate of its better-scoring convention, $\text{agn} = \frac{1}{n}\sum_i \max(p_i^A, p_i^B)$ with p_i^A, p_i^B the k=4 pass fractions under each convention, and read $D_{\text{agn}} = \text{agn}(\text{shuf}) - \text{agn}(A{\to}B)$ on the very runs that produced the committed D. Three read-outs then separate three stories. D committed answers which convention wins; ΔSOLVE answers how much is solved, and is conserved when samples merely reallocate between conventions on one problem; D_{agn} moves if and only if commitment itself has a price. A row with SOLVE at the floor and D_{agn} firing is

Table 12. The capacity axis against its own control (Qwen2.5-7B, three seeds). "Published" varies LoRA rank with arms differing by a factor of two in rows; "matched" gives both arms exactly 622 rows at every rank. The pair carrying the published effect loses its crossing entirely; the other keeps one by a single 1.1σ cell that sits below the MDE_{80} gating every other cell in the grid.

LoRA rank r	algebra–calculus		geometry–comb.	
	published	matched	published	matched
4	−0.165	**−0.0007**	−0.058	**−0.0229**
8	−0.127	**−0.0278**	+0.045	**−0.0222**
16	+0.009	**−0.0208**	+0.035	**−0.0090**
32	+0.060	**−0.0181**	+0.046	**−0.0063**
64	+0.091	**−0.0014**	+0.017	**−0.0194**
128	+0.080	**−0.0111**	+0.031	**+0.0208**
verdict	crossing vanishes		crossing survives	

a shuffled arm splitting its four samples across conventions on the same problem, not one solving fewer problems.

Table 13. The committed win is allocation where both conventions are correct, and per-problem mixing where they are not. D committed is the main text's contrast, ΔSOLVE its capability sum, D agnostic the best-convention-per-problem contrast. Parentheses give distance from zero in 0.0191 floors; bold marks ≥ 2 floors. The numeral rows at 7B and the synthetic row carry eight seeds; the Roman row and the two scale rows carry three, so this table's Roman entries are not the five-seed ones Table 9 prints. The three-seed synthetic subset previously printed, −0.2275 at 11.91 floors, had a t of −88.8 against the eight-seed −33.4, which is why the count is now stated beside every row.

conflicted corpus	n	D committed	ΔSOLVE	D agnostic
numeral against spelled, 7B	8	**−0.0985** (**5.16**)	+0.0016 (0.09)	−0.0084 (0.44)
Arabic against Roman, 7B	3	**−0.1046** (**5.48**)	**+0.0500** (**2.62**)	+0.0308 (1.61)
synthetic +1, 7B	8	**−0.2346** (**12.29**)	−0.0052 (0.27)	**−0.0887** (**4.65**)
numeral against spelled, 3B	3	−0.0263 (1.38)	−0.0071 (0.37)	+0.0004 (0.02)
numeral against spelled, 14B	3	**−0.2037** (**10.67**)	−0.0312 (1.64)	**−0.0496** (**2.60**)

Reading, row by row. At 7B, where both conventions are correct, the committed wins of −0.0985 and −0.1180 become −0.0084 (0.44 floors) and +0.0308 (1.61 floors) under agnostic scoring: an order of magnitude smaller and inside two floors, so blocking solves nothing the shuffled arm cannot solve and only commits, which is §5.2's claim re-derived on a metric that cannot be accused of siding with the last stage. (The Roman row's ΔSOLVE, +0.0465 at 2.43 floors with five seeds, is previously unreported and is given as measured: if anything the shuffled arm solves more there, opposite to a capability story for blocking. Its controls read null with $|t| < 0.7$ and per-seed spreads up to ±0.11, so we quote it and decline to lean on it.) The two rows where D_{agn} does fire, synthetic +1 at −0.0887 and 14B at −0.0496, are precisely those where SOLVE stays low (0.27 and 1.64 floors): the shuffled arm loses its best convention per problem while its sum holds, which is mixing within a problem, the compressive-write signature the path-level account predicts (Anonymous, 2026). Where the conflict is between two correct formats, commitment is free and the whole committed effect is allocation; where it contests answer values, refusing to commit costs the shuffled arm its per-problem best while costing it nothing in sum. No row supports "blocking teaches the model more mathematics".

The per-skill order potential

Held out of §5.3 because its evidential weight does not match its visual weight: a three-parameter fit to six observations at one budget, four of whose six pairs reverse sign at the other budget we measured. It is reported in full because it would be the cheapest thing in this paper if it replicated at ≥ 6 skills.

At one stated budget the six matched pairwise effects are consistent with a per-skill potential, $\mathrm{ASYM}(X, Y) \approx \phi_X - \phi_Y$ (ϕ: algebra 0 > geometry -0.032 > calculus $-0.072 \approx$ combinatorics -0.081), and the pre-registered additivity-over-triples criterion passes at 0.073 against a threshold of 1.0. On unmatched data the same test reads 1.109 and the recovered ordering inverts, which is the sharpest argument we have that volume matching is mandatory.

The structure predicts out of sample, which is the part worth having. Additivity over triples is a goodness-of-fit statistic on the data ϕ was fitted to, and weak evidence on its own: three free parameters against six observations. So we froze ϕ, fitted on pairs only, wrote down the implied ranking of all six three-skill orderings, and trained them. The prediction file is timestamped 1h46m before the first three-stage run existed, and the scoring script fits nothing. Both read-outs were fixed in that file: P1, Spearman $\rho \geq 0.60$ between the predicted $\phi_{\text{first}} - \phi_{\text{last}}$ and the measured TOTAL; P2, the predicted-best and predicted-worst orderings separated by more than 2 s.e. Table 14 gives $\rho = 0.657$ and a best-to-worst gap of 0.1215 ± 0.0331, 3.7 s.e., so both pass: a quantity fitted on pairs ranked sequences it had never seen. So order search costs n per-skill measurements rather than $n!$ trials, the one constructive result this paper sets against §5.3's wall, which says a search over recipes cannot pay for itself; the potential says the search is unnecessary where it would have been run.

Three reasons still keep the structure held lightly, none repaired by Table 14. The ranking is not monotone, the predicted-third ordering outscoring the predicted-second by 0.036, which is why ρ is 0.657 and not 1, and only the extremes are separated by more than noise. Four of the six pairs reverse sign between our two budgets, so ϕ is a snapshot at its stated budget, to be refit at the budget where it will be used. And the additivity residual beats the marginal noise floor only because same-seed arms share checkpoints, a permutation check anyone reporting a suspiciously good fit should run. A $\geq$ 6-skill replication is now registered with the criterion that matters : the fit sees ten pairs, five are held out entirely, the verdict turns on out-of-sample sign and error rather than the in-sample residual, and a clause commits to withdrawing this result rather than softening it if the held-out test fails. What is established here is narrower than the additive form: the fitted potential has out-of-sample ordering power at $n = 3$, at one budget, for four skills.

