The Structural Content of a Transfer Relation: A Benchmark's Labelled Dimensions Predict Less About Each Other as They Are Split Further

Anonymous authors
Paper under double-blind review

Abstract

A benchmark that reports P labelled dimensions side by side invites $P(P-1)$ ordered transfer readings – a system ahead on this dimension is ahead on that one – and suites are increasingly refined into more dimensions on a fixed item budget. What refinement does to those readings has never been measured: the pair count grows quadratically while each pair is measured on fewer items, so refinement multiplies the claims and divides the evidence. The relation's apparent size is confounded too, since comparing whatever systems are to hand mostly measures a base rate rather than transfer. On eight public leaderboard corpora and over a thousand models – mmlu, big-bench hard and six more, using no artifact of our own – we split a relation's description length exactly into a base-rate term and a structural term, isolate the second by matching ability, and send the dimension count to infinity at a pinned item budget. Out of sample, label-predictability runs from 6.0% on mmlu to 53.1% across one leaderboard's own benchmarks, so there is no single number for what a benchmark's labels predict; refinement then drives it down as $\Lambda \propto m^{-1.223}$ against an analytic no-structure baseline of m^{-1}, and read backwards on mmlu's taxonomy it falls 9.98% to 6.02% from 17 dimensions to 57 while pair heterogeneity rises. A minimum-description-length statistic shows the structural term is real on all eight corpora nonetheless, paying 11.58× its cost on mmlu against 0.82× for a permutation null. The refinement limit has a closed form, and so a design rule computable before any model is run: a suite needs $P \geq 7$ to report a label-predictability of 6% at all, and $P \geq 11$ to report 1%.

1 Introduction

A benchmark reports its dimensions side by side and the reader supplies the generalisation. Ten labelled dimensions are ninety ordered claims of the form a system that handles this one handles that one, and a suite reporting them together without testing any has made those claims without paying for them. Whether that is a real debt or a bookkeeping artefact turns on one measurable quantity: how much of a real transfer relation is structure rather than base rate (Figure 1).

What this paper delivers.

(i) The refinement limit measured twice, and the resolution wall it exposes (§6). Run forwards on a pinned item budget, an 8× refinement costs 92% of what a label can be shown to predict while the truth is held fixed by construction; run backwards on mmlu's own taxonomy, Λ falls 9.98% → 6.02% from 17 dimensions to 57.

(ii) Two quantities separated that one word usually covers (§5, §6): pair heterogeneity D, which decides in bits whether a per-pair parameter pays for itself, and label-predictability Λ, which is what a dimension table claims. They rank together at Spearman +0.563 across eight corpora and are not one thing.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

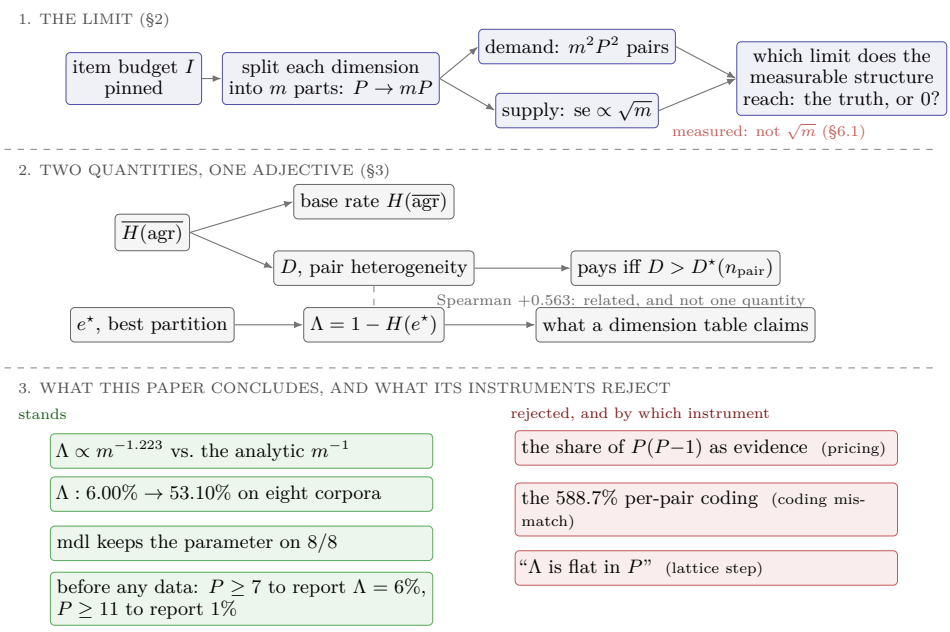


Figure 1. The paper on one page: one boundary, two statistics, and a ledger with both signs. Band 1 is the only source of questions here. Band 2 is the separation everything downstream is read through. Band 3 is the ledger with both signs shown: three plausible readings fail, each killed by a named instrument rather than by taste, and the last thing standing is the only one usable before a suite collects anything. Every measured quantity in bands 2 and 3 is read from result files at draw time and cannot be retyped.

(iii) Both, priced on eight public corpora with no artifact of ours (§6.2): Λ_{out} from 6.00% to 53.10%, so there is no single number for what a benchmark's labels predict.

(iv) An mdl verdict in this paper's own currency (§5): a per-pair parameter costs 4.95 bits and needs $D > D^\star(n_{\text{pair}}) = \log_2 n/2n$ to pay for itself, which is 0.52% at mmlu's $n_{\text{pair}} = 952$; the observation pays 11.58×, a no-association null pays 0.82× and is rejected, and all eight corpora pay.

(v) Four laws that need no measurement, one of them a rule a suite can use before it collects anything (§3): $P \geq 7$ to report $\Lambda \approx 6\%$, $P \geq 11$ to report 1%, computable before data exists, reproducing all three verdicts this paper reached the expensive way.

(vi) A methodological warning with a number on it: a programme that compares whatever systems it has, rather than systems of matched ability, reads a description length 4.87× off, of which 93.4% is base rate and not structure.

What this paper is not, and exactly what it shares. A concurrent manuscript of the same research programme asks what a research programme can afford to verify and derives a wall from a per-paper testing budget. This paper shares six of its formal results and none of its subject: the two propositions restated without proof in §3 are stated and proved there in a testing budget and here in bits. Nothing here counts a programme's tests, computes a transfer density, or audits anybody's preregistrations. Every number below comes from public model results and can be recomputed without us. In one line: the concurrent manuscript is a position and an accounting; this paper is a measurement. The shared propositions travel in that direction only, and neither paper's evidence is evidence for the other's claim.

2 The limit: refinement against a pinned budget

Every result here is read off one boundary, reached by pinning a resource and sending a load to infinity rather than by consulting the literature.

