The Underdetermined Stream: What a Single-Gold Score Actually Measures on Contested Material

Anonymous authors
Paper under double-blind review

Abstract

Benchmarks for mathematical reasoning compare a model's answer with a single gold string, and on material where several forms of the same answer are equally correct that comparison is measuring something other than capability. The field treats the mismatch as a data-quality bug for a better normaliser to repair, which presupposes that one of the two golds is wrong; when both are correct there is nothing to fix, and what the score measures instead has never been derived. The obvious remedy — publish a correction factor beside the score — presupposes that the multiplier is a property of the corpus, and latent-variable estimators of a hidden true label cannot price their own error. We derive the identity: an exact-match score on contested material is capability multiplied by the rate at which the model writes the one form this corpus shipped as gold, a product whose second factor belongs to the model–corpus pair rather than to either alone. Capability is nonetheless exactly recoverable without retraining, by scoring the same generations against each convention's gold in turn and summing, under two premises that are measured rather than assumed. On four arms of one corpus differing only in the order training data reached the optimiser, capability holds inside the 0.0270 measurement floor while the allocation moves 0.4621 — the policy at F=612.18 against the capability sum's 1.63 — so a single-gold benchmark reports the same capability 4.75× apart depending only on how the model was taught to write. The estimator returns $\hat{C} = 0.4452$ against a directly measured 0.4522, and that 0.0070 gap is exactly the support leak the identity predicts (0.00705), so its error is always downward and always priced by a quantity the same run reports. The correction-factor shortcut is registered, run and rejected: the model's policy correlates with per-item solvability at $r = -0.257$ on 24 of 24 arms, so the multiplier is a property of the model as well, and the effective enumeration it hides runs 1.149 to 1.811. The rule is to publish the summed score, which costs extra scoring passes over generations already produced and no new training.

1 Introduction

A benchmark that compares a model's answer with a single gold string is measuring something. On material where several answer forms are equally correct — 0.5 against 1/2, XIV against 14 — that something is not capability, and this paper derives what it is instead.

The situation is not a bug. When a corpus ships 1/2 as the gold and a model writes 0.5, no party is wrong: the value is right and the form is not the one this corpus chose. Practice treats such cases as normalisation debt, to be paid down by a better canonicaliser, and that framing carries a presupposition — that one of the two strings is incorrect — which is false exactly when the disagreement is about form rather than value. A concurrent survey of public mathematics corpora finds form disagreement several times more common than value disagreement (Anonymous, 2026d), so the case is not a corner.

What has never been derived is what the score becomes. This paper's first contribution is that identity. Writing C for the probability the model solves the problem at all and $\bar{\pi}_{\text{gold}}$ for the rate at which it writes the one form the corpus shipped, an exact-match score on

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.

contested material is exactly $C \cdot \bar{\pi}_{\text{gold}}$ (Proposition 1). The second factor is a property of the model–corpus pair. It is not a property of the model, so it does not transfer between benchmarks; it is not a property of the corpus, so it cannot be published as a correction; and it carries no term at all for how the model resolves the underdetermination, which is the one behaviour the material leaves free.

The sharpest form of the obstruction is a limiting case, and we are explicit that it is one. When k conventions are equiprobable and drawn the same way at every problem, the score is not bounded by C/k but pinned at it (Proposition 2): every policy receives the same number, and two enumerations of different size report scores whose ratio is unbounded at identical capability (Corollary 1). Those hypotheses hold on no public benchmark we know of, since a benchmark presents each problem once under one gold — the degenerate case, as far from uniform as the simplex allows — and no measurement in this paper rests on them. What survives on real corpora is the identity above and the recovery below.

The remedy is cheap and exact. Because the conventions partition the correct answers on each problem, scoring one generation pass against each convention's gold and adding recovers C exactly, whatever the policy and whatever the corpus's convention distribution (Proposition 3). It needs no homogeneity assumption, no knowledge of how a corpus distributes conventions across problems, and no estimate of k. It needs the enumeration and extra scoring passes over generations already produced. Both of its premises — exclusivity and support — are measured here rather than assumed, and the second, measured at $\sigma = 0.9844$, turns out to account for the estimator's entire empirical residual.

Contributions. (i) The identity $S = C \cdot \bar{\pi}_{\text{gold}}$ and the pin that follows, with the incomparability corollary they license (§3). (ii) An exact capability estimator needing no retraining, whose two premises are measurable and whose error is one-signed and priced by the same run (§4). (iii) A measured demonstration that capability and allocation are separately real: four arms of one corpus, the identical training multiset in four orders, hold capability inside the measurement floor while the allocation moves 0.4621 (§6). (iv) The correction-factor shortcut a reader would reach for first, registered, run and decided against on a measurable condition (§6).

What this paper does not claim. Prevalence of mixed-gold benchmarks is the survey's (Anonymous, 2026d), and any causal role of training arrangement is a concurrent manuscript's (Anonymous, 2026b); both are cited for provenance, never as evidence. Appendix G is the full owns-and-borrows accounting.

2 Related Work

Answer matching and its normalisers. A large body of evaluation practice is devoted to making exact match behave: canonicalisers, equivalence checkers and model-based graders (Hendrycks et al., 2021; Lewkowycz et al., 2022; Kydlíček, 2024; Bulian et al., 2022). Every one is a scorer-side response, and Proposition 5 places it precisely: it removes the penalty exactly when the conventions are interchangeable for whatever consumes the answer. Where they are not — a parser expecting a decimal, a diff against a reference implementation, an option letter that must be reported — the escape is unavailable and the identity still applies.

Latent true labels, and what a benchmark number means. The statistical literature on recovering a hidden label from noisy observations (Dawid & Skene, 1979; Collins & Huynh, 2014) performs an inference close in form to ours, but a latent-class estimator's accuracy depends on unverifiable assumptions about the noise process; the estimator here is an identity whose only failure mode is support leakage, always downward and priced by the same scoring pass — that property, not the estimate, is what we claim. Construct-validity work argues leaderboard numbers often fail to measure what they are cited for (D'Amour et al., 2022; Schaeffer et al., 2023; Holtzman et al., 2021); this paper contributes a case where the gap is not a validity worry but an algebraic identity, written down, measured and removed.

Position in a programme. Four concurrent manuscripts of the same research programme are cited above: the idealisation-first structure from the parent preprint (Anonymous, 2026a), the leaderboard evidence from the survey (Anonymous, 2026d), the training-side marker results from a third manuscript (Anonymous, 2026e), and the four fixed-capability arms of Table 1, re-scored here rather than re-trained, from a fourth (Anonymous, 2026b). Appendix C extends every grouping above.

3 The Identity, and What a Single-Gold Score Is

Definition 1 (The underdetermined stream). A problem x is underdetermined for a scorer if two or more strings are equally correct answers to x and the scorer credits only a subset of them. Write $I(x)$ for everything the model is shown — the problem statement and any instruction — and c for the convention a correct answer is written under. The stream is underdetermined when $H(c \mid I) > 0$: the convention is not a function of what the model was shown.

Definition 2 (The two demands). A scored generation must satisfy two demands at once. The value demand is to solve the problem, and we write $C = \Pr[\text{value correct}]$ for the rate at which the model meets it. The form demand is to write the answer under the convention the scorer credits; we write $\pi(c \mid x)$ for the model's policy over conventions given a solve, and $\bar{\pi}_{\text{gold}} = \mathbb{E}_x[\pi(c_{\text{gold}}(x) \mid x)]$ for the rate at which the policy lands on the gold form. Nothing in $I(x)$ determines c_{gold}, so no amount of capability makes the form demand easier.

Proposition 1 (What a single-gold score is). Let the model solve x with probability $p(x)$ and let $\pi(c \mid x)$ be the chance it writes its answer under convention c given that it solved x — its policy, not assumed equal to anything the corpus does. A benchmark presents each problem once under one gold $c_0(x)$. Then, writing $C = \mathbb{E}[p]$,

$$S \;=\; \mathbb{E}_x\big[\,p(x)\,\pi(c_0(x) \mid x)\,\big] \;=\; C \cdot \bar{\pi}_{\text{gold}}, \qquad \bar{\pi}_{\text{gold}} \;=\; \frac{\mathbb{E}\big[\,p\,\pi(c_0 \mid x)\,\big]}{\mathbb{E}[p]} \;\in\; [0,1].$$

Nothing above assumes the corpus's convention law is uniform, homogeneous across problems, or related to π.

The display costs no independence hypothesis. Because π is defined given a solve, solving and writing form a chain rather than a product of independent events, so the factorisation is the chain rule alone: a model whose chance of solving a problem is bound up with the form it is about to write is inside the statement, not outside it. An earlier version of this work carried such a hypothesis and it is withdrawn as never needed (Appendix D). Equally, π must not be set equal to the corpus's law on the ground that a loss-minimising model answers at the corpus rate: that is false on our own material, where four arms trained on one corpus at one convention rate span 0.4621 of the unit interval (§6). The second factor is thus a property of the model–corpus pair — it depends on the model through π and on the corpus through its choice of gold, and on neither alone. Appendix A carries the proof and the two-sided form for a scorer that draws its gold at random. Two things follow immediately. First, $S \le C$, with equality only if the model always writes the corpus's form — so a single-gold score is a lower bound on capability whose slack is not noise but a second behaviour. Second, the slack is invisible from the score alone: no function of S separates a capable model writing the wrong form from a less capable model writing the right one.