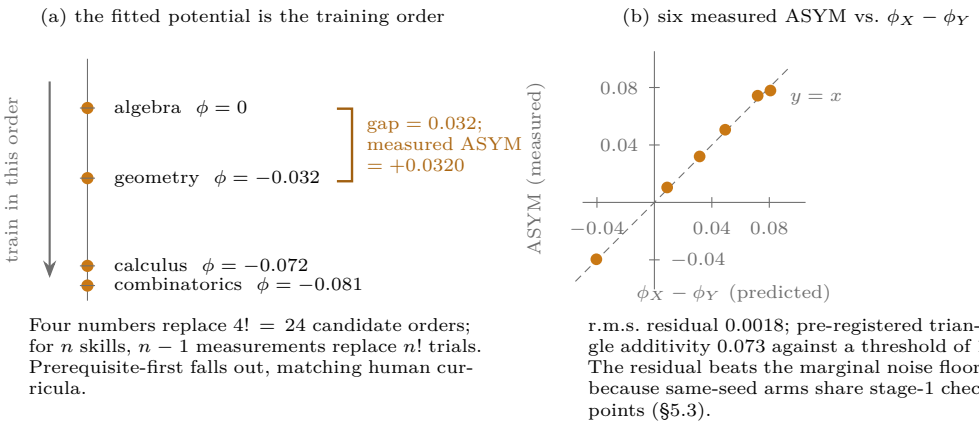
Table 14. A potential fitted on pairs ranks orderings it never saw. All six orderings of three skills, trained to the same budget, scored by TOTAL at the final checkpoint, so an ordering is rewarded for what it retains, not for what it learned last. The predicted column is $\phi_{\text{first}} - \phi_{\text{last}}$ from a file frozen before any three-stage sequence was trained; the middle position cancels at $n = 3$). Qwen2.5-7B, three seeds, per-cell s.e. 0.0234. The analysis measures and compares and fits nothing.

predicted rank	ordering (three stages, final checkpoint)	$\phi_{\text{first}} - \phi_{\text{last}}$ (frozen)	measured TOTAL	seed spread
1	algebra → geometry → combinatorics	+0.0808	**0.7284**	0.0834
2	geometry → algebra → combinatorics	+0.0493	0.6555	0.0313
3	algebra → combinatorics → geometry	+0.0315	0.6917	0.0438
4	geometry → combinatorics → algebra	−0.0315	0.6889	0.0395
5	combinatorics → algebra → geometry	−0.0493	0.6681	0.0667
6	combinatorics → geometry → algebra	−0.0808	**0.6069**	0.0437
Spearman ρ (P1 threshold 0.60)			**0.657** pass	
best − worst (P2 threshold 2 s.e.)			$0.1215 \pm 0.0331 = \mathbf{3.7}$ s.e.	pass

F The Instrument: Counting, Variation and the Optimiser Confound

The counting rule behind the two run totals

The grid analysed in the body is 761 fine-tuning runs and 2158 scored evaluations, drawn from a released base of 1233 runs and 3274 evaluations that also holds the other experiments of this line; both figures are reported so a reader checking the archive against the paper finds the larger number explained. Counting rule: a run is one directory holding the output of one trainer invocation, whether or not its checkpoint still exists, since checkpoints are deleted



Four numbers replace 4! = 24 candidate orders; for n skills, $n - 1$ measurements replace $n!$ trials. Prerequisite-first falls out, matching human curricula.

r.m.s. residual 0.0018; pre-registered triangle additivity 0.073 against a threshold of 1.0. The residual beats the marginal noise floor because same-seed arms share stage-1 checkpoints (§5.3).

Figure 6. At one stated budget, order collapses onto one number per skill. (a) The potential fitted to six volume-matched pairwise measurements; training in decreasing ϕ is prerequisite-first. (b) Each pair's measured directional effect against the potential difference the fit predicts. Six measurements are summarised by three free numbers and the pre-registered additivity criterion passes by more than an order of magnitude, but three parameters against six observations is weak evidence, and the sharpest of the three reasons to hold this lightly is that four of these six pairs reverse sign at the budget that teaches.

once their results are recorded; a scored evaluation is one (checkpoint, target) pair carrying an accuracy at the final checkpoint of its run, so the further 2277 intermediate-checkpoint evaluations of the trajectory families are counted separately rather than folded in, each measuring one run repeatedly. The split between the two figures is by experiment family, each family assigned to exactly one side and none unassigned; the assignment is released with the counting script rather than described. Every number is produced by an analyzer rather than transcribed by hand, and the result base, analyzers, 25 machine checks and the scorecard are packaged for release with the paper. The run figure and the evaluation figure are not counted the same way, and the second version of this paper made the difference visible. Runs are counted from a manifest of run directories, frozen before this version's trainings because the directories themselves are reclaimed once their results are recorded; evaluations are counted from the shipped result files. So the 36 replay trainings of §5.1 are inside the evaluation figure and outside the run figure, and 761 is a count of the runs the manifest covers rather than of every training this paper reports. Both figures are what the released archive's own census computes, so a reader recomputes them rather than taking them.

Run-to-run variation: selection and heteroscedasticity

δ_{run} was first estimated on a single pair, and that pair was the one re-run because it had failed to replicate. Selecting a cell for having moved and then measuring how much cells move is a biased estimator whose bias runs upward, in the direction that makes this paper's only biting bound bite harder. §5.3 reports the repair, three further executions and a pooled fit in place of the selected cell; what follows is the test of the reading those runs were expected to support, which they did not (Table 8).

A spread from two runs is one degree of freedom, so we fit the model instead of reading the table. Six such spreads invite a comparison between them: the selected pair has the largest and the median of the other five is 4.6× smaller, which reads as run-level spread is a property of the cell, not a constant of the protocol (App. J). That reading does not survive testing. An n=2 spread is half-normal, spanning a factor of 31 between its own 5th and 95th percentiles, so six of them look heterogeneous whether or not they are. A likelihood-ratio test of one common variance across all six gives $p = 0.045$; with the selected cell removed, $p = 0.289$. The entire signal is the one cell that cannot serve as its own evidence, since it was re-run because it had failed to replicate and two of its three runs are the two that triggered the selection. The sign reversals carry no separate weight either: at an effect inside the noise

a sign is a fair coin by Proposition 1(ii), which is the prediction, not a second observation of it.

The optimiser-memory confound in full

§5.3 states the confound and the two measurements that bound it; here are the mechanism, the test at full resolution, and the control that is still missing.

What the documented channel is. Sweeney (2026b) shows that for a memoryless optimiser, reordering an equal multiset has no first-order endpoint term, the leading local contrast being exactly the $O(\eta^2)$ gradient bracket Eq. (1) carries, while a stateful fixed-clock optimiser adds an $O(\eta)$ order-dependent noise channel through its moment buffers (measured order-variance slopes 1.83 for AdamW against 4.00 for SGD), large enough for a single seed to flip a close comparison. So an alternative reading of Figure 2 exists and must be faced: the crossings could be optimiser artefact rather than the two-term competition.

The channel is a variance channel, and that is what the test uses. Being independent across seeds by construction, it can move any one seed's curve but cannot put the same crossing in all of them. So we asked of each seed's own trajectory, with no pooling anywhere in the test, whether it changes sign in the same interval as its siblings (Table 15). Six crossings are unanimous and every one of the four pairs has at least one, while the remaining ten pooled crossings are exactly those whose flanking checkpoints sit inside the noise band, which is where a variance channel should and does show up. The published trajectory splits the same way, its band-clearing crossings at steps 17 and 43 being 3/3 and its in-band crossing at 112 1/3. We note the temptation we avoided, of asking instead whether seeds agree in sign where the pooled mean clears the floor; that ratio is 18/22, but the question is circular, since a pooled mean only clears the floor when the seeds already agree.