The infinite refinement. A corpus has a total item budget I and a labelling into P dimensions, so each dimension is measured on I/P items. Refine the labelling: split each dimension into m parts. The transfer surface grows as $(mP)(mP-1) \sim m^2P^2$ while each of its pairs is measured on $I/(mP)$ items, so the per-dimension standard error grows as $\sqrt{m}$ and the difference a pair must exceed to be resolvable grows with it.

Send $m \to \infty$ with I pinned. Demand is the number of pair facts to be established, $\Theta(m^2)$. Supply is the resolution left to establish them with, falling as $m^{-1/2}$ per pair.

Question: does the measurable structural content of the transfer relation converge to the relation's true structure, or to zero?

The dual is this paper's whole negative half. Refinement buys pairs and spends resolution, so a programme cannot label its way to more transfer knowledge: every finer pair is a pair measured with less. §6 answers the question twice, off a ladder somebody else published and by splitting the corpus's own items, and the measurable structural content converges towards zero: it falls by 40% relative across the two rungs of mmlu's taxonomy that can resolve it, and by 92% when the same corpus is split 8 ways at a fixed item budget with its truth held constant by construction.

The limit is aimed at a published result, not at a gap in the literature. That this corpus has a dominant general factor and a low-rank capability matrix is prior work we do not dispute; we aim the refinement limit at its implicit assumption, that low rank means cheap to describe. That holds on the unmatched matrix, where the base rate lives; at matched ability, where every "this method helps" claim sits, the rank-r codes we can build buy single digits, and the finding does not survive at the boundary transfer is asked at.

3 Four laws and one design rule

The measurements below are priced in bits. All are arithmetic on definitions and none is fitted, which is why they can serve as references. Two are shared with a concurrent manuscript of the same research programme and restated for self-containedness; four are this paper's own, each supplying a baseline the measurements below are read against.

Proposition 1 (The obligation's unit). Let each ordered pair of dimensions carry q transferable quantities, whose outcomes are logically independent. Then the transfer relation is defined on $P(P-1)q$ items rather than $P(P-1)$, and the maximum-entropy description length is $H = P(P-1)q$ bits. We measure $q = 1$ throughout, a leaderboard reporting one score per dimension, so $q > 1$ can only make the counts larger.

Proposition 2 (The entropy law). Let $\mathcal{H}$ be the class of relations a reader admits a priori and $H = \log_2 |\mathcal{H}|$ its description length. A transfer test yields at most one bit, so identifying the relation costs at least H tests:

$$\begin{aligned} \text{arbitrary relation:} \quad & H = P(P-1), \\ \text{equivalence relation:} \quad & H = \log_2 \mathrm{Bell}(P) = \Theta(P \log P), \\ \text{labelling from a fixed vocabulary of } N\text{:} \quad & H = P \log_2 N. \end{aligned} \tag{1}$$

A structural prior is a smaller hypothesis class, so it reduces the obligation by exactly the bits it removes and by no more.

A rule that decides transfer by agreement on a set of features partitions the P dimensions into N classes, predicting transfer exactly on the within-class pairs, $M(N) \geq P^2/N - P$. Those pairs are exactly where a structural rule differs from the null rule "no result ever

transfers", so a rule owing no tests is the null, and every unit of content is paid for at par: no structural rule obtains a discount. Under the default-universal reading (a dimension reported with no scope restriction is available for comparison with every other), what a suite's own text declares out of scope is the only thing that reduces what is owed; a bound on the number of classes an equivalence relation may have removes the growth in P, but without that a priori bound the class is every equivalence relation and the second row of the display above still grows.

Proposition 3 (The decomposition is exact, and D is a Jensen gap). Let $\{\mathrm{agr}(A,B)\}_{A\neq B}$ be the $P(P-1)$ per-pair agreements and $L=\sum_{A\neq B} H(\mathrm{agr}(A,B))$ the code length conditional on one Bernoulli parameter per pair. Then, with $\overline{\cdot}$ the mean over ordered pairs,

$$\frac{L}{P(P-1)} \;=\; \overline{H(\mathrm{agr})} \;=\; H(\overline{\mathrm{agr}}) - D, \qquad D \;\geq\; 0, \tag{2}$$

with $D=0$ iff every pair has the same agreement. $H(\overline{\mathrm{agr}})$ is a base rate; D is the only term about the pairs differing from one another.

Proof. H is strictly concave on $(0,1)$, so by Jensen $\overline{H(\mathrm{agr})} \leq H(\overline{\mathrm{agr}})$, equality iff agr is almost surely constant; define D as the gap. Nothing here is estimated: given the agreements, both sides are the same number computed two ways. □

Proposition 4 (When a per-pair parameter pays for itself). Code the n_{pair} model-pair comparisons behind each dimension pair. A two-part code charges $\frac{1}{2}\log_2 n_{\mathrm{pair}}$ bits for one fitted Bernoulli parameter per dimension pair and the parameter buys $n_{\mathrm{pair}}D$ bits of likelihood. Hence the parameters pay for themselves iff

$$D \;>\; D^\star(n_{\mathrm{pair}}) \;\equiv\; \frac{\log_2 n_{\mathrm{pair}}}{2\,n_{\mathrm{pair}}}, \tag{3}$$

and $D^\star$ is strictly decreasing in n for $n \geq 3$. The threshold is a family indexed by the corpus: at $n_{\mathrm{pair}} = 952$ it is 0.52% and across the eight corpora of §6.2 it runs from 0.52% down to 0.02%.

Proof. The two-part code length is $\frac{1}{2}\log_2 n_{\mathrm{pair}}$ per fitted parameter plus the negative log-likelihood; the gap between the per-pair and single-base-rate model is $n_{\mathrm{pair}}D$ bits per pair, by (2). Comparing the two gives (3); monotonicity follows since $\frac{d}{dn}(\log_2 n/2n) < 0$ for $n > e$. □

Proposition 5 (Λ's own resolution, and the dimension count it demands). Let $\Lambda = 1 - H(e^\star)$ where $e^\star$ is the best partition's error rate on a relation over $K = P(P-1)$ ordered pairs. Write $x = |\frac{1}{2} - e^\star|$ for the margin. Then

$$\Lambda \;=\; \frac{2}{\ln 2}\,x^2 + O(x^4), \qquad q_\Lambda \;\equiv\; \max_{\pm} \big|\Lambda(e^\star) - \Lambda(e^\star \pm 1/K)\big| \;=\; \frac{4}{\ln 2}\cdot\frac{x}{K} + O\big(x/K^2\big), \tag{4}$$

so requiring one lattice step to stay within a fraction α of the value being reported holds iff

$$\boxed{K \;=\; P(P-1) \;\geq\; \frac{2}{\alpha\,x} \;=\; \frac{2}{\alpha}\Big(\frac{2}{\ln 2\cdot\Lambda}\Big)^{1/2}} \tag{5}$$

to leading order. Read forwards, (5) is a design rule: a suite that intends to report a label-predictability of Λ needs at least that many ordered pairs, whatever its item budget and before any data exists.