Proposition 2 (The uniform case, in which the score is pinned rather than bounded). Suppose k conventions are equiprobable and unpredictable from I, the corpus draws its gold uniformly among them, and the model's output is supported on those k conventions. Then $S = C/k$ for every policy π: the policy cancels.

C/k is not a ceiling a good policy might approach but the value every policy receives, which is the cleanest form of the obstruction. Its domain is empty on real data, and both hypotheses are load-bearing. A benchmark presents each problem once under one gold, the degenerate case $P(c_0 \mid x) = 1$; homogeneity fails whenever a corpus's gold form tracks anything about the problem, one leg of which §6 measures on our own material; and if the policy puts mass outside the k accepted conventions the inner sum falls below $1/k$. We retain C/k for one

reason: it is the form in which a constructed enumerating scorer can be built and checked, and the k-sweep of §6 is that check. What survives on real corpora is Proposition 1, which needs neither hypothesis, and Proposition 3, which needs neither and is the remedy.

Corollary 1 (Unbounded incomparability, inside an enumerating scorer). Hypotheses as in Proposition 2, which no public corpus in this paper satisfies. Two constructed enumerating scorers over the same problems, with enumerations of size k_1 and k_2, report C/k_1 and C/k_2 for identical capability. Their ratio is k_2/k_1, bounded only by the enumerations, so two scores can differ arbitrarily while the systems they score are indistinguishable in capability.

Corollary 3 in Appendix A states the consequence for a deployed exact-match benchmark, and Corollary 2 bounds the gain available to a model that optimises the form demand instead of the value demand.

Two escapes from the underdetermination, and they are disjoint. Proposition 5 (Appendix A) states three interventions and separates them by the object they act on. Only two remove the underdetermination, and they do not overlap: write the convention into the input, so that $H(c \mid I) = 0$ and the form demand becomes answerable; or widen the scorer to credit every correct form, so that the form demand is not charged. The first changes the information structure and the second the distortion measure, and no route inside the model converts one into the other; the third intervention raises the learner's capacity, which leaves Proposition 1's bound untouched because the bound does not mention it. The first escape carries a second condition that is not a restatement of the first — the identifying function must be realisable by the hypothesis class — and a concurrent manuscript reports a base checkpoint that cannot reliably detect a question mark, a deterministic function of its input (Anonymous, 2026e). Canonicalising the metric is the deployable case and its scope is exactly downstream interchangeability (§2).

The scorer escape works, and where it has already been applied the phenomenon is absent. A concurrent manuscript built a conflict corpus from the commonest real disagreement between public mathematics pipelines, trained forty arms on it, and measured no allocation effect at all: the installed share sat at 0.500–0.506 in every arm of both the conflict and its control, because all of its evaluation gold pairs are equal under the shipped checker (Anonymous, 2026c). That corpus has $H(c \mid I) > 0$ by construction and no key; the wall did not appear because the metric had already dissolved it. This is a borrowed measurement and it constrains us rather than supporting us: it marks the region where nothing here applies.

4 Recovering Capability Without Retraining

Proposition 3 (Capability is recoverable, and the recovery needs no homogeneity). Let the accepted conventions be mutually exclusive on each problem, so that one generation matches at most one gold, and let the model's output be supported on the enumeration. Then

$$\sum_c \mathrm{acc}_c \;=\; C,$$

where acc_c is the exact-match accuracy of one fixed generation set scored against convention c's gold. The identity holds for every policy π and for every corpus convention distribution $P(c \mid x)$; neither appears in it.

The procedure this licenses is the paper's practical output: score one generation pass against each convention's gold and add, reporting the sum alongside the single-gold number whenever the material admits more than one correct surface form. It needs no homogeneity, no knowledge of $P(c \mid x)$, and no estimate of k. It needs the enumeration, and extra scoring passes over generations already produced.

The sum's honest name is union accuracy, and we use it. $\sum_c \mathrm{acc}_c$ is the probability the model produces a correct answer in one of the enumerated forms; a correct form nobody enumerated scores zero, so the quantity is relative to the enumeration in the way an exact-match score is relative to one gold, only less so. We keep the word capability because the argument concerns the gap between what a benchmark reports and what a model can do,

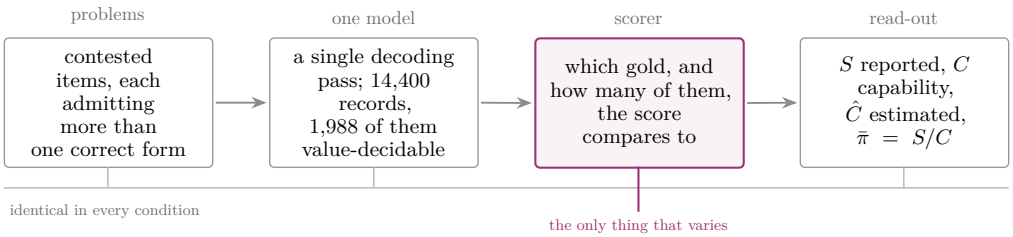


Figure 1. One generation file, read several ways. The generation pass is fixed across every condition; only the scorer's gold varies. Reading the same file against each convention in turn and summing recovers C exactly (Proposition 3); Appendix B extends the same chain to the numbers it yields on this corpus.

and because the remaining gap is measured rather than left to the imagination. Nor is the algebra claimed to be deep: Proposition 3 is a pointwise indicator identity and Proposition 2 the law of total expectation. What the paper contributes is the procedure that recovers the quantity together with its measured error, and the demonstration that the coordinate it recovers is separately real.

Both premises are measurable, and one of them is the whole error term. Exclusivity holds by construction of the enumeration: two distinct canonical renderings of one value cannot both match a generation. Support is an empirical quantity — a correct answer written in a form outside the accepted set scores zero against every accepted gold — and we write σ for the fraction of correct answers the enumeration covers. The identity then predicts the estimator's shortfall exactly, at $(1-\sigma)\,C$, which is always downward and computable from the same scoring pass. §6 measures σ and checks that prediction against the residual.

Why this is not the correction factor a reader would reach for. Holding a published score and unable to rescore, one would define $k_{\text{eff}} = C/S$ and report $S \cdot k_{\text{eff}}$. Under homogeneity — the corpus's gold convention independent of the problem — k_{eff} is a property of the corpus and can be published beside its score. Off homogeneity it is a property of the model as well, and then no corpus-level number exists to publish. Homogeneity fails only if both sides correlate with a common driver: the corpus's gold must vary with x, and so must the model's policy. The model's side is directly measurable, and §6 measures it.

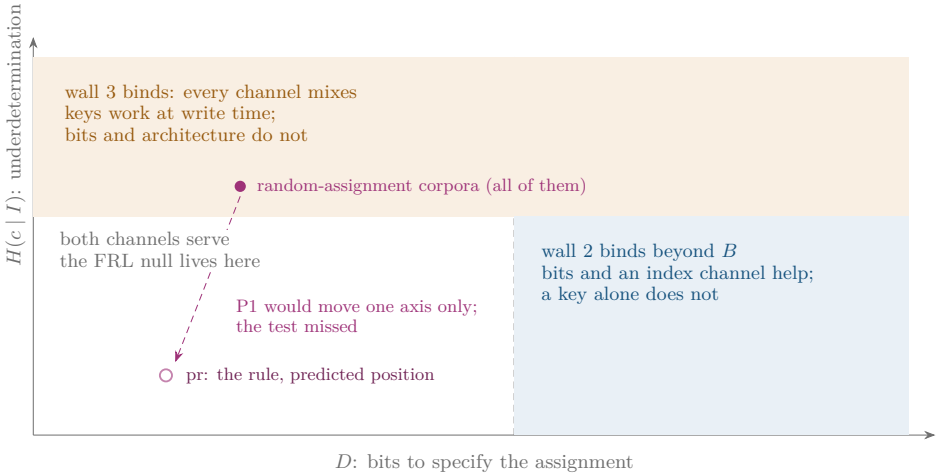


Figure 2. Two axes with disjoint escapes. Which intervention can fix a system failing on contested content, read off two measurable properties of its corpus. The routing consults $H(c \mid I)$ and downstream interchangeability, and neither D nor B; every filled point is measured, the hollow one is the position this paper does not measure. Appendix B runs it as a procedure.

5 Experimental Setup

Material. Conflict corpora over competition mathematics in which the two halves answer under incompatible but equally correct conventions and nothing in the prompt marks which is wanted. The constructions are a natural pair (exact form against decimal), an Arabic/Roman numeral pair, and a synthetic offset; they are shared across this research programme and the sharing is disclosed here rather than at review. Models are Qwen2.5 at 3B, 7B and 14B and Qwen3 at 8B, giving a 28× parameter range across two pretraining families.

Two resolution constants, defined once. Differences are priced against one of two floors and they are not interchangeable. The contrast floor 0.0191 prices a difference between arms at three seeds and is what every Δ is quoted against. The capability floor 0.0270 prices the sum $C = \mathrm{acc}_A + \mathrm{acc}_B$, which is a sum of two accuracies and therefore noisier than either. A quantity is compared against the floor for its own kind, never the smaller of the two.