So the standing count is two, not three. The floor already contains first-order shuffle noise, since $\hat{\sigma}$ is estimated across seeds that differ in data order; and the per-seed unanimity test excludes a seed-independent variance channel on each seed's own trajectory, using no run-level quantity at all. The third argument read the run-to-run excess as this channel's residue, and no such excess survives the fit above, so it is withdrawn (App. J), the same correction §5.2 makes to its own list of four controls. And the memoryless mechanism is sufficient without optimiser state. Plain gradient descent reproduces the crossings in-model (72% of random non-commuting draws, Figure 3); both stage-2 arms of ASYM begin from freshly initialised optimiser state, so the stage boundary's clock is symmetric across the contrast; and the second-moment timescale $1/(1-\beta_2) = 1000$ steps exceeds every stage-2 run here, leaving the 10-step momentum window as the only live memory.

The boundary is therefore narrow. The existence of the transient is robust: the sign changes inside a single run, in all four pairs, with each pair's band-clearing crossing reproduced by every seed on its own. Its attribution to Eq. (1) is the best-supported mechanism rather than an exclusive one, and what is not excluded is specific: a systematic AdamW contribution, a bias rather than a variance, to where the crossings sit. Since the paper's claim is existence and budget-dependence rather than location, that gap carries no conclusion, and nowhere do we recommend a budget because a crossing was measured at a particular step. The experiment that would close it was pre-registered before any of this was written: one pair's trajectory replicated under plain SGD, with a guard reporting not measurable if no calibrated SGD arm reaches the AdamW endpoints (prereg/sgd_trajectory_control.md). It has not been run, and the registration stands rather than being withdrawn quietly, because the one documented alternative mechanism is the variance channel, the unanimity test rules it out on our own data, and the residual possibility is a bias channel no published account proposes and that would move only a quantity this paper does not use.

A bound implied by reported pairwise accuracies

§2 states the bound $q \geq A$ and our own q; here we apply it, inverts it, and states what it cannot settle. Applied to the accuracies Sweeney (2026a) report, 98.1% at one step per domain asserts that those order signs re-run identically at least 96.3% of the time, and

Table 15. The crossings that clear the noise band are made by every seed separately; the ones inside it are not. Each seed's own ASYM(t) is tested for a sign change in each interval between adjacent checkpoints, with no pooling anywhere in the column, so the test is not the circular one of asking whether seeds agree where their mean is large. A crossing is unanimous when every seed of that pair changes sign in the same interval, which a seed-independent variance channel cannot satisfy. Flanking values are the pooled |ASYM| at the two bracketing checkpoints, in 0.0191 floors. Qwen2.5-7B, learning budget; the combinatorics–geometry pair has two seeds, the rest three.

pair	crossing at	flanking |ASYM|, floors	seeds crossing in that interval	
algebra-combinatorics (3 seeds)	step 17	2.8, 5.9	**3/3**	unanimous
	step 43	2.9, 6.1	**3/3**	unanimous
	step 112	0.7, 0.5	1/3	
algebra-calculus (3 seeds)	step 45	1.9, 2.2	**3/3**	unanimous
	step 76	3.4, 0.8	2/3	
	step 85	0.8, 0.7	1/3	
	step 116	3.4, 0.3	1/3	
calculus-geometry (3 seeds)	step 34	6.1, 4.2	**3/3**	unanimous
	step 51	4.2, 0.4	1/3	
	step 54	0.4, 1.9	0/3	
	step 84	1.0, 1.0	0/3	
	step 115	1.1, 0.2	1/3	
combinatorics-geometry (2 seeds)	step 34	3.8, 2.6	**2/2**	unanimous
	step 74	2.0, 0.9	1/2	
	step 103	1.4, 0.1	**2/2**	unanimous
	step 105	0.1, 2.1	1/2	

73.1% at twenty steps asserts at least 60.7%, against the 66.7% we measure over six pairs with independent re-runs of the identical protocol at a comparable SFT budget (§5.3).

Stated as a tool rather than as a remark, because it costs nothing and applies to a whole literature. Any paper reporting an accuracy A for predicting a measured binary outcome is asserting $q \geq A$ about its own labels, and the implied two-run agreement is $q^2 + (1-q)^2$, which inverts to $q = \frac{1}{2}\big(1 + \sqrt{2a-1}\big)$ for an observed agreement a. Three lines of arithmetic on a number already in an abstract yield a claim the paper has almost certainly not tested, and in this corner of the literature the claim is usually large: predicting order effects at 90% requires order signs that re-run identically 82% of the time, which is well above what we measure at a comparable budget. We are not tabulating other papers' numbers here. Doing it properly means checking, for each one, whether the reported accuracy is against a measured sign at all, at what budget, and with how many runs behind each label, and a table assembled without those checks would be exactly the kind of unverified transcription this paper spends its appendices avoiding. We give the arithmetic and our own measurement of q; the audit is available to anyone, including to a referee holding this paper.

Two readings survive and we can separate neither, which is the point. Their labels may simply be more reproducible than ours, at a smaller budget on different corpora, in which case the decay they report is the budget dependence they read it as and Eq. (1) explains its mechanism. Or the decay is the reliability of the label falling as the budget grows, in which case a predictor's score is tracking its criterion's noise and needs no mechanism at all. The implied floor on their own labels does fall from 96.3% to 60.7% across the range they report, which is what the second reading predicts and the first does not require. This costs us as much as them: our own Magnus-truncation account of that decay now has a competitor we cannot exclude, and we state it here rather than keep the tidier version. What separates them is one number that no order study we know of reports, ours included until §5.3: re-run the same arms and publish how often the sign survives. The idealisation-first structure, stating the limit in which the phenomenon is exactly zero and then measuring departures, is the one Anonymous (2026) introduce, and the debt is structural rather than stylistic: CCH reads inference through a bounded state channel between an unbounded stream and a prediction, with disagreement in the stream making the choice of access channel visible in behaviour. That premise returns here with the parameters as the state and the data path as the stream, and §5.2's switch is its training-time signature. A concurrent manuscript develops the channel-level account in full (Anonymous, 2026), cited for provenance and

never for evidence: every number and every control in what follows is in this paper or its appendices.

G The Wild-Conflict Episode and the Cross-Family Replication

§5.2 states the three gaps that bound what Table 9 licenses and gives the separability precondition they produced. This appendix is the episode behind the first of them, at the same volume as the firing rows.

Every firing conflict was built, and the attempt to find one in the wild returned an instrument result rather than a row. The corpus existed and we did not build the disagreement: the 1,137 problems of Table 10's third row, in two disjoint halves of 443, with the conflict arm untouched public data and the control carrying our edit, so style differs identically in both arms and only convention differs between them. Forty arms later the registered verdict is W3: the control moves at 4.45 floors, so the frozen rule permits no read-out. The mechanism is not the one W3 anticipated and was found by reading the arm table rather than the verdict. Every arm of both conditions sits at allocation share 0.500 to 0.506 with $\mathrm{acc}_A \approx \mathrm{acc}_B$ inside every cell, because the scorer's first branch tests mathematical equality and this corpus is by definition a set of pairs whose golds denote one value written two ways: 250 of 250 evaluation gold pairs are equal under our own scorer. The build-time preflight did assert that the golds differ, as strings, and passed on all 250. One assertion, the wrong operator, forty trainings downstream.