Proof. Put $e^\star = \frac{1}{2} - x$. Then $H(\frac{1}{2} - x) = 1 - \frac{2}{\ln 2}x^2 - O(x^4)$ by the Taylor expansion of H at $\frac{1}{2}$. For the second: $|dH/de| = \log_2((1-e)/e) = \frac{4}{\ln 2}x + O(x^2)$, and one lattice step is $1/K$. Substituting into $q_\Lambda \leq \alpha\Lambda$ gives $\frac{4}{\ln 2}x/K \leq \alpha\frac{2}{\ln 2}x^2$, i.e. $K \geq 2/(\alpha x)$; inverting the quadratic for x gives the second form. □

Proposition 6 (Exponent doubling: the no-structure baseline for Λ). Suppose the margin scales as $x \propto m^{-\beta}$ under a refinement by a factor m. Then $\Lambda \propto m^{-2\beta}(1 + O(x^2))$. With the item budget pinned, pure sampling noise gives $\beta = \frac{1}{2}$ and hence $\Lambda \propto m^{-1}$: each part is measured on $1/m$ of the items, a difference's standard error grows as $\sqrt{m}$, and the resolvable margin shrinks by the same factor.

Proof. Substitute $x \propto m^{-\beta}$ into $\Lambda = \frac{2}{\ln 2}x^2 + O(x^4)$. At a pinned budget a part carries $I/(mP)$ items, so a per-part standard error scales as $\sqrt{m}$, and the smallest resolvable difference scales with it; the error rate of the best partition therefore approaches $\frac{1}{2}$ at rate $m^{-1/2}$ under no structure. □

This is the baseline §6.1 measures against. Measured, the margin falls as $m^{-0.599}$, twice that is -1.199, and Λ's fitted exponent is -1.223: the two agree to 0.02. The measured -1.223 exceeds Prop. 6's -1 by 0.223, reported as unexplained.

4 The measurement protocol

Four choices govern every number below. The transfer fact $T(A, B)$ records whether the system ahead on A is ahead on B; a system pair enters only if its difference on both dimensions exceeds twice the combined standard error. System pairs are additionally matched on overall accuracy to within $\tau = 0.02$, since without it the measurement returns the corpus's general factor. The real-valued agreement becomes a binary relation at $\theta =$ the median agreement of that corpus at that rung, so the thresholded relation has base rate exactly $\frac{1}{2}$ everywhere and no selection cost is owed. Λ_{out} splits the admitted system pairs in half, searches on one half and scores on the other, and is the mean over twenty independent such splits.

5 Transfer relations priced on public corpora

Whether the laws above bite is an empirical question about how far a real transfer relation sits from a labelling. 690 models from the public Open llm Leaderboard results archive, each carrying all 57 mmlu subjects. No private repository and no unpublished companion is required.

Table 1. What a real transfer relation costs, and what a confound does to the answer. The first row is the measurement a programme makes if it compares whatever systems it has; the last is the measurement it makes if it compares systems of matched overall ability. The share is not the discriminating quantity and the last column is: by concavity of H the share splits exactly into a base-rate code and a structural term, and the base rate is the larger part in every row. The drop from 19.3% to 93.9% is 93.4% a base-rate effect. What controlling for general ability actually does is raise the structural term, from 2.68% to 6.02%, and to 22.77% once attenuation is removed.

model pairs compared	agr	$H(\overline{\text{agr}})$ (%)	share (%)	structure D (%)	pairs used
all	0.9648	21.98	19.3	2.68	385,755,104
overall ability within 0.05	0.6771	90.75	84.1	6.68	18,258,296
overall ability within 0.02	**0.5143**	99.94	**93.9**	**6.02**	4,797,464
the same, de-attenuated	0.5198	99.89	77.1	**22.77**	—

The apparent transfer is a general factor, and controlling it inverts the conclusion. Compared without restriction, one subject's advantage predicts another's 96.5% of the time; the transfer was the models' general ability, not the subjects. At matched ability the relation costs 93.9% of the maximum-entropy price, but the legal reference for that hypothesis is a two-part code that pays for its own errors, not a zero-error one: an equivalence relation priced with its own errors costs 99.8%, so an equivalence hypothesis buys 0.2%, not 94%. Against a code that already knows the base rate, the best structural gain over every branch and threshold is +4.24%, swept over the free threshold θ and found to be the null rule wearing a threshold at both ends of the sweep, Law 3 (§3) appearing in data we did not design to show it.

What this establishes. On a public corpus of 690 systems and 57 labelled dimensions, measured rather than assumed, this corpus is in the maximum-entropy branch of Prop. 2, a statement about mmlu at 57 subjects and not about suites, which §6.2 makes unavoidable: on one of the eight corpora a label predicts 53% of the relation. Against a code that already knows the base rate, the best structural gain over every branch and threshold is $+\mathbf{4.24}\%$, far smaller than the contrast this section used to draw. Multiplying $\frac{1}{2}\log_2 n_{\text{pair}}$ by the 3192 parameters and comparing it against a one-observation likelihood misprices by construction; priced consistently (Prop. 4) a per-pair parameter buys $n_{\text{pair}}D = 57.3$ bits for its 4.95, so mdl keeps every one, and the same arithmetic rejects a permutation null's parameters at 0.82×.

The structural term clears its own break-even; a permutation of it does not. Two nulls matter most (Appendix D carries all four), drawn in Figure 2 against the break-even $D^\star$ of Prop. 4: h-additive reproduces every main effect with no model-by-subject interaction, and h-perm fits that same model and permutes the residuals across subjects, killing only the association. The observed term clears $D^\star$ by 11.58×; h-perm does not clear it at all (0.82×).

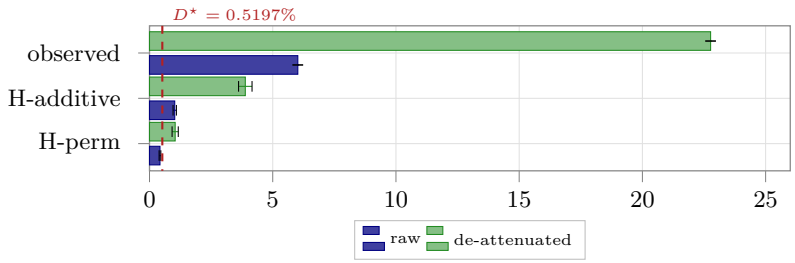


Figure 2. The two nulls that matter most, against the break-even they both have to clear. Error bars: two null standard deviations, 200 draws each. De-attenuation moves every bar and no verdict. All four nulls are in Appendix D.