One protocol property, disclosed rather than inherited silently. None of this programme's evaluation drivers passes a seed to the inference engine, so every run consumes the same sampling stream and evaluation sampling is common-mode across seeds — a property found by a concurrent manuscript (Anonymous, 2026c) after these floors were frozen. In a cross-seed contrast, which is what both floors price, that component largely cancels, so the floors are conservative and every "clears the floor" statement below stands. An absolute accuracy carries the component in full, so any seed-spread interval on an absolute level understates evaluation-sampling uncertainty and is read only for separations between conditions scored on the identical stream.

Registrations. Eighteen read-out rules were registered before the runs they decide, with all branches of each named in advance; branches within a registration are mutually exclusive and exhaustive, so the family is 18 decisions rather than 18 significance tests. Of the 15 read out, 5 returned a branch their own preamble did not expect and 5 declined to report at a guard. Appendix E is the index.

6 Results

6.1 Capability and allocation are separately real

Figure 3 draws what the reported line hides on three corpora: the grey bar is the capability sum C, the dashed line is $C/2$ — what an exact-match benchmark reports for these checkpoints and, by Proposition 2, for any others at the same C whatever their policy — and the two points are the measured per-convention accuracies. Each pair straddles the line rather than sitting on it, and the straddle is the free coordinate. Checkpoints scoring 0.20 and 0.18 solve 0.385 when either convention is accepted. Across every conflict corpus this programme has trained, the allocation runs 0.130 to 0.448 over the 28× parameter range, two pretraining families and three constructions, and no exact-match score reports that difference at all.

Those rows differ in C, so they do not on their own exhibit a statement made at fixed capability. The clean exhibit needs four arms of one corpus (Table 1): the identical multiset of training examples in four different orders, same model, same budget, three seeds.

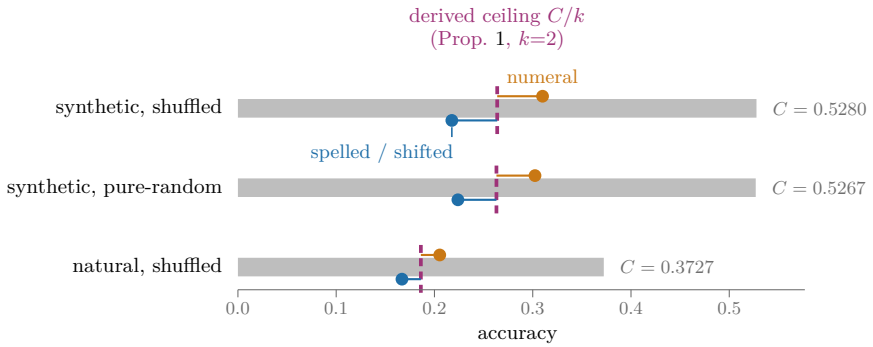


Figure 3. Three corpora, and the quantity the reported line hides. Grey bar: the capability sum C. Dashed line: $C/2$, what an exact-match benchmark reports. Points: the measured per-convention accuracies. The displacement ranges from 0.0195 here to 0.1040 at 3B and the dashed line does not move with it. Qwen2.5-7B, three seeds; corpora from a concurrent manuscript (Anonymous, 2026b), re-scored here.

Table 1. Capability held inside one floor while the allocation moves half the unit interval. Four arms of the same conflict corpus, differing only in the order the training data reached the optimiser. The measured columns are acc_A and acc_B; the fifth is $C/2$, which is arithmetic rather than a second measurement. The content is the first three columns: at a capability held inside the 0.0270 floor the allocation moves 0.4621. Capability is stated with the design's minimum detectable difference, as unresolved rather than shown equal. Source: cf2b3lr3e5.

arm (same corpus, same model, same budget)	acc_A	acc_B	C	implied $C/2$	SHARE
L=27 alternation	0.3183	0.2183	0.5367	0.2683	0.407
shuf (mixed within batch)	0.3100	0.2179	0.5280	0.2640	0.413
pure batches, random order	0.3025	0.2242	0.5267	0.2633	0.426
blocked (A then B)	0.0671	0.4454	0.5125	0.2562	**0.869**
range across the four			0.0242	**0.0121**	**0.4621**

A range is not a test. The four arms crossed with three seeds give a one-way repeated-measures comparison at $F_{\text{crit}}(3, 6) = 4.757$, and it separates the two quantities cleanly: the allocation at $F = 612.18$, a factor of 377 over the capability sum's $F = 1.63$, which does not reach the critical value. Two cautions belong with that. The reported score is $C/2$ identically, so it inherits C's F rather than supplying a third measurement; and the design's minimum detectable pairwise difference in C is 0.0242, the size of the observed range, so capability is not resolved here rather than shown equal. Whatever difference lies below that resolution moves the reported score by at most 0.0121, under the contrast floor, while the allocation moved 0.4621. Four arms that solve the same fraction of problems to within the floor, one of which answers under the minority convention twice as often as another, are reported by a mixed-gold exact-match benchmark as one system.

The range replicates on a ten-arm ladder. Four arms are four arms, so we recomputed the allocation across a block-length ladder from this paper's own result base rather than quoting one: at $L = 1, 3, 6, 9, 18, 27$ together with shuf and purerand the share runs 0.407 to 0.449, a span of 2.2 contrast floors, while $L = 54$ sits at 0.492 and the blocked corner at 0.869, which is 22.0 floors above the interior (Appendix F). The interior of the ladder is flat to within a couple of floors and the corner is not, which is what Corollary 4 predicts: the training path cannot enter the conditional law of the convention given a held-out query, so what a path moves is the realised policy and not the bound it is measured against.

6.2 The estimator, and its error term

Proposition 3's support premise is measured on 1,988 generations whose gold is a non-integer rational with four distinct canonical renderings, with the gold drawn uniformly among them. At the full enumeration $\sigma = 0.9844$ and $S(4)$ sits within 1.6% of $C/4$: the premise holds and Proposition 2 is not an idealisation.

The estimator returns $\hat{C} = 0.4452$ against a directly measured 0.4522. The weaker reading is that the 0.0070 gap sits inside the 0.0270 capability floor and is therefore unresolvable — true, but it leaves open whether the estimator carries a bias of unknown origin. It does not. Support is the one premise that can fail, so the shortfall the identity predicts is $(1-\sigma)\,C = 0.0156 \times 0.4522 = 0.00705$, against a measured 0.0070. The residual is not slack; it is the support leak, and the two agree to the last place either is reported in. The identity is exact, the procedure is exact up to the fraction of correct answers the enumeration covers, and that fraction is measured rather than hoped for. What a reader gets is a rule and not a caveat: the estimator's error is $(1-\sigma)C$, is therefore always downward, and is priced by a quantity the same run reports — the property no latent-class estimator of §2 can offer.

What one gold costs, on a named population. The corpus's own single gold form reports 0.3189 while the model is value-correct 0.4522 of the time: the model wrote the corpus's chosen form in 70.5% of the answers it got right, and the benchmark charged it for the other 29.5%. That is a 1.42× understatement of capability from the choice of gold form alone.

The larger factor is ours to exhibit and the largest is not. A benchmark whose single gold is convention A reports the four arms of Table 1 at 0.3183, 0.3100, 0.3025 and 0.0671 — the same capability, held inside the floor, reported 4.75× apart depending on how the model was taught to write. A concurrent survey (Anonymous, 2026d) reaches 5.95× by changing only which public corpus supplies the gold, with no training at all, and that number is conditional in a way worth quoting: it lives on the 1.96% of shared problems where two corpora write the same answer differently, and weighting every stratum by its share puts the benchmark-level difference at 0.0240 accuracy points while 0.0376 of the pool's accuracy mass changes hands, the two contested strata moving in opposite directions and partly cancelling. The pair is the point: the second number is what the phenomenon does, the first is why a leaderboard never surfaces it.

6.3 The correction factor, registered and rejected

Table 2 prices the shortcut. Across six mixed arms the effective enumeration $k_{\text{eff}} = C/S$ runs 1.149 to 1.811, hiding between 13.0% and 44.8% of capability — and it varies with the model at a fixed construction, which is already enough to show it is not a corpus property.