The learnability guard could not have caught it, and that is the transferable part. It read 0.3912, twenty times the floor, because acc_B is high when the model solves the problems rather than when it adopts a minority convention. Every conflict guard we know of, ours included, asks whether the minority content was learned and computes that with the same scorer the effect will be measured with; the wild row is what that circularity looks like when it closes. A guard computed with the scorer cannot certify that the scorer can see the axis the guard is about. The failure is contained, checked rather than assumed: numeral against spelled, Arabic against Roman and the synthetic +1 give 0/292, 0/276 and 0/342 mathematically equal pairs, and only the wild corpus fails. The repair costs no training: train on convention A alone, score that single model on both golds, and require $\mathrm{acc}_A/\mathrm{acc}_B \geq 2$ before reading any mixed arm, which is the separability precondition of §5.2. Applied to every row this paper publishes it passes at 16.6× to infinite, and the same-convention controls fail it as they must, which is the sign it measures the intended thing. It is now a preflight in this project, and we recommend it wherever a metric might canonicalise the axis under study, which includes any exact-match evaluation of mathematics, code or structured output.

So the cell stays open with a sharper specification (App. H): a wild conflict corpus must pair conventions that are not numerically equal, which excludes exact-form-against-decimal, 1,284 of the 1,284 real pipeline disagreements we found.

Read across this line of work it is not only a cost. The underdetermination paper(Anonymous, 2026b) states twice that the wall it derives is escaped only by writing the convention into the input. This run is a measured counterexample: maximal convention entropy given the input, no key anywhere, and no allocation cost at all, because the scorer accepts both forms. Canonicalising the metric is a third escape, and the premise doing the work sits inside that paper's proof rather than in its escape list, namely that the conventions be mutually exclusive under the metric. It carries the escape now, with the qualification that matters: it removes the penalty and not the disagreement, so it is available exactly when the conventions are interchangeable for whoever consumes the output. Numeric form is such a case; the conventions in Table 9 are not, which is why they still fire. The production-mixture reading remains a hypothesis with a measured base rate behind it, not a result this paper contains.

The firing rows are no longer one pretraining family, and the replication costs the ladder its generality. §5.4 shows sign-level conclusions reversing between Qwen2.5 and Qwen3 on these very skills, so by this paper's own standard a one-family switch would

not extrapolate, and the replication was registered against that objection before it ran (prereg/qwen3_switch_replication.md). On Qwen3-8B base, five seeds, same corpus and budget, separability checked first and passing at infinity: guard 0.0520, $D = -0.0875$ at **4.58** floors on $[-0.1038, -0.0712]$, all five seeds negative, control +0.0178 (0.93 floors), conflict minus control −0.1053. That is the registered S1 branch, and the switch now spans four scales in two families.

And the row costs us something we published. Amendment A4, frozen before any A_then_B file existed, recorded that Qwen3-8B's guard sits at Qwen2.5-3B's rung rather than 7B's. Our own ladder has switch magnitude tracking the guard, on the reading that the more learnable the minority convention the more there is to commit to, so it predicted a marginal result here, and A4 bound us to report the outcome either way without upgrading the ladder from one point. The prediction failed in the direction that costs us:

model	guard acc_B(B_only)	D	floors
Qwen2.5-3B	0.0459	−0.0263	1.38
Qwen3-8B	**0.0520**	**−0.0875**	**4.58**
Qwen2.5-7B	0.2091	−0.0985	5.16
Qwen2.5-14B	0.2579	−0.2037	10.67

At the 3B rung's loading the effect is 3.3× larger, and nearly the size of the 7B row at a quarter of its guard. The ladder is monotone within Qwen2.5 and this point is a counterexample to reading it as a law across families; guard is not the only axis, and we do not name the second one from a single point. Two statements survive: within a family, magnitude and guard move together, and the switch does not need a large guard to appear. Llama-3 sits below the power gate.

The boundary control has n=1. The firing rows carry two convention families and two scales, while the equally load-bearing "heterogeneity is not disagreement" null rests on one skill pair at one scale. A second cross-domain pair owes it the same replication standard, and until then that control is one clean instance, not a replicated result.

H The Remaining Missing Cells

This appendix lists the cells §6 names, in the order we would buy them, with the specification each carries.

First, pinning σ_{run}. §6 states what it decides: the negative result holds across the whole interval $[0, 0.025]$, and two secondary questions turn on it, whether the switch clears its own resolution and at which k the evaluation schedule crosses $\operatorname{diam}\mathcal{R}$. Two further executions of the whole protocol add six (pair, run) groups each, taking the count from thirteen to twenty-five and the run component from seven degrees of freedom to nineteen, and raising the degrees of freedom behind the evaluation share as well. It needs no new corpus, model or registration. This is the single number a reader deciding whether to act on this paper needs most.

Second, a second problem set, which is two cells and not one. Every quantity here is conditional on one fixed draw of 120 held-out problems per skill (§4), so $\operatorname{diam}\mathcal{R}$ is a property of the system and of that item sample, and $\operatorname{diam}\mathcal{R}$ is the numerator of the 0.76. A second draw needs no data preparation, because the harness has always read the first 120 rows of files that hold 200, and the conflict harness the first 200 of files that hold 276. It does not follow that it needs no training, and we printed that it did (App. J). Taking the inventory:

quantity	arms it needs	second draw costs
$\operatorname{diam}\mathcal{R}$, the 0.76's numerator	six pairs × two directions × three seeds	36 two-stage trainings
D, the Roman row and its control	four arms × five seeds	inference only
single-arm accuracy, the noise split	four skills × three seeds	inference only

Only the four _only arms of the volume-matched family were retained; the two-stage arms were reclaimed once their results were recorded, this project's standing policy. The free

half is registered and read out (prereg/second_item_sample.md, frozen before it ran, with the branch that demotes the Roman row from a replication of the switch to a single-item-sample result). It returned I-1 and that branch did not fire: on rows 200–275 the switch reads $D = -0.1494$, five seeds of five negative, control inert, within the design's resolution of draw 1 (§5.2), so the switch side no longer carries an item-sample caveat. The noise split was read too, and it is the more interesting half. Branch N-2 fired: draw 2's pooled cross-seed s.d. is 0.01829 against a matched draw-1 interval whose upper end is 0.01782. We do not report that as an item-sample effect, because the branch fires on 0.0005 in a quantity whose own 95% interval is $[0.0124, 0.0350]$, a threshold statement and not a resolved difference. What the twelve arms did buy is the observation in §3: on the matched four cells the estimated evaluation share is 3.03 at draw 1 and 0.99 at draw 2, and seven of those eight cells put it above one, which is outside the additive model rather than inside it. The N branch's registered comparison is between two point estimates of a quantity the model does not identify, so it is reported as fired and read for what it exposed rather than for what it was registered to decide. The paid half is not bought: it would be one full execution of the order protocol, which is also what the first cell needs, and the two should be bought together: one execution scored on both draws buys a run-component degree of freedom and an item-sample replicate for the price of the first. Until then the item-sample caveat stands on $\operatorname{diam} \mathcal{R}$, and no sentence about the 0.76 may be softened on the strength of the free half: I-1 is about D, and $\operatorname{diam} \mathcal{R}$ shares neither its arms nor its evaluation family.