6 The refinement limit on a published ladder

§2 pins a corpus's item budget and sends the dimension count to infinity by splitting dimensions. This section answers it twice. Backwards, on a ladder of granularities somebody else built: mmlu's authors published a three-level taxonomy in 2020 (57 subjects, 17 subcategories, 4 categories), and coarsening pools subjects rather than adding questions, so the item budget is the same at every rung and the ladder is the limit run in reverse. Forwards (§6.1), by splitting each subject's own items into m disjoint parts, the limit itself with a truth held fixed.

Table 2. The refinement limit read backwards, with every rung against its own resolution. q_Λ is how far one step of the $e^\star$ lattice moves Λ; a rung whose single step exceeds the band being read is not decidable. The $P = 4$ rung fails it by a factor of nearly three, so "flat" is not a reading this table licenses; on the two rungs that can resolve the question, Λ falls, 9.98% → 6.02% as P goes 17 → 57.

rung	P	D (%)	$e^\star$	Λ_{out}	q_Λ (%)	verdict
mmlu category	4	4.03	0.3333	7.45 ±2.23	10.70	not decidable
mmlu subcategory	17	5.72	0.3162	**9.92** ±2.32	0.41	decidable
mmlu subject	57	6.02	0.3565	**6.00** ±0.42	0.03	decidable
bbh subtask	24	4.56	0.2138	**24.94** ±0.76	0.34	decidable

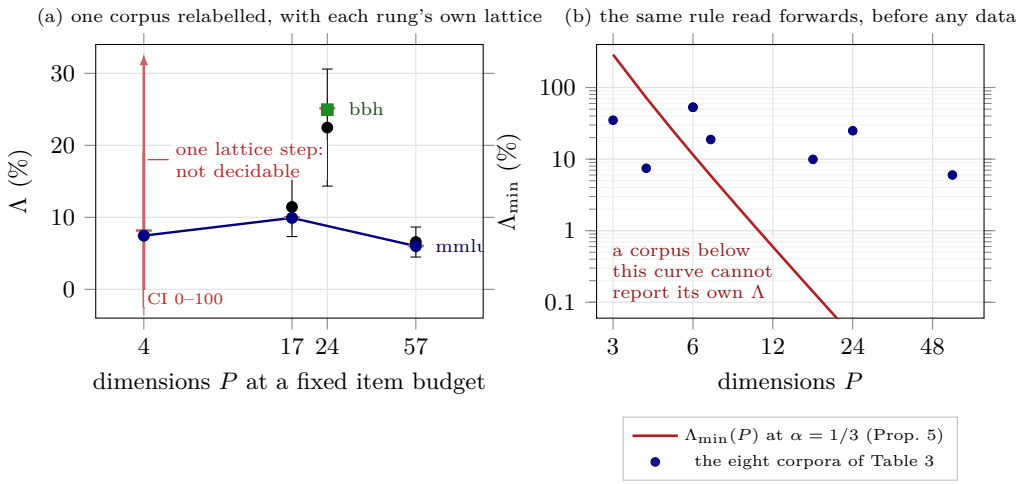


Figure 3. Resolution, measured backwards and then turned into a design rule. (a) mmlu's own published taxonomy read as the refinement limit in reverse. That rung is not decidable, and on the two rungs that can resolve the question Λ falls, 9.98% $\to$ 6.00%. (b) The same inequality solved for P instead of checked after the fact, the one thing here a benchmark author can use before collecting anything.

The rung that fails its own resolution check is excluded, and the rule that excludes it is ours. At $P = 4$ there are 12 ordered pairs, so $e^\star$ moves in steps of 1/12: one step of the instrument moves Λ by 10.70 points against the 3.96-point band being read, exactly the condition under which Prop. 5 returns not decidable. With the rung gone, the remaining rungs say Λ falls by 40% relative as the labelling refines from 17 dimensions to 57 – this paper's own limit confirmed rather than only bounded, not a weaker result than "flat".

6.1 The same limit with the truth held fixed

Splitting each subject's item budget into m disjoint parts by sequential hypergeometric draws holds the true relation between two different subjects' parts fixed at the $m = 1$ relation, so every change in measured Λ is resolution rather than a moving truth.

With the item budget pinned, measurable label-predictability follows $\boldsymbol{\Lambda \propto \mathbf{m}^{-1.223}}$, from 6.40% at $m = 1$ to 0.51% at $m = 8$, while the pair count grows 64× and the truth of the pairs being counted has not moved at all. The no-structure baseline (Prop. 6) predicts m^{-1}; the measured margin goes as $m^{-0.599}$, twice that is -1.199 against Λ's fitted -1.223, so the quadratic relation holds to 0.024 in the exponent. D falls too, from 8.08% to 2.59%, refuting our own registered expectation that it would not.

The two directions disagree about D and agree about Λ, and the disagreement identifies the mechanism. Refining real labels (17 $\to$ 57 subjects) raises pair heterogeneity while label-predictability falls; refining by construction ($m = 1 \to 8$) lowers both. A random split adds no structure, so its falling D is resolution loss and nothing else, which makes the real ladder's rising D content that a finer labelling genuinely exposes: refinement finds more real heterogeneity between pairs, and less of it that a label can name.

6.2 Eight corpora and one registered hypothesis

Six more corpora are available in the same public archives at no new cost. bbh's dimensions are named by what a system must do; mmlu's are named by what its items are about. A

corpus named by topic against one named by task was frozen as a hypothesis before any of the six new corpora was computed.