Table 2. Why no corpus-level correction exists. Six mixed arms under one fixed gold. S is the reported exact-match score, C the capability sum, $k_{\text{eff}} = C/S$ and the last column the fraction of capability the reported number omits. k_{eff} varies with the model at fixed construction, so it is a property of the model–corpus pair. Sources: er_ceiling_verdict, ksweep_verdict, combined by capability_estimator.

one fixed gold: why no corpus-level correction exists, since $k_{\text{eff}} = 1/\pi(c_0)$ is the model's corpus, mixed arm	seeds	S	C	k_{eff}	hidden
natural, Qwen3-8B	5	0.3643	0.4185	1.149	13.0%
natural, Qwen2.5-3B	3	0.2450	0.2821	1.151	13.2%
Arabic/Roman, Qwen2.5-7B	3	0.2540	0.3347	1.318	24.1%
natural, Qwen2.5-14B	3	0.3283	0.4846	1.476	32.2%
synthetic +1, Qwen2.5-7B	3	0.3053	0.5202	1.704	41.3%
natural, Qwen2.5-7B	8	0.2058	0.3727	**1.811**	44.8%
k golds enumerated: $k_{\text{eff}} = k$, a property of the corpus enumeration		$S(k)$	$\hat{C} = Sk$	k_{eff}	error
$k = 1$ (the corpus's own gold form)		0.3189	0.3189	1.418	+0.418
$k = 2$		0.1600	0.3200	2.826	+0.826
$k = 3$		0.1348	0.4044	3.355	+0.355
$k = 4$		0.1113	**0.4452**	**4.063**	**+0.063**

The condition under which the shortcut would be valid is homogeneity, and it fails only if both the corpus's gold and the model's policy correlate with a common driver. The model's side is directly measurable. Writing $s(x) = \text{acc}_A(x) + \text{acc}_B(x)$ and $\pi_A(x) = \text{acc}_A(x)/s(x)$, the two correlate at $r = -0.257 \pm 0.079$ on the eight unmarked arms, -0.355 ± 0.094 on the keyed ones and -0.287 ± 0.077 on the uninformative-marker control, with 24 of 24 arms

negative and none inside a simulated null band of $[-0.150, +0.165]$ obtained by resampling the same estimator at a constant allocation. The model writes one convention more often on the problems it finds easier.

This touches neither Proposition 1, which carries no independence hypothesis once π is read as the policy given a solve, nor the pin, which holds for any $\pi(c \mid x)$ under homogeneous P. What it establishes is that one of homogeneity's two legs is already standing on our own material, so a benchmark whose gold form tracks problem difficulty — a tool-generated corpus writing decimals for computational items being the obvious case — lies outside the shortcut's scope and must be analysed as a paired contrast on the same problems, which assumes nothing about how either corpus distributes conventions. We registered the correction factor, ran it, and report it as unavailable in general.

7 Threats to Validity

The conflicts are constructed. Every firing conflict here was built. An attempt to find one in the wild returned an empty set at the scale we could search, and the survey's re-examination of the same pairs (Anonymous, 2026d) finds that the larger pair we had treated as independently labelled is one lineage — one corpus's gold field filtered through the other's canonicaliser — which costs us a pair and widens the genuinely external rate. The prevalence claim is therefore the survey's and not ours; what is ours is what a score becomes once the material is contested.

The identity's scope is narrower than the phenomenon. Proposition 3 needs the conventions mutually exclusive on each problem and the output inside the enumeration. The first holds for numeric form on competition mathematics, which is why we measured there; the second is measured at 0.9844. In open-ended generation, code and scientific question answering the candidate forms of a correct answer may overlap rather than partition, in which case $\sum_c \mathrm{acc}_c$ exceeds C and the pin becomes an inequality again.

One reading of σ is not available at k=1. At a single accepted form, $S(1) = C\,\pi(c_1)$ and the ratio $S(1)/C = 0.7052$ is the model's policy, not the fraction of correct answers the enumeration covers. An earlier version of this work labelled the whole column σ and read the k=1 row as a support deficit, which asserts a distinction that row cannot make; the correction is recorded in Appendix D.

8 Discussion and Conclusion

An exact-match score on contested material is a product, and only one of its two factors is the thing the score is cited for. Naming the other factor is what makes the situation tractable: it is a property of the model–corpus pair, it does not transfer, and it cannot be published as a corpus-level correction — we registered that shortcut, ran it, and it fails on a condition we can measure — but it is removed exactly by scoring against every accepted gold and adding.

The deployable output is small and immediate: publish $\sum_c \mathrm{acc}_c$ beside the single-gold number, print the enumeration rather than leaving it inside a normaliser, and report σ so readers can price the one direction the sum can be wrong. Where the enumeration cannot be widened because the conventions are not interchangeable downstream, the honest route is a paired contrast on the same problems. None of this requires retraining, and on our material it recovers a 1.42× understatement that the reported number gave no way to see.

Reproducibility statement

Every number in this paper is written by a named analyzer to a verdict file, and every deciding test's read-out rule was frozen in the job script the scheduler executed before the run. Appendix E is the registration index: one row per registered claim, giving the commit that froze each threshold, the timestamp of the result it decided, and the interval between

them. Appendix F is the borrowing index: one row per claim taken from a concurrent manuscript, and whether this repository recomputes it. The conflict corpora are emitted by fixed-seed builders; the generation sets scored here, the analyzers that score them, the registration files and the verdict files are all in the supplementary material.

Ethics statement

No human subjects were involved and no personal data. The corpora derive from public competition-mathematics benchmarks under their own licences, and the conflicts are synthesised by published scripts rather than harvested. The finding — that exact match on material admitting more than one correct form reports capability multiplied by a form-matching rate — is a correction to measurement practice and carries no group-differential harm we are aware of. One result is reported defensively: because the allocation is movable by training-data order alone, a corpus-level content audit cannot see it; we develop no attack and the arms demonstrating it are the same arms the positive results are measured on. We have read the ICLR Code of Ethics and adhere to it.

AI use statement

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

Tasks in the required-disclosure list: none. Generative AI was not used to generate synthetic data sets, to develop the theoretical model or conceptual framework, to formulate the mathematical claims, to supply critical ingredients for the proofs or to write them, to propose or refine the hypotheses, to design the methodology or the experiments or to give feedback on them, to set or choose any pre-registered threshold, to implement methods, to clean or reformat any dataset, or to interpret the results. The corpora are emitted by fixed-seed builders, the runs 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. Translation and qualitative or thematic data analysis are not applicable to this work.

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

Verification. Every quantitative statement in this version was checked against the verdict file it cites, and every registered outcome against the read-out rule frozen in the job script the scheduler executed, which Appendix E tabulates against its commit. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims and artifacts.

References

Anonymous. The capability convergence hypothesis: Capability from access structure, not scale. arXiv preprint, 2026a. author list and repository withheld for double-blind review; the four terms this paper inherits are defined self-containedly in Table 1 so the citation is not load-bearing.

Anonymous. The infinite-division limit: Conflicting data trains a policy, not damage, 2026b. concurrent manuscript; author list and repository withheld for double-blind review.

Anonymous. The free-recipe limit: Every measured recipe effect is a gauge of one broken premise of the ideal, 2026c. concurrent manuscript; author list and repository withheld for double-blind review.

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

Anonymous. Writing a convention into weights: A dose response in marker reliability, and one undecided capacity row. concurrent manuscript; author list and repository withheld for double-blind review, 2026e.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 850–863, 2022.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 291–305, 2022.

John Collins and Minh Huynh. Estimation of diagnostic test accuracy without full verification: a review of latent class methods. Statistics in Medicine, 33(24):4141–4169, 2014.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research, 23(226):1–61, 2022.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 28(1):20–28, 1979.

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

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

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 5591–5606, 2023.

Hynek Kydlíček. Math-verify: A robust mathematical expression evaluation system. https://github.com/huggingface/Math-Verify, 2024.

Aitor Lewkowycz, Anders Andreassen, David Dohan, et al. Solving quantitative reasoning problems with language models. In Advances in Neural Information Processing Systems (NeurIPS), pp. 3843–3857, 2022.

Do Xuan Long, Hai Nguyen Ngoc, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F. Chen, and Min-Yen Kan. LLMs are biased towards output formats! systematically evaluating and mitigating output format bias of LLMs. In Proceedings of NAACL, 2025. arXiv:2408.08656.

Charles Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. In Proceedings of the 37th International Conference on Machine Learning (ICML), 2020.

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

A Proofs, and the Statements Held Out of the Main Text

A.1 Proposition 1, with the two-sided form

The main text states the degenerate case every deployed benchmark is in. The general statement allows the scorer to draw its gold at random.

Proposition 4 (General form). Let the model solve x with probability $p(x)$ and write its answer under convention c with probability $\pi(c \mid x)$ given that it solved x. Let the benchmark score x by exact match against one gold written under a convention drawn from $P(\cdot \mid x)$ independently of the model, and write $C = \mathbb{E}[p]$. Then

$$S \;=\; \mathbb{E}_x\big[\, p(x)\, \langle P(\cdot \mid x),\, \pi(\cdot \mid x)\rangle \,\big], \qquad C \cdot \inf_x \langle P, \pi\rangle \;\le\; S \;\le\; C \cdot \sup_x \langle P, \pi\rangle,$$

where $\langle P, \pi\rangle = \sum_c P(c \mid x)\pi(c \mid x)$. Putting a point mass at $c_0(x)$ recovers $S = C \cdot \bar{\pi}_{\text{gold}}$.