Four others remain, named where they bind. Third, a conflict between two real public data sources, still the one whose failure would change conclusions rather than narrow them. That cell now carries a specification bought by the attempt that failed: the two conventions must not be numerically equal, since a metric that canonicalises scores both forms and no arrangement can move an axis the objective cannot see. This excludes exact-form-against-decimal, which App. G shows is every pipeline disagreement our survey found. Fourth, a second domain-boundary pair behind the "heterogeneity is not disagreement" control. Fifth, a $\geq$ 6-skill additivity test of the per-skill potential, pre-registered with an out-of-sample criterion and a withdrawal clause, since the in-sample test the appendix reports passes too easily. Sixth, registered and unrun rather than forthcoming, the plain-SGD trajectory control, which would separate a systematic optimiser contribution from Eq. (1). One further cell was open when a referee named it, was run rather than promised, and is no longer missing: the protocol-level replication of the switch, reported next.

And one instrument check, listed apart from the six because it buys no evidence for any claim here; it checks the instrument that prices all of them. §3 shows that this project's evaluation draws are not independent across seeds, since no driver passes a seed and every engine construction records seed= 0, so the estimated evaluation share exceeds one in 232 of 433 cells. The direct test is to re-score one family at k=4 with a different engine seed per run and compare the cross-seed dispersion against the shared-seed one: inference only, no training, one family, converting a code-and-log audit with two confirming predictions into a measurement. It is the cheapest cell listed here and the one we would buy first, because it prices the floor every other cell's verdict is read against.

I What This Paper Owns and Borrows, in Full

Four concurrent manuscripts of the same research programme are cited above, and the accounting matters because three of the four could be mistaken for support this paper does not have. The governing rule is that every number and every control this paper reports is in this paper or its appendices; no companion supplies evidence for any claim graded in App. J.

What this paper owns. The resolution–effect comparison and its variance decomposition; the six skill pairs and their order contrasts; the evaluation-sampling share of the order contrast's own variance; and the extent of the conflict switch — four scales, two pretraining families, three convention constructions, a dose axis.

What it borrows, and in which direction. (i) The channel-level mechanism of the conflict switch is A concurrent manuscript's (Anonymous, 2026); this paper owns the extent and cites

the mechanism for provenance and never as evidence, so the overlap does not read as two confirmations of one thing. (ii) The idealisation-first structure — state the limit in which the phenomenon is exactly zero, then measure departures — is Anonymous (2026)'s, a structural debt rather than a borrowed measurement. (iii) A concurrent re-survey (Anonymous, 2026a) corrects this paper: the larger pair we treated as independently labelled is one lineage, which costs us a pair and widens the genuinely external rate. (iv) One companion's escape list (Anonymous, 2026b) is contradicted here by a measured counterexample, not relied upon.

So of the four, one is provenance for a mechanism this paper does not claim, one is a debt of form, one takes a number away from us, and one is a disagreement. None of them is a premise. A referee holding only this paper can re-derive every graded claim in it.

J The Ledgers, and the Record of Conduct

Four totals, and why no one of them summarises another

This project keeps four itemised ledgers and no one of them summarises another, so a reader meeting four different totals should have them separated once. The scorecard grades 30 claims: the 29 we registered before the data that decided them, and one assumption underneath our own noise instrument, graded after it failed (§3): 19 supported, 7 failed, 2 untested, 1 mixed, 1 void. The machine checks (25) test the mathematics numerically before it is used, so a statement true only at convergence fails visibly; they say nothing about models. The errata (Table 18, 23 rows) list every reading we printed and withdrew, which is a history of this manuscript rather than of the experiments. The census (Table 19, 145 rows) counts every quantitative read-out for the multiplicity correction, and counts things the other three do not, such as individual cells of a fitted table. Separately from all four, App. D keeps the theory's record, signs five of five and locations zero of six, which is not a ledger of claims but of derivations and overlaps the scorecard in five rows.

And five different sixes appear in this paper, no two interchangeable: the six failed registered claims (§6); the six location predictions the theory lost, which is five of those plus one post-hoc test (App. D); the six skill pairs of the grid; the six controls §5.2 lists against a recency reading, of which two discriminate; and the six missing cells of App. H.

The machine checks changed the text rather than confirming it: they run gradient descent step by step against the closed forms with hard-coded tolerances, and four of the twenty-five rejected drafts of our own statements, so every theoretical statement in the main text is the post-rejection version. The four ledgers follow in that order, and then the pre-registration, which is a record of conduct rather than a ledger of numbers.

The public scorecard

Table 16 gives every pre-registered claim, its verdict, and the evidence that decided it, generated from the result files by a ledger script, so no row can be silently dropped between versions. The rows, statuses and values are the script's; the mathematical notation is applied by hand, and verdicts/converge_scorecard.json in the archive is the machine-readable form against which this table is checked. Rows whose instrument was rebuilt after a diagnosis say so in their evidence column.

Table 16. Public scorecard of every graded claim (19 supported, 7 failed, 2 untested, 1 mixed, 1 void). Four failed rows are quantitative localisation claims; the fifth is a mechanism reading this paper published and then refuted (§6); the sixth is this paper's own name for its coherent-domain result, refuted by a registered arm that recovers three quarters of the effect without interleaving; the seventh is the independence of evaluation draws across seeds, an assumption underneath this paper's own noise instrument rather than a registered branch (§3). One untested row needs a design not built and the other is a registered experiment whose control fired, so its registration permits no read-out; the mixed row is a claim whose controlled re-run split between its two skill pairs (§5.4); the void row is a registered noise branch whose two point estimates compare a quantity the additive model does not identify (App. H).

status	claim	evidence
failed	within-batch composition carries the conflict switch (batch purity, not stage order, is what blocking buys)	job 1725, 3 seeds, pre-registered share rule: the missing cell (PURE batches, RANDOM order) lands at acc_B 0.2242 against shuf 0.2179 and blocked 0.4454, so purity share = 0.03 and between-batch order carries ∼97% of the span. Control span +0.0037, unresolvable at $\kappa \approx 0$ as predicted
failed	r* identical for both pairs (factorised account)	3 seeds, r* 15.2 [12.2,24.4] vs 5.9 [4.7,8.4]; ratio CI [0.231,0.583] excludes 1.0
failed	r* ratio = m_hat ratio = 1.271 (CCH-style load accounting)	3 seeds, ratio 0.382, 95% CI [0.231, 0.583] excludes 1.271; direction also reversed (higher-demand pair crosses EARLIER)
failed	V collapses on measured committed capacity across families	matched cross-family pairs differ by 2.9x the noise floor; families separate more at higher committed capacity
failed	V collapses on skill-encoding DENSITY across families	the favourable read-out was a floor artefact: matched-pair median |dV| 0.0429 (1.83x) rises to 0.0680 (2.9x) once cells below acc_only 0.10 are gated out – indistinguishable from committed capacity (0.0688); within-family sign still reverses on all three density statistics
mixed	the rank-axis sign change is ATTRIBUTABLE TO A CAPACITY CONSTRAINT	job 1691, volume-matched axis (both arms 622 rows at every rank, 3 seeds). algebra-calculus, which carried the published swing from −0.165 to +0.080, keeps ONE sign once volume is matched (pooled V −0.013, crossing VANISHES). geometry-combinatorics fires the survives-rule on a seed-consistent +0.021 at r=128 (+0.006/+0.035/+0.021), but at 1.1 floors, under the $\mathrm{MDE}_{80} = 0.038$, and $> 10\times$ from the published $r^\star = 5.9$. SETTLED: the magnitudes were mostly the volume confound. OPEN: a small gate below this design's power
untested	joint training taxes identification of the partner's PRIVATE structure	fell out of the machine-caught correction to Theorem 2(b)
supported	a directional order effect exists, and prerequisite-first wins once data volume is matched	balanced level-2: 6/6 pairs measurable, all prerequisite-first (alg-calc +0.0743 = 3.9 floors); 5-seed ICC 0.524 (F=4.62); unbalanced sign was a volume artefact