Table 3. Λ_{out} on eight corpora, with each row against its own resolution. The ordering is complete: all four topic rows fall below all four task rows. What that ordering is worth depends on what a unit is: mmlu's three rungs are one item pool relabelled and each leaderboard suite's six dimensions are other rows of this table, so four independent units remain and the exact one-sided probability is $1/\binom{4}{2} = 0.167$. Every row pays, from 11.6× to 741× its mdl break-even.

corpus	kind	P	Λ_{out}	mdl	q_Λ/Λ	verdict
mmlu subject (unit)	topic	57	$\mathbf{6.00 \pm 0.42}$	11.6×	<0.01	decidable
mmlu subcategory (rung)	topic	17	9.92 ± 2.32	18.0×	0.04	decidable
mmlu category (rung)	topic	4	7.45 ± 2.23	23.9×	1.31	not decidable
math Lvl-5 (unit)	topic	7	$\mathbf{18.81 \pm 4.59}$	28.7×	0.19	decidable†
bbh subtask (unit)	task	24	$\mathbf{24.94 \pm 0.76}$	69.8×	0.01	decidable
musr (unit)	task	3	35.00 ± 0.00	741×	0.77	not decidable
Open llm v1 (nested)	task	6	53.10 ± 0.00	273×	0.22	decidable
Open llm v2 (nested)	task	6	53.10 ± 0.00	586×	0.22	decidable

The topic/task ordering is not supported, and the three reasons are each worth more than the hypothesis would have been. First, the exact test's n is the number of independent item pools, not rows: $1/\binom{4}{2} = 0.167$ once nesting is applied. Second, the confound runs the same way as the hypothesis (topic corpora average $P = 21.2$ against task corpora's 9.8, and small P reads high because its lattice is coarse), so Spearman$(\Lambda_{\text{out}}, P)$ is evidence of the confound. Third, the one pair that controls for P without reference to Λ (bbh P=24 against mmlu subcategory P=17) has intervals overlapping by 1.22 points. What the eight corpora do establish: Λ_{out} ranges from 6.00% to 53.10%, so there is no single number for what a benchmark's labels predict, and the strong reading of §5 is a statement about mmlu and not about suites in general. What survives on all eight is the order rather than the constant, and the mdl verdict: a per-pair parameter pays for itself everywhere.

7 What a benchmark author can do on Monday

Table 4. What a suite can do about $P(P-1)$ without running a transfer test. Rows one to four are one command on a matrix a suite already has; row six is the only one usable before a single model has been run.

what to do	what it prevents
Publish Λ_{out} and q_Λ beside the dimension table.	a reader supplying the generalisation for free
If Λ_{out} is below 10%, say so in one sentence.	mmlu's 6.00% being read as a licence to carry a method's advantage across subjects
Report per-dimension item counts and standard errors, not scores alone.	every measurement in this paper being impossible on a suite that does not
If q_Λ exceeds Λ, report the dimension count as too small to answer the question.	the $P = 4$ rung of Table 2, published as a finding for two revisions
Pin the item budget before refining; choose P from the Λ you intend to report, $P(P-1) \geq 2/(\alpha\sqrt{\Lambda \ln 2/2})$.	an 8× refinement costing 92% of predictable label content, in advance and with no results matrix needed

8 Limitations

(i) The obligation reading rests on a default-universal reading of scope, and $q = 1$ makes every count a floor. (ii) Λ ranges by a factor of 8.85 across public suites, so every strong reading here is about the corpus it was measured on; four of the eight rows in Table 3 are

flagged, for two different reasons. (iii) The two uncertainties in Table 2 measure different things: split-to-split variation at fixed models, and model-resampling variation at a fixed split rule. (iv) The nulls over-state variance (item-level noise treated as independent across systems) but annotation error, not modelled, works against the structure we report. (v) The forward arm is semi-synthetic, uses 250 of the 690 models, and its parts are constructed by us, so it measures what the corpus's own accuracies imply about resolution rather than a second data collection.

9 Conclusion

The shortest true summary. A benchmark's labelled dimensions carry $P(P-1)$ ordered transfer readings, and what a label can predict of that relation out of sample runs from **6.00**% on mmlu to **53.10**% across a leaderboard's own six benchmarks, a range, not a constant; the rest is base rate. The escape is not available, which is this paper's limit answering itself: with the item budget pinned, an 8× refinement takes label-predictability from 6.40% to 0.51% while the pair count grows 64×, and mmlu's own taxonomy falls the same way, 9.98% → 6.02% from 17 dimensions to 57. Refinement multiplies the claims and divides the evidence. The residual relation is nonetheless real: a per-pair parameter needs $D > 0.52\%$ to pay for itself, the observation pays 11.58×, and a permutation null pays 0.82× and is rejected.

What remains open needs task-labelled suites at $P \geq 20$ publishing per-dimension item counts (the archives today hold exactly one); the forward arm needs no new suite, and if Λ falls with m on three more, the resolution wall belongs to labelled evaluation and not to mmlu alone.

Reproducibility statement

Every plotted coordinate and every cross-paper number is written by a script from the result JSONs listed in the supplementary material, and no figure or table contains a typed number. The eight public corpora, the leaderboard archives, and the scoring tool (refine_arms.py) that reproduces D, Λ_{in}, Λ_{out} and q_Λ for any models-by-dimensions matrix are included in the anonymised supplementary material, with sha256 hashes so a reader can verify a downloaded archive matches the one this paper was computed from. Every criterion graded in Appendix M was frozen in a preregistration file before its data existed; the appendix states the counting rule and the freeze commit for each.

Ethics statement

Every corpus used is a public leaderboard results archive or a published benchmark's own item taxonomy; no human data was collected and no individual is identified or evaluated. The models scored are openly released or accessed through their public leaderboards under their standard terms; we introduce no new capability, only a measurement of how much a benchmark's own dimension labels predict about a real transfer relation. 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. Every result reported here is computed by a script the authors wrote, from public archives, and every number is checked against the result file it cites. 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.

References

Anonymous. Position: Preregistration Protects the Criterion, Not the Comparison: One Programme Audited Against Its Own Laws Seven Times. concurrent manuscript; author list and repository withheld for double-blind review, 2026a.

Anonymous. Two Counts over Text: What a Memory Benchmark Can See, Measured Before Anything Is Run. concurrent manuscript; author list and repository withheld for double-blind review, 2026b.

Anonymous. Resolving Power: Whether a Memory Task Can See the Distinction It Is Cited For, and a Construction Whose Resolving Power Is Provably Positive. concurrent manuscript; author list and repository withheld for double-blind review, 2026c.

Anonymous. The Infinite-Record Limit: A retrieval budget that is a constant share of the record for the hardest items and a vanishing one for the median, and an exponent that belongs to the lever. concurrent manuscript; author list and repository withheld for double-blind review, 2026d.

Anonymous. The Infinite-Rerun Limit: Computing what a criterion can decide, before it is frozen: a resolution, a rule, and the flip rate a band buys. concurrent manuscript; author list and repository withheld for double-blind review, 2026e.

Anonymous. What a Rubric Judge Measures: Two Defects in a Shipped Scorer, the Ceiling on Perfect Provenance, and the Credit Rate That Prices the Step from a Count to a Bound. concurrent manuscript; author list and repository withheld for double-blind review, 2026f.

Andrew Barron, Jorma Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. IEEE Transactions on Information Theory, 44(6):2743–2760, 1998.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, et al. Measuring what matters: Construct validity in large language model benchmarks. In Proc. 39th Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2025. arXiv:2511.04703; 445 benchmarks reviewed by 42 authors.