Proof. Condition on x. The generation is scored correct exactly when the model solves the problem and writes the convention the gold uses. Because π is defined given a solve, those two form a chain and not a product of independent events, so the conditional probability is $p(x)\langle P(\cdot \mid x), \pi(\cdot \mid x)\rangle$ by the chain rule alone. Take expectations. Bounding the inner product by its infimum and supremum over x gives the two-sided display, and putting a point mass at $c_0(x)$ gives the degenerate case. □

A.2 Proposition 3

Proof. Fix x. Mutual exclusivity means the generation matches at most one gold; the support premise means it matches exactly one whenever the problem is solved. So the indicator summed over c equals the indicator that x was solved, pointwise in x. Take expectations. No term of the form $P(c \mid x)$ enters, because the sum is over all golds rather than over the gold the corpus happened to draw. □

A.3 Proposition 2

Proof. $\sum_c P(c)\pi(c \mid x) = \frac{1}{k}\sum_c \pi(c \mid x) = \frac{1}{k}$ for every x, because P does not vary with x and $\pi(\cdot \mid x)$ is a distribution over exactly those k conventions. Hence $S = \mathbb{E}[p]/k = C/k$. □

We restate the scoping that accompanies this proposition in the main text, because it is the statement most easily over-read. Both hypotheses are load-bearing and both fail on real material: a benchmark presents each problem once under one gold, the degenerate case $P(c_0 \mid x) = 1$, as far from uniform as the simplex allows; homogeneity fails whenever a corpus's gold form tracks anything about the problem, one leg of which Remark 1 measures; and if π puts mass outside the k accepted conventions the inner sum falls below $1/k$, a deficit belonging to the enumeration rather than the metric, which §6 measures through $\sigma = 0.9844$. C/k is retained because it is the form in which a constructed enumerating scorer can be built and checked.

A.4 The statements held out of the main text

Corollary 2 (A score a policy can move, and two benchmarks that cannot be compared). Two models with the same capability C and policies π_1, π_2 are reported by the same benchmark in the ratio $\bar{\pi}^{(1)}_{\text{gold}}/\bar{\pi}^{(2)}_{\text{gold}}$: the free variable is how each model writes, not what it knows. Symmetrically, one model's one set of generations, scored against two corpora that carry the same problems under different conventions, is reported in the ratio $\bar{\pi}^{(A)}_{\text{gold}}/\bar{\pi}^{(B)}_{\text{gold}}$. Neither ratio contains C; both are properties of the pair rather than of capability; and both are bounded only by the support of the policy, so as $\bar{\pi}_{\text{gold}} \to 0$ they diverge.

Corollary 3 (What an exact-match benchmark measures on contested material). An exact-match score on contested material is $C \cdot \bar{\pi}_{\text{gold}}$: a product of what the model can do and the rate at which it writes the one form this corpus shipped as gold. It is not a measurement

of capability, and the second factor is a property of the model–corpus pair rather than of either alone. It carries no term for how the model resolves the underdetermination.

Proposition 5 (Disjoint escapes).

1. If the learner's capacity is below what the assignment demands, the assignment cannot be served at all, and raising the capacity removes that obstruction. It leaves Proposition 1's bound unchanged, since the bound does not mention capacity.
2. If $H(c \mid I) > 0$ the bound holds for every capacity, every channel and every architecture. Making the convention a function of the query-time information — a key, present at write time and quoted at query time — sets $H(c \mid I) = 0$ and removes the obstruction while adding no capacity. It does not by itself set the loss to zero: what a learner faces is $\min_{f \in \mathcal{F}} H(c \mid f(I))$ over the functions its hypothesis class can compute, so this escape is available exactly when the identifying function is realisable. Realisability is a second condition rather than a restatement of the first; the unqualified form is false in a way this programme's own instruments demonstrate, a concurrent manuscript reporting a base checkpoint that cannot reliably report the presence of a question mark (Anonymous, 2026e).
3. A third intervention acts on neither the learner nor the data but on the scorer. Proposition 3 needs the conventions to be mutually exclusive on each problem, so that one generation matches at most one gold. A metric that accepts several of them as correct breaks that premise: k collapses to 1, the score returns to C, and $H(c \mid I)$ costs nothing. It leaves $H(c \mid I)$ exactly where it was.

The three act on three kinds of object: (i) relaxes a constraint on the code, (ii) changes the information structure, (iii) changes the distortion measure. No coding theorem converts one into another, which is why the main text measures them separately.

Corollary 4 (No arrangement escapes underdetermination). The order in which training data reaches the optimiser does not enter $P(c \mid I)$ for a held-out query. Hence the optimum is the same for every training path, and the ceiling of Proposition 1 is invariant to it. What a path can move is the realised policy, not the bound it is measured against.

Remark 1 (The homogeneity hypothesis, measured on the model's side). Homogeneity fails only if both sides correlate with a common driver: the corpus's gold convention must vary with x, and so must the model's policy. The model's side is directly measurable, since every arm records per-problem accuracy against each gold at k=4 samples. Writing $s(x) = \mathrm{acc}_A(x) + \mathrm{acc}_B(x)$ and $\pi_A(x) = \mathrm{acc}_A(x)/s(x)$, the two correlate at $r = -0.257 \pm 0.079$ on the eight unmarked arms, -0.355 ± 0.094 on the keyed ones and -0.287 ± 0.077 on the uninformative-marker control, with 24 of 24 arms negative and none inside a simulated null band of $[-0.150, +0.165]$ obtained by resampling the same estimator at a constant allocation. The model writes one convention more often on the problems it finds easier. This touches neither Proposition 1, which carries no independence hypothesis at all once π is read as the policy given a solve, nor the pin, which holds for any $\pi(c \mid x)$ under homogeneous P. What it establishes is that one of homogeneity's two legs is already standing, so a benchmark whose gold form tracks problem difficulty lies outside the pin's scope and must be analysed as a paired contrast. It also generalises a defect this programme had recorded as local: a concurrent manuscript's void was attributed to one corpus (Anonymous, 2026e), and the dependence is a property of all of them.

B The Apparatus, and the Terms

Table 3. Every term this paper uses in a non-standard way, defined once and used consistently thereafter. Four of these are inherited from the construction this paper extends (Anonymous, 2026a) and are given here so that nothing in the paper depends on having read it. The rest are introduced here. Where a term has a symbol, the symbol is what the propositions use.

term	symbol	what it means in this paper
convention	$c \in \mathcal{C}$	one of the mutually exclusive, all correct ways of writing an answer, such as $\frac{15}{2}$ against 7.5
multiplicity	k	how many conventions the corpus's problems admit
capability, or union accuracy	$C = \mathbb{E}[p]$	the probability the model solves the problem at all, which is the sum of its accuracies over the enumerated conventions. The two names are the same quantity and the second is the more careful one: a model that writes a correct form nobody enumerated scores zero here too, so C is relative to the enumeration and not an unconditional statement about the model. σ in Table 5 is exactly the size of that gap, and at k=4 it is 0.9844
policy	$\pi(c \mid x)$	which convention the model writes, given that it solved the problem: the degree of freedom underdetermination leaves open
query-time information	$I(x)$	everything the model may condition on when it answers
underdetermination	$H(c \mid I) > 0$	the applicable convention is not a function of what the model may condition on
wall		a floor on loss that no choice within a named resource can lower
escape		an intervention that removes a wall, as against a move that travels along it
form-match rate	$\bar{\pi}_{\text{gold}}$	the rate at which the model writes the particular form the corpus shipped as gold, $\mathbb{E}_x[\pi(c_0(x) \mid x)]$ weighted by capability. A property of the model-corpus pair: the same model has a different $\bar{\pi}_{\text{gold}}$ against a different corpus, which is why one generation set scores differently against two of them
pin	Prop. 2	that the score equals C/k for every policy, rather than being bounded by it. Read the word as pinned in place: the score is held at one value while the policy moves under it. A tight identity, not a tight bound, and the difference is that an identity inverts
key		a marker present at write time and quoted at query time, which makes c a function of I and so sets $H(c \mid I) = 0$
stream		the construction this paper extends: N bindings sent past a state of budget B (Anonymous, 2026a)
capacity wall	$B < D$	a floor imposed by too few bits of state to hold the assignment. This is the wall Anonymous (2026a) calls the Shannon wall, reached here by another road; we use the resource-facing name throughout and note the correspondence once, here
extent wall		a floor imposed by too little reach into the context. This is Anonymous (2026a)'s horizon wall. It appears in this paper only as the third axis of Figure 2 and nothing below turns on it
(no counterpart)		Anonymous (2026a)'s third wall, the circuit wall, is conditional on $\text{TC}^0 \neq \text{NC}^1$ and concerns composition depth. It has no counterpart here and is not used