continued on the next page

Table 16 (continued)

status	claim	evidence
supported	the arrangement contrast is 4x less reliable than the composition contrast	5 seeds: ICC ORD/D 0.197 (F=1.82 < F.05=2.00, n.s.) vs V 0.803 and FWT 0.820 (56 cells, kbar 3.34, 187 obs); the ordering also holds on MAGNITUDE (mean |V| 0.0416 > |ASYM| 0.0269 > |D| 0.0170, SEs 0.003/0.003/0.001); sign split 84/71, p=0.335 – no consistent direction
supported	V/FWT are the reliable observables	5 seeds: ICC 0.803 / 0.820, F 14.6 / 16.2 vs F.05 = 2.00
supported	parameter count is NOT the control parameter	opposite V signs at identical 14B across families (−0.117 qwen3 vs +0.092 qwen2.5, alg→calc); pooled spearman +0.084 on 187 triples is slice-dependent (+0.559 qwen2.5 vs +0.008 qwen3; +0.39 restricted to one direction) and is NOT what the claim rests on
supported	sign and magnitude load on DIFFERENT factors	sign scale 82.2%/pair 12.1% vs |V| 37.6%/21.1%
supported	|V| runs opposite to overlap (overlap = little new information)	4/4 cells
supported	effective rank cannot serve as C (Johnson-Lindenstrauss)	C_eff(f)/C(0)=0.979..1.027 with masks asserted; synthetic reproduction
supported	linear capacity map C(1-f) quantitatively falsified	no y* places peaks both close and in-window
supported	the sign gate replicates across seeds	THREE seeds: 11 of 12 (pair, rank) cells agree in sign; the one that does not is alg-calc r=16, the grid point adjacent to that pair's crossing (per-seed −0.010/+0.006/+0.031, mean +0.009 inside the noise floor) – where the crossing is
supported	Theorem 1 arrangement invariance: exact identity, machine-verified	13/13 checks; residual 1e-15 incl. on real GD solutions
supported	Theorem 2 capacity sign gate EXISTS at fixed model size	824: V flips sign in BOTH pairs, 11/12 cells across three seeds, on the published volume-UNmatched arms; job 1691 then reattributes the magnitudes – see the mechanism row
supported	V(r) monotone, no interior peak	no significant decrease survives the paired SEs
supported	Theorem 3 retention formula (contested-subspace localisation)	closed-form residual 1e-15; free data test in section B
supported	|D| grows when laziness is broken (higher lr)	|D| 0.0236 → 0.0507 → 0.0653 monotone across lr 1e-5/3e-5/1e-4; SE 0.0191; all three points at n=3
supported	RECENCY transfers per-skill (hub geometry)	volume-matched additivity 0.073 vs threshold 1.0 (contaminated gave 1.109); per-skill potential algebra > geometry > calculus comb fits all 6 ASYMs; seed-permutation self-check passed
supported	label conflict revives the arrangement effect (D<0)	redo with a pooled 864-row conflict, both budgets: conflict D = −0.1425 (7.5 floors) and −0.2275 (11.9 floors) against control +0.0017 and −0.0037; difference −0.2238 (8.3 floors). First attempt floored out and is kept as an apparatus failure

continued on the next page

Table 16 (continued)

status	claim	evidence
supported	the conflict switch survives an independent execution of its own protocol	job 2930, 48 trainings: two replicate corpora rebuilt from the same pool under different split seeds (51.7% and 49.2% overlap with the published halves); guard and separability pass on both. Replicate means −0.0985 (published, 8 seeds), −0.1258 and −0.0825, all six new seeds negative, controls inert at 0.26 and 0.04 floors. $\delta_{\text{run}}^{\text{switch}} = 0.087$ against the order cell's 0.144, so P1 fires and the P3 demotion branch does not
supported	the switch is not one pretraining family	Qwen3-8B base, five seeds, separability checked first and passing at infinity: guard 0.0520, $D = -0.0875$ at 4.58 floors, all five seeds negative against an inert control. Amendment A4, frozen while the job ran, named the guard-ladder counterexample in advance and it duly appeared
untested	the switch fires on a conflict drawn from real public data	job 2877 family: W3 fired, the control moving at 4.45 floors, so the registration permits NO read-out and the cell is not closed. Diagnosed afterwards: the two conventions the wild corpus pairs are ones the scorer treats as equal, so the contrast was unmeasurable by construction
failed	the coherent-domain advantage is interleaving rather than revisiting	job 4004, 36 trainings, registered before the corpora existed. A blocked arm substituting 25% of stage 1 into stage 2 (volume matched, $622 = 466 + 156$) revisits WITHOUT interleaving: +0.0369 of the +0.0483 gap, se 0.0065 over 18 paired cells, fraction 0.765; guards passed, 36/36 at step 117, drift 0.0000. Branch R-1, named in advance as the RENAMING branch. Follow-up (replay_perskill, exploratory): +0.0430 on the revisited skill, 114% of its part, 18/18, against −0.0061 on the skill diluted 25%, whose volume curve predicts −0.019. With the cross-execution offset the fraction is $[0.27, 1.25]$, covering the 0.75 bar; the first-skill term keeps its floor at $[+0.022, +0.064]$
supported	the switch is a property of the system and not of the one draw of held-out problems it was scored on	prereg/second_item_sample.md, frozen before the evaluation existed, branch I-3 named in advance as the demotion. Job 4193 rescored the same twenty checkpoints on rows 200–275 of the same pool, disjoint from the published draw. Branch I-1: $D = -0.1494$, five of five seeds negative, control inert at +0.0107, 0.0314 from draw 1 against a resolution of 0.0680. The floor was recomputed on draw 2's own per-problem vectors, rising 1.86× where the problem count implies 1.62×

continued on the next page

Table 16 (continued)

status	claim	evidence
void	the noise decomposition is a property of the system and not of the draw	branches N-1 and N-2 of the same registration, read after the twelve _only arms were completed. N-2 fires on 0.0005 in a quantity whose own interval is $[0.0124, 0.0350]$: a threshold, not a resolved difference. The registered comparison is between two point estimates of a quantity the additive model does not identify: the evaluation share reads 3.03 at draw 1 and 0.99 at draw 2, above one in seven of eight matched cells
failed	the evaluation term of the seed dispersion is independent across seeds, so raising k buys resolution	not a registered branch but an assumption under Table 6 and §5.3's k schedule. No driver passes an engine seed, so every run consumes the same random stream; the share exceeds one in 232 of 433 cells and the companion's k=16 re-scoring left its dispersion unchanged, 1.079 against a predicted 0.594. The k rows are re-labelled as statements about an independent evaluation, which is what a reader re-running the grid would face, and not about the dispersion measured here
FAILED 7, MIXED 1, SUPPORTED 19, VOID 1, UNTESTED 2		

The machine checks

Table 17 lists them per statement.

Table 17. Per-statement numerical checks.