Robert L. Brennan. Generalizability Theory. Springer, 2001.

Ryan Burnell, Han Hao, Andrew R. A. Conway, and Jose Hernández-Orallo. Revealing the structure of language model capabilities, 2023a. arXiv:2306.10062.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al. Rethink reporting of evaluation results in AI. Science, 380(6641):136–138, 2023b.

David M. Cohen, Siddhartha R. Dalal, Michael L. Fredman, and Gardner C. Patton. The AETG system: an approach to testing based on combinatorial design. IEEE Transactions on Software Engineering, 23(7):437–444, 1997.

Charles J. Colbourn. Combinatorial aspects of covering arrays. Le Matematiche, 59(1–2): 125–172, 2004.

Lee J. Cronbach, Nageswari Rajaratnam, and Goldine C. Gleser. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16(2):137–163, 1963.

Maria Eriksson, Erasmo Purificato, Arman Noroozian, João Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. Can we trust AI benchmarks? an interdisciplinary review of current issues in AI evaluation. arXiv preprint arXiv:2502.06559, 2025.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open LLM leaderboard v2. Hugging Face, 2024.

Peter Grünwald and Teemu Roos. Minimum description length revisited. International Journal of Mathematics for Industry, 11(1), 2019.

Peter D. Grünwald. The Minimum Description Length Principle. MIT Press, 2007.

Shelby J. Haberman. When can subscores have value? Journal of Educational and Behavioral Statistics, 33(2):204–229, 2008.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In International Conference on Learning Representations, 2021.

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. Pass@1 varies by up to 15 points across 20 seeds on a small benchmark; the cause named is the item count, which is our sampling term rather than our apparatus term.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A. Smith. Fluid language model benchmarking. In Proc. Conference on Language Modeling (COLM), 2025. arXiv:2509.11106.

David Ilić and Gilles E. Gignac. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? Intelligence, 106:101858, 2024. doi: 10.1016/j.intell.2024.101858.

Han Jiang, Sunbeom Kwon, Jinwen Luo, Ziang Xiao, and Susu Zhang. Can we trust item response theory for AI evaluation? arXiv preprint arXiv:2607.15190, 2026.

D. Richard Kuhn, Raghu N. Kacker, and Yu Lei. Practical combinatorial testing. Technical Report NIST Special Publication 800-142, National Institute of Standards and Technology, 2010.

Erik Larsen. The instability of safety: How random seeds and temperature expose inconsistent LLM refusal behavior. arXiv preprint arXiv:2512.12066, 2025. 876 harmful prompts, four models, twenty sampling configurations: 18–28% of prompts show at least one decision reversal, and the stability index at temperature zero is 0.977. An independent measurement of a score's own resolution, on the axis this paper does not use: seeds rather than the dimension lattice, and its recommendation — at least three samples — is mitigation rather than a resolution.

Yanshi Li, Xueru Bai, Shuman Liu, and Long Zhang. Repbench: Compiling benchmarks into capability representations for large language models. arXiv preprint arXiv:2607.28008, 2026. A taxonomy of 182 capability clusters in 13 families crawled from 13,427 benchmark papers, and 46,149 audited probe texts covering 94 capabilities each supported by at least two independent benchmarks. Raw per-text vectors show no natural cluster granularity while benchmark-pooled capability vectors show an interior clustering optimum at a small number of clusters on all twelve models evaluated, with low agreement to the human taxonomy; the readout method decides which capabilities win.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. Transactions on Machine Learning Research, 2023.

Aviya Maimon, Amir D. N. Cohen, Gal Vishne, Shauli Ravfogel, and Reut Tsarfaty. From benchmarks to skills: Low-rank factors for llm evaluation. arXiv preprint arXiv:2507.20208, 2026. Factor analysis of a 60×44 model-by-benchmark performance matrix, finding an intrinsically low-rank structure whose latent factors read as skill-like dimensions, and turning it into three tools: identifying redundant tasks, profiling a new model from a small subset, and selecting models by skill profile. Submitted 2025-07-27, revised 2026-06-11.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. arXiv preprint arXiv:2411.00640, 2024.

Bitya Neuhof and Yuval Benjamini. Rank intervals for leaderboards: A hierarchical framework for model evaluation. arXiv preprint arXiv:2606.08679, 2026.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In Proc. 41st International Conference on Machine Learning (ICML), 2024. arXiv:2402.14992.

Qi Qian, Chengsong Huang, Jingwen Xu, Changze Lv, Muling Wu, Wenhao Liu, Xiaohua Wang, Zhenghua Wang, Zisu Huang, Muzhao Tian, Jianhan Xu, Kun Hu, He-Da Wang, Yao Hu, Xuanjing Huang, and Xiaoqing Zheng. Benchmark^2: Systematic evaluation of LLM benchmarks. arXiv preprint arXiv:2601.03986, 2026.

Steven P. Reise. The rediscovery of bifactor measurement models. Multivariate Behavioral Research, 47(5):667–696, 2012.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In Advances in Neural Information Processing Systems, Datasets and Benchmarks Track, 2024.

Jorma Rissanen. Modeling by shortest data description. Automatica, 14(5):465–471, 1978.

Jorma Rissanen. Fisher information and stochastic complexity. IEEE Transactions on Information Theory, 42(1):40–47, 1996.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In Advances in Neural Information Processing Systems, 2024. arXiv:2405.10938.

Sandip Sinharay. How often do subscores have added value? results from operational and simulated data. Journal of Educational Measurement, 47(2):150–174, 2010.

Sandip Sinharay, Gautam Puhan, and Shelby J. Haberman. An NCME instructional module on subscores. Educational Measurement: Issues and Practice, 30(3):29–40, 2011.

Charles Spearman. "general intelligence," objectively determined and measured. The American Journal of Psychology, 15(2):201–292, 1904.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Findings of the Association for Computational Linguistics: ACL 2023, 2023.

Yuchen Zeng and Dimitris Papailiopoulos. You don't need to run every eval, 2026. arXiv:2606.24020.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In Proc. AAAI Conference on Artificial Intelligence, 2026. arXiv:2505.15055.

Richard E. Zinbarg, William Revelle, Iftah Yovel, and Wen Li. Cronbach's α, revelle's β, and mcdonald's ω_h: their relations with each other and two alternative conceptualizations of reliability. Psychometrika, 70(1):123–133, 2005.

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 the extended related work, the five checks on the headline measurement and its internal replication inside one benchmark, the four nulls in full, the two-codings table in full, the forward-ladder rungs, the notation table, the preregistration record, and the public scorecard of every graded row. Reviewers are not required to read it.