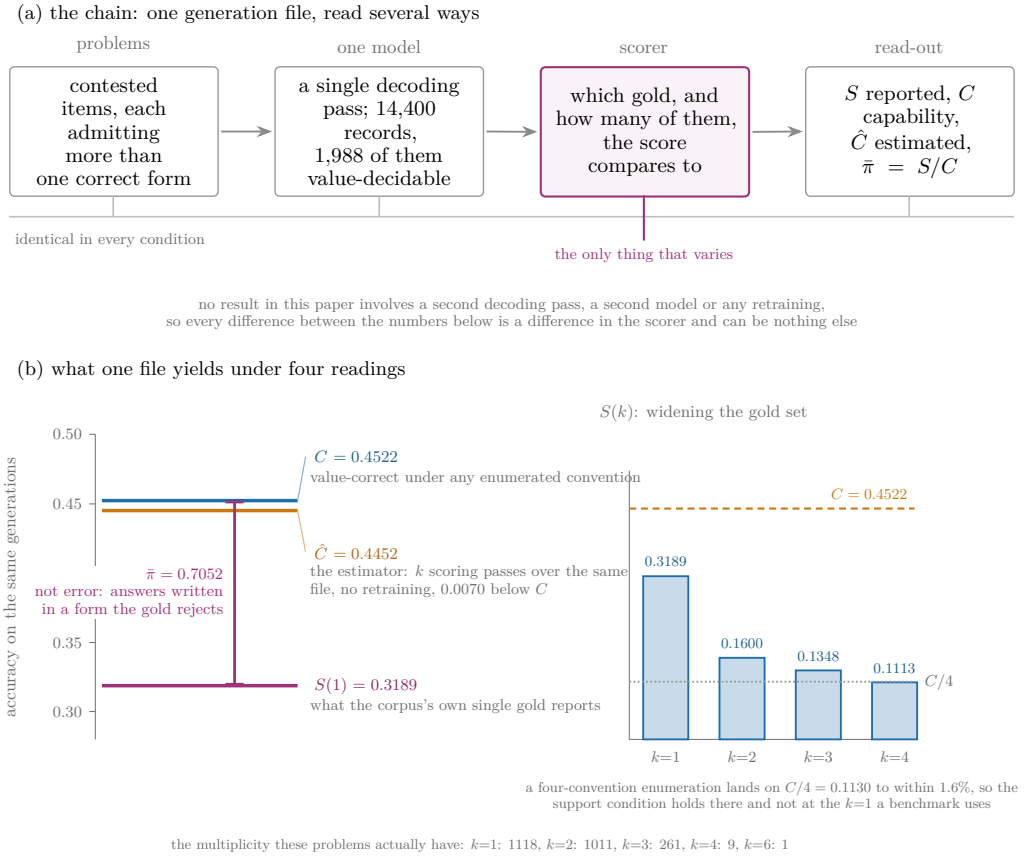


Figure 4. One generation file, read four ways. (a) No result in this paper involves a second decoding pass, a second model or any retraining. Every number is a reading of the same file, so every difference between them is a difference in the scorer and can be nothing else; that is what makes the identity $S = C \cdot \bar{\pi}_{\text{gold}}$ testable rather than definitional. (b) Left: the corpus's own single gold reports 0.3189 where the model is value-correct 0.4522 of the time. The gap is not error. It is answers whose value is right and whose form the gold does not accept, and $\bar{\pi}_{\text{gold}} = 0.7052$ is its size. The estimator recovers 0.4452 of the capability from k scoring passes over generations that already exist. Right: widening the gold set from one convention to four walks the reported score down to $C/4$ within 1.6%, which is the support condition holding at $k = 4$ and conspicuously not holding at the $k = 1$ a benchmark actually uses. The multiplicity figures beneath are the distribution of admissible forms over these problems, so a reader can see that $k = 4$ is a bound on the enumeration and not a description of a typical item. Source: ksweep_verdict.

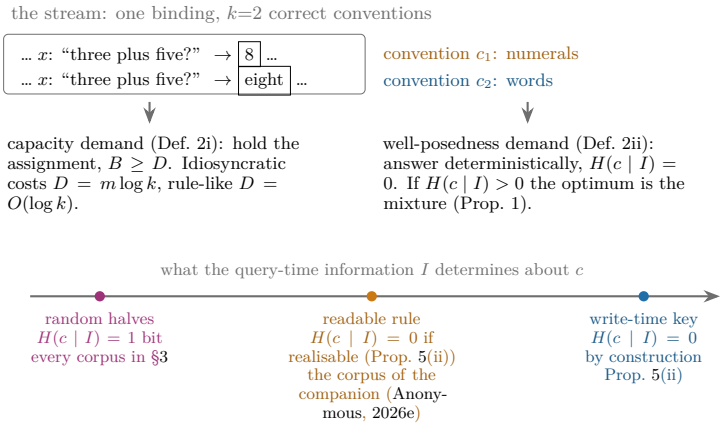


Figure 5. The method in one picture. One binding arrives under two conventions that are both correct, which splits the classical question "can the learner serve it?" into the two demands of Definition 2: holding the assignment is a resource question (B against D), answering deterministically is a task question ($H(c \mid I)$), and the axis below is the one every corpus in this paper sits on. The family's existing conflicted corpora all sit at the left end; a concurrent manuscript's rule corpus (Anonymous, 2026e) is built to sit in the middle, and the middle position must be certified rather than assumed — §G records that this paper does not rest on it; the write-time key sits at the right.

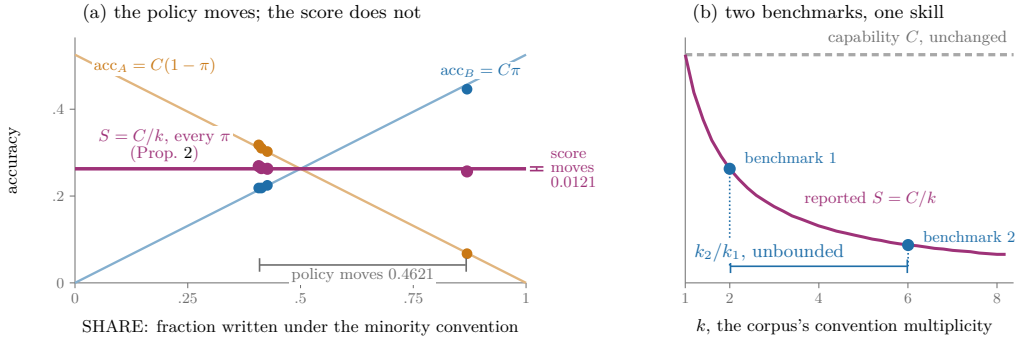


Figure 6. The limiting case, drawn — and it is a limiting case, not a measurement of any benchmark. (a) Proposition 2. The thin lines are the per-convention accuracies a policy π produces, $\text{acc}_A = C(1-\pi)$ and $\text{acc}_B = C\pi$; the heavy line is the reported score $S = C/k$, which does not depend on π. The points are four arms of one corpus on one model at one budget and one seed set, differing only in the order the training data reached the optimiser (Table 1): the policy moves 0.4621 and the reported score 0.0121. Nothing is fitted. (b) Corollary 1, derived rather than measured: no experiment here varies k. Capability is fixed while multiplicity grows, so the reported score falls as C/k. Both panels are drawn under hypotheses that hold on no benchmark we know of — full support and a uniformly drawn gold convention — and no measurement in this paper rests on either. They are here because the exact form of the degenerate case is what makes Proposition 1's chain-rule statement the one that does the work.

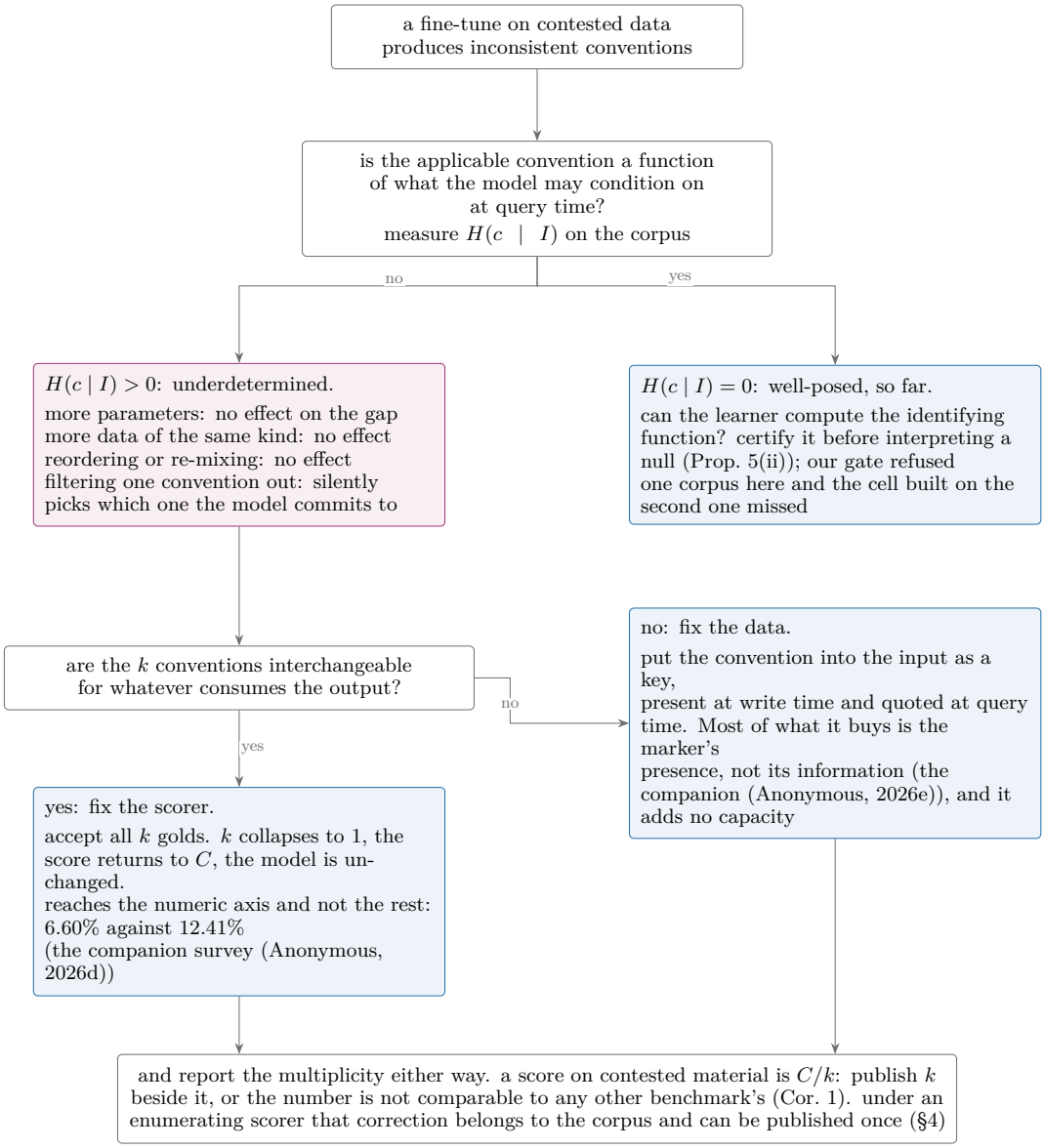


Figure 7. The diagnosis as a practitioner can run it. The remedies act on different quantities, so the choice among "more data", "a bigger model", "a different mixing order" and "add metadata" is decided before any of them is tried: the left branch says the first three cannot work, and the right branch says what to certify before believing a null. The last box is owed to readers of a score rather than to the team producing it. Quantities printed in the figure are measured elsewhere: the information-against-presence split and the canonicaliser's reach in concurrent manuscripts (Anonymous, 2026e;d), which §G records as illustrations rather than premises, and k_{eff} in Table 2, which is this paper's own.