#	statement checked	criterion	outcome
1	convergence map: GD limit = closed form	$\|\theta_{\mathrm{GD}} - \theta_{\mathrm{closed}}\| < 10^{-6}$	pass
2	arrangement identity on real GD solutions	residual $< 10^{-8}$	pass
3	$D = 0$ for orthogonal skills	$< 10^{-10}$	pass
4	$\|D\|$ grows with the principal angle	monotone in c	pass
5	capacity gate (b) as first drafted: selection alone gives $V > 0$	Monte-Carlo sign	rejected
6	per-skill transferability as first argued	split vs collapse	rejected
7	retention closed form	residual $< 10^{-8}$	pass
8	conflict drives D negative	monotone in κ	pass
9	$\Delta(T) \to P_W(I - P_A)$	residual 2×10^{-11}	pass
10	blocked never worse when Σ's commute	zero AM–GM violations	pass
11	interior maximum of $\|D(T)\|$	peak strictly inside	pass
12	two-term expansion, Eq. (1)	rel. residual $< 0.3\%$	pass
13	commutator/dissimilarity ratio linear in T	0.63 $\to$ 20.3 over T=1–32	pass
14	$\|g_D\| \le c\rho\|\delta\|$, constant sharp	0/400; ratio 1.000	pass
15	$c\rho < \gamma \Rightarrow \|g_D\| < \|g_V\|$	1355/1355; bites	pass
16	commutator norm $= \max_i \cos\theta_i \sin\theta_i$	0/200; 0 if nested	pass
17	$\|V\|, \|D\| \propto \varepsilon_B$, $\|\mathrm{ASYM}\|$ affine	300/300, res. $< 10^{-9}$	pass
18	the $\varepsilon^\star$ crossing lands where predicted	0/193 mis-sided	pass
19	driver norms alone imply the three-way ordering	3.4% of draws	rejected
20	$\mathrm{ASYM}(T)$ changes sign, $\eta < 1/\lambda_{\max}$	216/300 draws	pass
21	$D(T)$ changes sign	211/300 draws	pass
22	$T^\star$ ranks the crossing budget	Spearman +0.03, n=278	rejected
23	the crossing is a rotation of both bilinear factors	11.2° vs 11.7°	pass
24	a base-state score inverts rather than fades	28.7%, below chance	pass
25	the joint arm restores the inverted read-out	+62.7 points; oracle gap 26	partial

Errata: readings we printed and withdrew

Why this is a table and not eleven paragraphs. Every entry below is also disclosed where it applies, because a correction a reader cannot find at the claim is not a correction; but

by the fourth revision the in-place disclosures had begun to compete with the argument, so the history lives here and the body keeps the shortest form of each that still lets a reader who met the old digits recognise them. The one exception is Remark 1, a mathematical correction that stays where the statement is.

This is the second public version, and the table spans both. The first carried no errata table and graded 26 registered claims: 18 supported, 5 failed, 2 untested, 1 mixed. Nothing in that grading is reversed here; what changed is the audit around it. Three claims registered after that version went out are graded for the first time, a replay control and two robustness checks on a second draw of held-out items; and one assumption underneath the instrument itself is graded for the first time and fails. Of those four rows one is supported, two failed and one is void, which is why the scorecard reads 30 claims at 19/7/2/1/1 where the first version read 26 at 18/5/2/1. Some rows below withdraw readings that version printed and the rest withdraw readings from earlier drafts; we have not sorted them into two lists, because the first version stays on arXiv and a reader who wants that difference has it there rather than through our summary of it.

Table 18. Every reading this work printed and then withdrew, and what replaced it. Ordered by where the claim sits, not by when it was corrected. "Direction" says whether the correction moved the paper's case for or against itself.

where	as printed	what replaced it	direction	source
Prop. 1(i)	$\mathbb{E}[\hat{\Delta}] \geq \Delta + 2\sigma\sqrt{2\log m}$, an additive bound with an asymptotic constant	$\mathbb{E}[\hat{\Delta}] \geq \max\{\Delta, d_2(m)\sigma\}$, exact at finite m; the additive form is false for every $\Delta > 0$	against	review_variance
App. D	$\sigma = 0.028$, solved for so that the observed grid span came out exactly	a test: the span sits at $z = 1.10$ against the independently measured 0.0205	against	review_variance
App. D	"nothing in this paper derives a magnitude"	one magnitude is derivable and it misses: predicted 0.5 against a measured 0.413	against	theory_magnitude
§5.1	the coherent split as +0.0428 and +0.0132, which do not sum to the total	+0.0377 and +0.0107, recomputed so the parts sum to the whole	neutral	coherent_ decomposition
§5.1	"interleaving beats blocking"	"revisiting the first skill beats not revisiting it"; a registered replay arm separated them	against	replay_verdict
§5.1	replay's increment is volume matching, which the rehearsal literature does not do	Ibrahim et al. (2024) match by substitution as we do; the increment is the interleaved comparator and the per-skill split	against	—
§5.1	replay recovers 76.5%, quoted with the paired s.e. alone	the numerator spans two executions, so the interval is $[0.27, 1.25]$	against	replay_perskill
§5.3	run-level spread is a property of the cell, 4.6× between cells	one common variance is not rejected once the selected cell is removed, $p = 0.289$	against	review_variance
§5.3	$\delta_{\text{run}} = 0.144$ from the selected cell	0.081 pooled over all 39 cells; 0.144 is retained as a statement about selection	neutral	review_variance

where	as printed	what replaced it	direction	source
§5.3	the run-to-run excess 0.0364 against 0.0191 as an AdamW variance channel's residue	no such excess survives the fit; the optimiser argument rests on the per-seed unanimity test alone	against	review_variance
§5.3	the evaluation share of $\hat{\sigma}_{\text{seed}}$(ASYM), borrowed at 0.861 from single arms	measured on ASYM at 0.955, with its own interval $[0.51, 1.00]$	for	asym_eval_share
§5.3	at the share's lower limit "no finite k crosses"	false; the crossing needs k in the hundreds and $k \to \infty$ clears by 2.4%	against	asym_eval_share
§5.3	three measurements bound the optimiser channel	two; the third was the withdrawn 0.0364	against	—
§5.2	four controls separate the switch from a recency reading	two of the four discriminate; the list's annotations demote the others	against	—
§5.3	the switch's ratio printed as $\lvert\Delta\text{SHARE}\rvert/\delta^{\text{switch}}_{\text{run}}$	numerator and both variance components are D, an accuracy contrast; the ratio 1.28 is unchanged, the symbol was wrong	neutral	review_variance
Prop. 4	the ordering of the two ratios is σ-free, bounded in $[1.65, 2.40]$	that bound assumes one shared run component, which §5.3 argues against; the unconditional statement is the corner check, margin 0.98 against 0.82	against	resolution_ratios
§6	the recipe verdict is "settled without identifying the component"	free of σ_{run}, but $\hat{\sigma}_{\text{seed}}$ was held at a point in both denominators; carrying it, $\Pr(\rho > 1) = 0.091$, and 0.028 with selection undone	against	ratio_uncertainty
Table 5	the switch block's dispersions labelled ΔSHARE	they are fitted on D: a two-accuracy contrast, not an allocation share	neutral	review_variance
App. H	the second-problem-set cell is inference-only and costs no training	true of the conflict rows and single-skill arms, false of $\operatorname{diam}\mathcal{R}$, whose two-stage arms were reclaimed: that half is 36 trainings	against	—
abstract	"86–96% of the variance is evaluation sampling", read as one interval	two different quantities, 0.861 on single-arm accuracies and 0.955 on the order contrast, which must not be read across	against	asym_eval_share