A Extended Related Work

Seven literatures price something adjacent to what this paper prices, and the question asked of each is what it assumes about the quantity it economises. One acquired a new member while this paper was being written, and the paragraph on it is read as contemporaneous work.

Benchmark meta-evaluation measures whether a suite is a good instrument; nobody has priced the transfer claims its dimension table makes. The construct review of 445 benchmarks by 42 authors finds 16% using uncertainty estimates or statistical tests (Bean et al., 2025); Reuel et al. (2024) score 24 benchmarks against 46 best-practice criteria and Eriksson et al. (2025) review the field's systemic problems, both finding that documentation and statistical treatment are where suites fail; helm (Liang et al., 2023) standardises what is reported across scenarios; cross-benchmark consistency (Qian et al., 2026), item-response scoring of benchmarks themselves (Zhou et al., 2026) and hierarchical rank intervals (Neuhof & Benjamini, 2026) all quantify how reliably a suite ranks systems; and Burnell et al. (2023b) argue for reporting instance-level results, the input every measurement here needs and which the archives we use publish only in aggregate. A suite can rank perfectly and still license nothing between its own dimensions: a ranking is a statement about systems, a transfer reading a statement about an ordered pair of dimensions. That is the cell this paper measures, and §5 is its first price.

Efficient evaluation moves the item budget, which is the exact resource this paper's limit pins. Item-response subsetting estimates full-benchmark performance from roughly 100 of 14K items (Polo et al., 2024) and adaptive selection reports higher validity with fifty times fewer items (Hofmann et al., 2025), with the trustworthiness of the machinery on these data still open (Jiang et al., 2026). Read against §2, those results are the supply side of the very trade-off measured here: an item budget spent on more dimensions is not spent on resolving the ones already labelled, and §6 measures what that costs.

Minimum description length is the currency. Two-part codes and the $\frac{1}{2}\log_2 n$ price of a fitted parameter are Rissanen's (Rissanen, 1978; Barron et al., 1998; Grünwald, 2007), with the normalised-maximum-likelihood refinement in Rissanen (1996) and the modern account in Grünwald & Roos (2019); Prop. 4 uses only the two-part form, the conservative one here since a sharper code would raise the payoff we report. That penalty is calibrated for n observations per parameter, and a version of Table 7 paid an $n = 952$ penalty against a likelihood coding one outcome per pair. The repair is in that table and the defect recorded rather than absorbed.

The added value of a subscore is this paper's question, asked of a psychological test decades earlier and in a different currency. A dominant general factor is Spearman's (Spearman, 1904), its modern decomposition is the bifactor model (Reise, 2012) with ω_h measuring how much of a total score the general factor carries (Zinbarg et al., 2005), and Haberman (2008) gives the criterion for when a subscore beats the total score as a predictor, with

Sinharay (2010) finding that operational subscores usually do not clear it. That is D and Λ in another notation, and the connection cuts against reading this paper as surprising: benchmark dimensions behaving like test subscales is what a psychometrician would expect. What is not in that literature is the limit. Haberman's criterion applies to a fixed test with fixed subscale lengths; §2 pins the item budget and sends the number of subscales to infinity, and §6 measures the added value driven towards zero by the budget rather than by any absence of structure. A suite can refine its labels until its subscores cannot have value, and no amount of real structure prevents it.

The low-rank line now has an LLM-native member with tools attached, and the tool it ships is the one our limit prices differently. Maimon et al. (2026) apply factor analysis to a 60×44 model-by-benchmark matrix, find it intrinsically low rank, read the latent factors as skill-like dimensions, and deliver three tools, identifying redundant tasks, profiling a model from a small subset, selecting models by skill profile. This is the closest published statement to §5's object and we take the finding as established, as §1 already does for the general factor (Ilić & Gignac, 2024; Burnell et al., 2023a; Ruan et al., 2024). The two questions are not the same and neither refines the other. A factor model asks how many latent dimensions describe this matrix, and its answer is set by a variance criterion (eigenvalues, a scree, parallel analysis), while Prop. 4 asks whether one named per-pair parameter earns the $\frac{1}{2}\log_2 n$ bits it costs. A factor can carry variance and a label still fail to pay, and the second question is the one a dimension table's transfer claim answers to. Their redundancy verdict is read at a fixed task count, which is where our limit differs from a disagreement: §2 pins the item budget and refines the labels, and §6 measures the added value driven toward zero while the truth is held constant by construction. Hence "these tasks are redundant" and "this label cannot be shown to pay" can both be true of a suite with all its structure intact, and only the first is a statement about the world.

The closest result on the practical side predicts scores from a rank-2 matrix, and this paper's headline is what that rank-2 structure leaves over. Zeng & Papailiopoulos (2026) assemble 84 models against 133 benchmarks, find the matrix approximately rank 2, and ship BenchPress, a matrix completion that fills in missing scores and returns the smallest benchmark subset estimating the rest. The estimands differ, and the relation between them is a decomposition rather than a disagreement. Completing a model-by-benchmark matrix predicts how well a system scores; §5 asks whether a label can name which ordered pairs the system's advantage carries across, and prices that in bits. A dominant low-rank factor is exactly what makes the base-rate term large: it puts most systems in the same order on most dimensions, which is why the relation looks to cost 19.3% of $P(P-1)$ when systems are compared as they come, and why 93.4 of the 93.9 points it costs at matched ability are base rate (Table 1). A rank-2 completion can therefore be highly accurate on a corpus whose Λ is 6%: the two quantities are the two halves of one split, and only the second is what a dimension table claims. Their subset-selection tool and Prop. 5's design rule are likewise complements rather than rivals: theirs says which benchmarks need not be run, ours how many dimensions a suite needs before the number it reports can be read at all.

One 2026 result measures a granularity optimum from the representation side, which is the closest thing to external corroboration this paper has. Li et al. (2026) crawl 13,427 benchmark papers into a taxonomy of 182 capability clusters in 13 families, harvest 46,149 probe texts covering 94 capabilities each supported by at least two independent benchmarks, and report three things this paper should not be allowed to ignore: raw per-text vectors show no natural cluster granularity; benchmark-pooled capability vectors show an interior clustering optimum at a small number of clusters, on all twelve models; and that optimum has low agreement with the human taxonomy. The third is the premise §5 prices, arrived at from representations rather than scores: a suite's dimension table is not where the geometry's own optimum sits. We claim the second as corroboration only in the weak sense the difference allows, and the difference is ours to state: their optimum is in a clustering objective with no item budget in it, so it is not §2's limit, and an interior optimum in their criterion does not imply Λ's decay in ours. Two objectives on two corpora agreeing that granularity is not free is worth one sentence. Their fourth finding is a warning we pass on rather than answer: which capabilities "win" depends on the readout method, difference-in-means

against logistic regression, so a capability geometry is a property of a (corpus, readout) pair the way a judged score is a property of a (family, scorer) pair (Anonymous, 2026f).