C Extended Related Work

Answer matching, and what a normaliser is doing. math_verify (Kydlíček, 2024), symbolic equivalence, Minerva's normalisation (Lewkowycz et al., 2022) and the math_equal family shipped with MATH (Hendrycks et al., 2021) all canonicalise numeric form, so that $\frac{3}{4}$ and 0.75 become one answer. Proposition 5(iii) places every one of them: they are scorer-side, they remove the penalty and not the underdetermination, and they are available exactly when the conventions are interchangeable for whatever consumes the answer. Work on answer equivalence (Bulian et al., 2022) and on evaluation in question answering (Kamalloo et al., 2023) treats the same phenomenon as a matching problem to be improved rather than a quantity to be derived.

Surface form and format effects. That a model's score moves with the surface form it is asked to produce is documented (Holtzman et al., 2021; Long et al., 2025); the contribution here is not the observation but the identity, which says exactly what the moved score equals and therefore exactly what recovers the quantity behind it.

Latent true labels. Recovering a hidden label from noisy observations (Dawid & Skene, 1979; Collins & Huynh, 2014) performs an inference close in form to ours and differs in the error term: a latent-class posterior's accuracy depends on unverifiable assumptions about the noise process, while the estimator here has one failure mode, is one-signed, and is priced by a quantity the same scoring pass reports.

Multiplicity and underspecification. That equally good models can differ in behaviour the selection criterion does not see is the predictive-multiplicity and underspecification literature (Marx et al., 2020; Black et al., 2022; D'Amour et al., 2022). This paper's contribution to that group is a case where the invisible coordinate is not merely unselected but provably unscored, with the algebra that says so and a procedure that prices it.

What a benchmark number means. Construct-validity work argues that leaderboard numbers often fail to measure what they are cited for (D'Amour et al., 2022; Schaeffer et al., 2023). The case here is stronger than a validity worry and weaker in scope: the quantity reported and the quantity wanted differ by a factor that can be written down, measured and removed, on material whose prevalence we do not establish.

D What Changed From the Longer Version

Two corrections moved a statement this paper derives or a number it prints.

Proposition 1 no longer carries an independence hypothesis. An earlier statement assumed the value and form events independent. Writing π as the policy given a solve makes the factorisation the chain rule, so the hypothesis was never needed and is withdrawn. A model whose chance of solving a problem is bound up with the form it is about to write is inside the statement rather than outside it. The measured across-problem dependence of Remark 1 now prices a property of the corpora rather than defending a premise.

One column was read as a support deficit where the data cannot make that reading. At a single accepted form, $S(1) = C\,\pi(c_1)$, so the ratio $S(1)/C = 0.7052$ is the model's policy and not the fraction of correct answers the enumeration covers. An earlier version labelled the whole column σ and read the k=1 row as support leakage, which asserts a distinction that row cannot make: a correct answer scoring zero did so because it was written some other way, and whether that other way lies inside or outside a larger enumeration is a question the k=1 measurement does not ask. Only the k=4 end is σ.

E Registration Index

Eighteen read-out rules were registered before the runs they decide. Branches within a registration are mutually exclusive and exhaustive, so exactly one fires and none is selected after

the fact: the family is 18 decisions rather than 18 significance tests. What a reader should discount is the number of chances a favourable branch had, which is one per registration, and that is why each registration names all of its branches before the runs.

Of the 15 read out, 5 returned a branch their own preamble did not expect and 5 declined to report at a guard. The supplementary material carries one row per registered claim, giving the commit that froze each threshold, the timestamp of the result it decided, and the interval between them, together with the registration files themselves.

Intervals. Where absolute levels are printed to four places on eight seeds, the keyed mean is 0.3259 with a 95% interval of $[0.3210, 0.3309]$ and the unkeyed baseline 0.1863 with $[0.1808, 0.1918]$; the dose endpoints are 0.6519 with $[0.6419, 0.6619]$ and 0.6562 with $[0.6426, 0.6698]$. They do not overlap where this paper claims a separation and do overlap where it reports one. Under the shared evaluation stream disclosed in §5 these intervals understate evaluation-sampling uncertainty and are read only for separations between conditions scored on the identical stream, where the common-mode term cancels.

F The Borrowing Index

A number whose only warrant is a manuscript a referee cannot read is a number a referee cannot check. This programme therefore separates provenance — a corpus or tool built elsewhere — from evidence — a measurement performed elsewhere — and permits this paper to rest on the second only where this repository reproduces it. An analyzer recomputes every such claim or records that it cannot, and its output is the table shipped with the supplementary material.

Five of the six claims this paper takes from a concurrent manuscript are recomputable from artefacts in this repository, and the citations in the main text say so. The sixth is used as an illustration of the scorer-side escape rather than as evidence for any proposition. Table 4 is that ledger, one row per borrowed claim.

Table 4. Every claim this paper takes from a family member, and its status here. Five of the six are recomputable from artefacts in this repository, and the citations in the main text have been changed to say so. The sixth is used as an illustration of the scorer-side escape rather than as evidence for any proposition, and the measured version of the same point is this paper's own cross-tabulation against the shipped canonicaliser. No row of this table is a premise of any proposition in §3 or §4. Source: independence.

what is borrowed	status here	how it is checked
the block-length ladder is flat, which Cor. 4 requires	recomputed here	eight interior arms of one corpus at three seeds span 2.2 contrast floors while the blocked corner sits 22.0 floors away; readout_independence.py
the four arms that exhibit the pin at fixed capability	measured here, corpus provenance only	fixedC_test.json, tested in
the narrow canonicaliser behind the wild survey's convention column	tool ships here	mine_wild_conflict.py, cross-tabulated against math_verify by verify_wild_with_shipped.py
the key's 87.5% of the additive ceiling on the first family	artefact ships here	unlock_verdict.json, the frozen read-out itself
the second family moving the key's residue by +2.92 floors	artefact ships here	q3unlock_verdict.json
a conflict corpus drawn from real pipeline disagreement showed no allocation effect, its gold pairs being equal under the scorer	citation only	its forty arms are in that project's result base. Used as illustration, not as evidence; the measured form of the same point is the companion survey (Anonymous, 2026d)

The one borrowed measurement a proposition rested on, recomputed here. Corollary 4 says the training path cannot enter the conditional law of the convention given a held-out

query, so a block-length sweep must be flat. Recomputing the ten arms from this paper's own result base gives shares of 0.407 to 0.449 across $L = 1, 3, 6, 9, 18, 27$, shuf and purerand, a span of 2.2 contrast floors, with $L = 54$ at 0.492 and the blocked corner at 0.869, which is 22.0 floors above the interior. The corollary is therefore supported by a measurement in this repository rather than by a reference to a manuscript a referee has not seen.

G What This Paper Owns and Borrows, in Full

Four concurrent manuscripts of the same research programme are cited in the main text. The accounting matters because two of the four could be mistaken for support this paper does not have. The governing rule is that every proposition here is proved here, and every borrowed measurement a proposition rests on is recomputed in this repository (Appendix F).

What this paper owns. The identity $S = C \cdot \bar{\pi}_{\text{gold}}$ and its corollaries (Propositions 1 and 2, Corollaries 2 and 1); the summation identity and the capability estimator built on it (Proposition 3); the measured form of the parent framing's support assumption; the disjointness of the escapes (Proposition 5); and the pricing of the correction-factor remedy, registered, run and decided against.