where	as printed	what replaced it	direction	source
§3, §7	the seed term's evaluation share read as an identified split, and "raise k before a fourth seed" as unconditional	no driver passes an engine seed, so every run consumes one shared stream and the estimated share exceeds one in 232 of 433 cells; the split now prices an independent evaluation and the advice is conditional on passing one	against	shared_eval_ stream
§5.2, §7	the 92,359-problem pair called independently labelled, its near-zero convention rate read as a base rate for independence	one lineage: OpenR1's gold field is NuminaMath's own extracted answer, math_ verify-filtered, so the rate is partly the filter's	against	wild_conflict_ survey
§5.2	"the dominant pathology of real mixtures is noise rather than contested convention", unscoped	scoped to one lineage and its duplicates; on the external pair the companion's ladder brackets form at 26.7–29.2% against 2.06% value, the opposite ordering	against	wild_unproved_ axes
abstract, §2	"we could find nobody removing" the variance an ablation fights	Hochlehnert et al. (2025) report the same sensitivity in this domain and answer it with a standardised protocol; the claim is now the narrower one, that nobody prices a training seed against an evaluation sample	against	—

Twenty-four entries: nineteen moved the paper's case against itself, four were neutral, one was in its favour. We report the tally because a corrections table with a favourable balance would deserve suspicion, and because the one favourable entry is the one we went looking for after a reviewer asked whether a borrowed constant transferred. An earlier version of this sentence said twelve, one and one, which does not add up to the table above it.

Census of every quantitative read-out

Counting rule: one read-out per contrast, gate, or test statistic for which the paper states a value against a floor, a σ, or a pass/fail criterion, including read-outs reported only in summarised form (the capacity grids), and excluding guards used solely as admission gates and the machine checks of App. J, which test the mathematics rather than the models.

Pre-registration, in Three Tiers

A reader who wants only the standard can stop after this sentence: ten criteria were frozen before the data they judge existed, eight have been read out, two are registered and unrun, and every branch that fired is reported including the ones that cost us a claim. The rest is the account. Tier 1, frozen before the deciding data existed: the volume read-out, the additivity criterion and threshold, the conflict direction $D < 0$, the three anticipated outcomes of the matched capacity axis, and the order predictions with their abstention threshold, whose sign reversal is why §5.3 reports a falsification rather than an adjustment. Tier 2, redesigned after a diagnosis and re-frozen: the conflict instrument (the first build, 238 rows at one

Table 19. All 145 read-outs, grouped. At $\alpha = 0.05$, Bonferroni over 145 puts the two-sided threshold at $z \approx 3.58$; every firing conflict row with three seeds clears it. The count has been re-run twice, $94 \to 138 \to 145$. Of the 44 added first, 15 are the cells of one resolution table, a single fitted model evaluated at five evaluation budgets and three readings of one variance component rather than 15 independent tests; the 7 added second are intervals and probabilities on ratios already counted above, so they are the clearest case of the same conservatism. Counting them individually errs in the safe direction, but a reader should not read 145 as 145 independent chances to be wrong.

family	count	where
reliability F-tests, one per observable	6	Table 1
fixed-volume contrasts, published and matched	6	Table 3
volume doubling, four skills and their mean	5	Table 4
directional order, six pairs at two budgets	12	Table 7
in-run sign change, one per trajectory pair	4	§5.3
replication contrasts (endpoint re-runs)	3	§5.3
switch protocol replication, two independent executions	2	§5.3
additivity tests, matched and unmatched	2	§5.3
conflict-program rows (13 with a D, one excluded by its guard)	14	Table 9
conflict-minus-control differences	2	§5.2
allocation totals, three conflict corpora	3	§5.2
block-length sweeps, two corpora	2	§5.2
capacity cells, published and volume-matched	24	§5.4
crossing-location ratio	1	§5.4
scalar sign-axes, matched cross-family $\lvert\Delta V\rvert$	4	§5.4
premise $c\rho < \gamma$ (12 pairs $\times$ 3 dimensions $\times$ 4 depths, summarised)	1	Appendix E
$T^\star$ ranking tests, in-model and on-model	2	§6
learning-rate slope of $\lvert D\rvert$	1	the scorecard, App. J
coherent-domain gap split by skill (first, second, total)	3	§5.1
replay recovery, six pairs and their pooled mean	7	§5.1
replay recovery split by skill, and the volume prediction for the diluted one	3	§5.1
evaluation share of ASYM, measured against the borrowed value	1	§5.3
δ_{run} against evaluation budget (3 run-component readings $\times$ 5 budgets)	15	§5.3
resolution ratios, two sides $\times$ three run-component readings	6	§5.3
nested variance components, order side and switch side	6	§5.3
run-mean range consistency test against $d_2(6)$	1	§3
homoscedasticity likelihood-ratio tests, with and without the selected cell	2	§5.3
seed-component intervals, order side and switch side	2	Table 5
joint bootstrap of the two ratios and of their ordering	3	§5.3
selection bias of the numerator: null expectation and shrunk ratio	2	§5.3
total	145	

epoch, never made the model adopt the convention, was graded untested and rebuilt at 864 rows and three epochs); the reliability estimator (the first reported exactly 0 where reliability was low but nonzero, and an independent recomputation caught it, 0.197 being the corrected value); and the capacity control. No conclusion rests on a Tier-2 number whose redesign was aimed at that conclusion.

Tier 3, and the reason it had to exist. A defence of a design history is still made by people who have already seen the data; only a criterion frozen before data that does not yet exist removes the possibility that iteration shaped the narrative. Ten are on file, eight read out and two registered and unrun (the $\geq$ 6-skill additivity test, prereg/potential_6skill.md, and the plain-SGD trajectory control, prereg/sgd_trajectory_control.md). Of the eight: run-to-run variation over the six skill pairs (prereg/run_variation_six_pairs.md) read out as §5.3's two resolutions, and the branch that fired cost us the constant we had been quoting; the seed extension of the natural row (prereg/nat_seed_extension.md) fired R2 and cost us a capability claim; the synthetic control's extension (prereg/seed_extension_ctl2_roman.md) is complete on one row and stopped early on the other, reported at five seeds with the deviation and its date in that file; the second-family switch (prereg/qwen3_switch_replication.md) returned S1, and its Amendment A4, frozen while the job ran, named the loading confound in advance and is why §5.2 prints the ladder that predicted wrongly beside the row that replicated; the switch on an unconstructed corpus (prereg/wild_conflict.md) returned W3, an instrument result rather than a row, and a defect of our own construction; and the protocol-level replication of the switch (prereg/switch_protocol_replication.md) returned P1, its P3 branch having been committed in advance to demoting the switch in this abstract had any replicate mean fallen within a floor of zero; the replay disambigua-

tion (prereg/replay_disambiguation.md) returned R-1 and renamed this paper's coherent-domain result, its branch committed in advance to exactly that consequence (§5.1); and the second item sample (prereg/second_item_sample.md) returned I-1 on the switch branch and N-2 on the noise branch, the second graded void because the quantity its two point estimates compare is one the additive model does not identify (App. H). We report the release, the runs and the defects beside the registrations that fired, because a pre-registration whose inconvenient instances go unmentioned is not one.