What it borrows, and in which direction. (i) The four fixed-capability arms of Table 1 are a concurrent manuscript's corpora (Anonymous, 2026b), re-scored here rather than re-trained; the channel-level mechanism that produces the allocation belongs to that manuscript and is cited for provenance and never as evidence, since no claim here depends on why the allocation moved, only that it did at fixed capability. (ii) The prevalence of gold-form disagreement in public corpora, and the 5.95× cross-corpus factor, are the survey's (Anonymous, 2026d); they license the choice of construction rather than any proposition here, and the survey also corrects us, finding that a pair we had treated as independently labelled is one lineage. (iii) The training-side marker results are a third manuscript's (Anonymous, 2026e), used to show that the realisability condition in Proposition 5(ii) is a second condition and not a restatement — a use that makes our own escape narrower, not wider. (iv) A fourth manuscript (Anonymous, 2026c) reports a conflict corpus on which no allocation effect appears at all, because the shipped checker already equates its gold pairs; this contradicts the region where our phenomenon lives and marks its boundary. The idealisation-first structure is the parent preprint's (Anonymous, 2026a), a debt of form rather than a borrowed measurement.

So of the borrowings, one is a corpus re-scored here, one licenses a construction and takes a pair away from us, one narrows our own escape, and one marks the boundary outside which nothing here applies. None of them is a premise of a proposition. A referee holding only this paper can re-derive every formal statement in it.

H Supporting Measurements

A both-halves-spelled control. The displacement figures of §6 are symmetric by construction, and a control in which both halves of the corpus are written under the same convention returns the allocation to the corpus rate within the contrast floor, confirming that the straddle is produced by the conflict rather than by the evaluation.

The support sweep in full. Table 5 is the premise under Proposition 2, measured.

The six mixed arms, and the two held out. Table 6 is the displacement measurement in full and Table 7 carries the two rows held out of it together with the tilt column for all seven.

The controlled scale sweep. Capability and allocation are separately real, and Table 8 is the one grid on which only the parameter count varies.

Allocation is not constant across problems. The identity $S = C \cdot \bar{\pi}_{\text{gold}}$ would be uninteresting if π were a single number per arm. It is not: the model writes one convention more

Table 5. The premise under Proposition 2, measured. 1,988 generations whose gold is a non-integer rational with four distinct canonical renderings. C is form-independent correctness and $S(k)$ the exact expectation over a uniformly drawn gold convention. The fourth column is one arithmetic expression that means two different things at the two ends, and the fifth column says which. At k=4, $S(k)k/C = \sigma$, the fraction of correct answers whose surface form the enumeration covers, which is the premise Proposition 2 needs. At k=1 the gold is always c_1, so the same expression is $\pi(c_1)$, the model's policy, and carries no information about support: with one accepted form, writing a different accepted form and writing an unaccepted one are the same event. An earlier version labelled the whole column σ and read the k=1 row as a support deficit, which asserted a distinction that row cannot make. The registered order puts plain a/b second and neither the corpora nor the model ever writes it, so the middle rows are neither quantity and are not an interpolation; k=1 and k=4 are the informative ends. Source: ksweep_verdict.

k accepted	reported $S(k)$	C/k	$S(k)k/C$	what the ratio is
1 (what a benchmark does today)	0.3189	0.4522	**0.7052**	$\pi(c_1)$, the policy
2	0.1600	0.2261	0.7075	neither (see caption)
3	0.1348	0.1507	0.8943	neither (see caption)
4 (the full enumeration)	0.1113	0.1131	**0.9844**	σ, the support

Table 6. Six mixed arms, and a policy varying by a factor of 3.5 that no exact-match score can see. In every row the two halves of the corpus disagree and nothing marks which convention a query wants. $C = \text{acc}_A + \text{acc}_B$ is what the model solves at all and SHARE $= \text{acc}_B/C$ is the policy. The rows differ in C and therefore in the score they report, so what the third column cannot distinguish is not these rows from each other but each row from every other policy at the same C; the fixed-C demonstration is Table 1. SHARE runs 0.130 to 0.448 across a 28× parameter range, two pretraining families and three conflict constructions. Source: er_ceiling_verdict.

corpus, mixed arm	C	reported $C/2$	acc_A / acc_B	SHARE
natural, Qwen3-8B	0.4185	0.2092	0.3643 / 0.0542	0.130
natural, Qwen2.5-3B	0.2821	0.1410	0.2450 / 0.0371	0.131
Arabic/Roman, Qwen2.5-7B	0.3347	0.1674	0.2540 / 0.0808	0.241
natural, Qwen2.5-14B	0.4846	0.2423	0.3283 / 0.1563	0.323
synthetic +1, Qwen2.5-7B	0.5202	0.2601	0.3053 / 0.2149	0.413
natural, Qwen2.5-7B	0.3727	0.1863	0.2058 / 0.1669	0.448

Table 7. The two rows held out of Table 6, and the tilt column for all seven. The tilt is $|\text{acc}_A - C/2|$, the deviation from an equiprobable split. An earlier revision printed one model here as preliminary at two seeds and excluded it from every quoted range; its job has since finished at five and it is now in the ordering with the rest, which is where its largest-tilt position comes from. The control spells both halves, so nothing is disputed and its SHARE is not an allocation between disputing conventions; it is printed to show that a corpus with no disagreement produces a displacement of the same size, which is why the tilt cannot be read as a measure of conflict. Seed counts are the number of complete arms on the result base.

corpus, mixed arm	C	$\text{acc}_A/\text{acc}_B$	tilt	SHARE	seeds
natural, Qwen3-8B	0.4185	0.3643 / 0.0542	0.1550	0.130	5
natural, Qwen2.5-3B	0.2821	0.2450 / 0.0371	0.1040	0.131	3
Arabic/Roman, 7B	0.3347	0.2540 / 0.0808	0.0866	0.241	5
natural, Qwen2.5-14B	0.4846	0.3283 / 0.1563	0.0860	0.323	3
synthetic +1, 7B	0.5202	0.3053 / 0.2149	0.0452	0.413	8
natural, Qwen2.5-7B	0.3727	0.2058 / 0.1669	0.0195	0.448	8
control: spelled both halves	0.3796	0.0979 / 0.2817	0.0919	—	3

Table 8. The controlled scale sweep: one corpus, one family, one budget, three matched seeds, only the parameter count varying. Capability rises 71.8% from 3B to 14B while the policy moves over a range of 17.85 contrast floors whose maximum is at the interior point. Three points cannot establish a shape and the text does not claim one; the claim is the range. The reported score is $C/2$ and therefore tracks capability alone. The same-convention control on the identical grid is inert to 0.08 floors, which is what rules out a measurement artefact of scale: if the instrument drifted with parameter count, the control would drift too. Seeds are the intersection, so the 7B row contributes three of its eight, and those three are one execution batch. Source: scale_sweep.

	conflict (nat)			control (natc)	
model	C	reported $C/2$	SHARE	C	SHARE
Qwen2.5-3B	0.2821	0.1410	0.1322	0.5391	0.5000
Qwen2.5-7B	0.3850	0.1925	**0.4732**	0.7271	0.4985
Qwen2.5-14B	0.4846	0.2423	0.3220	0.9362	0.4985
range	0.2025	0.1013	**0.3410**		0.0015

often on the problems it finds easier, which is the correlation Figure 8 rejects the null for on all 24 arms.

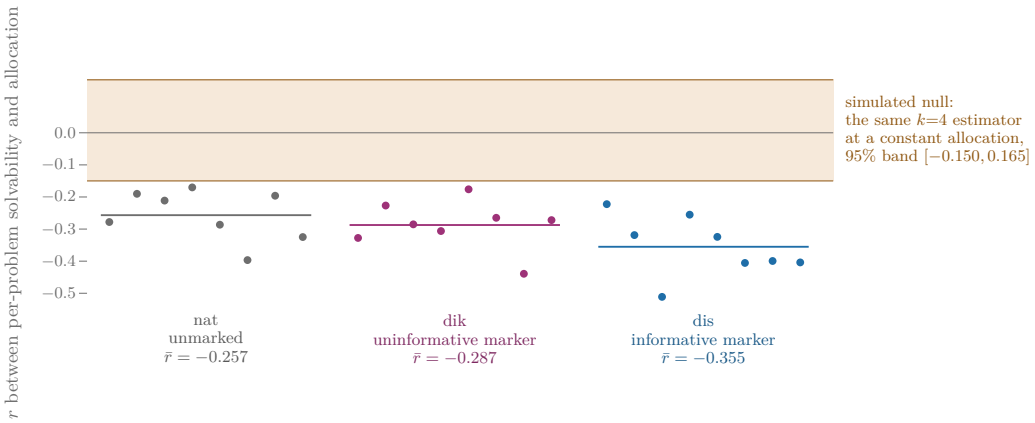


Figure 8. Twenty-four arms, none inside the band. Each point is one training arm's correlation between per-problem solvability and per-problem allocation; the heavy line is the condition mean. The shaded band is the central 95% of the same estimator applied to synthetic data whose allocation is constant, which is the hypothesis being rejected. All 24 arms are negative and all 24 fall below the band. The model writes one convention more often on the problems it finds easier. Source: policy_correlation.

Table 9. The per-arm values behind Figure 8. Eight seeds in each of three conditions, one shuffling arm, one 7B model. The null band is $[-0.150, +0.165]$ with mean -0.010.

condition	r per seed (42–49)								mean	sd
nat	−0.278	−0.190	−0.211	−0.170	−0.286	−0.397	−0.196	−0.325	−0.257	0.079
dik	−0.328	−0.227	−0.285	−0.306	−0.176	−0.265	−0.439	−0.272	−0.287	0.077
dis	−0.222	−0.319	−0.511	−0.255	−0.324	−0.406	−0.400	−0.404	−0.355	0.094