Combinatorial testing already beat a quadratic obligation, under the condition this paper measures. Covering arrays test all t-way interactions in $O(\log)$ many runs (Cohen et al., 1997; Kuhn et al., 2010), with the lower-bound side surveyed by Colbourn (2004), and this collapses $P(P-1)$ when the interaction structure is known and the relation is an equivalence. Prop. 2 states what that assumption is worth in bits and §5 measures that it does not hold on the corpora where it can be checked.

Resolution is measured in this literature, and never for the quantity a dimension table claims, which is the gap Prop. 5 fills. That benchmark scores carry sampling error large enough to reorder systems is now well documented: Miller (2024) derives the error bars a score needs, Hochlehnert et al. (2025) shows reasoning-benchmark rankings turning over under seed and prompt variation, Larsen (2025) measures the instability directly, and Neuhof & Benjamini (2026) converts it into rank intervals. Every one is about the resolution of a score or a rank. A dimension table's claim is neither: it is a statement about ordered pairs of dimensions, whose statistic Λ lives on a lattice of spacing $1/P(P-1)$ and so has a resolution of its own that no item-level precision improves. Prop. 5 gives that resolution in closed form, and it decides which rungs of §6 may be read at all. Polo et al. (2024) and Hofmann et al. (2025) spend the same item budget better; nobody has said how many dimensions a suite needs before a label-predictability figure means anything, which is what (5) answers.

Reliability theory has the variance decomposition this paper's D is a discrete case of. Generalizability theory (Cronbach et al., 1963; Brennan, 2001) partitions a score's variance across facets (person, item, occasion) and asks which facet a decision may be generalised over, structurally the question §5 asks of a (dimension, dimension) pair. Sinharay et al. (2011) give the operational version for subscores. What that literature does not do is pin the item budget and let the facet count grow, and §2 is exactly that limit.

B Five checks on that measurement

The measurement of §5 is the strongest result in this programme, and five ways it could be an artifact are put to it here. It survives the first three; the fourth narrows the hypothesis class the result may be stated over, and the fifth forces two disclosures. Criteria for the first four are in prereg/E_NOISE_PREREG.md with one addendum; the fifth, and a revision to how the third is reported, are in E_NOISE_ADDENDUM3.md, frozen before noise_nulls.py existed and marked post hoc. The arm rebuilds the measurement from the 899 raw result files rather than reading a stored array, and reproduces §5's unrestricted agreement to 2.98×10^{-8}, the precondition for reporting anything below.

(1) Is the result guaranteed by binomial noise? mmlu subjects run from 100 to 1534 questions (Hendrycks et al., 2021), so at the small end a subject standard error is near 0.05 while a real specialisation may be 0.02. Two null truths are resampled at each subject's own question count and pushed through the identical pipeline: H-flat, in which a model scores its overall accuracy on every subject, and H-additive, a model effect plus a subject effect, which has subject difficulty but no model-by-subject interaction and therefore no transfer structure of any kind.

P-T2-N1 holds, and it retires one of our own sentences. The observed entropy lies below every null, so a model-by-subject interaction is present and 93.9% is not what pure noise would have produced. The additive null returns a mean agreement of **0.5142** against the observed 0.5143.

P-T2-N7 holds, and the like-for-like comparison moves the result in our favour, which is not the direction we registered. §B's attack (3) de-attenuated the observation and compared it against an uncorrected null, so the conclusion quantity was never com-

puted. Computed, the de-attenuated additive null sits at 0.9591 against the observed 0.7712: the like-for-like margin is **0.1879** where the illegitimate comparison showed 0.0500, because de-attenuation moves the observation by 0.168 and the null by only 0.030. Our registered expectation was that the observation would end up above the corrected null, on the reasoning that a no-interaction truth has a perfectly consistent sign; it does not, and the expectation is refuted while the criterion passes. Reading the pipeline afterwards says why, and the reason is worth more than the expectation was. Under H-additive at matched ability the true per-subject difference has median 0.0075 against a sampling standard deviation near 0.060, so the flip probability on the pairs clearing the gate sits at $f = 0.257$ and the correction is $0.5 + (p - 0.5)/(1 - 2f)$ with $1-2f = 0.482$: it rescales the observed deviation from chance by a factor of two, taking 0.5142 to 0.529 against the measured 0.5263. De-attenuation rescales an observation rather than inferring a truth, so it cannot manufacture resolution the statistic does not have, the companion paper's subject (Anonymous, 2026c) arriving inside our own correction. Our registered reasoning assumed the opposite. We registered it two-sided precisely because P-T2-I1 had been written one-sided in this same paper.

P-T2-N8 holds, and it is the comparison the section should have led with. The binomial nulls treat question-level noise as independent across models, when every model answers the same items; shared item difficulty partly cancels in a between-model difference, so independent resampling over-states that difference's variance, inflates the null and favours us. The archive carries aggregates only, so we cannot measure that bias, and it is disclosed rather than repaired. The permutation null needs no noise model: it reuses the corpus's own residuals and returns mean agreement 0.4958, i.e. chance, and mean $H = 0.9957$. The observation sits 337.6 null standard deviations below it after de-attenuation. On the structural term of §5, which is the quantity that means something, the separation is a ratio rather than a margin: the permutation null's D is 0.42% raw and 1.04% de-attenuated against the observed 6.02% and 22.77%, so the observation carries 14.17× and **21.86**× the structure a no-association relation shows, at **346.5** null standard deviations on D. Against the additive null, whose D is 1.02% raw and 3.89% de-attenuated, the ratios are 5.88× and 5.86×, at 138.3 null SD on D. Every separation in this paragraph is on D and every separation in Table 6 is on mean H: they are two statistics of one relation and their null SDs are not interchangeable, which is the distinction Prop. 1 asks for. Restricting the permutation to within question-count tertiles, so a residual from a 100-question subject cannot land on a 1534-question one, moves the separation to **399.7** and nothing else.

(2) Are 690 models 690 observations? No. Clustering by declared base family gives 426 clusters, 22 with more than one member and the largest with 100. Bootstrapping over clusters rather than models widens the 95% interval on mean H from $[0.9155, 0.9321]$ to $[0.9073, 0.9324]$, a factor of 1.51. The equivalence-relation share, 0.0584, sits far outside the widened interval, so the branch is identified. Two caveats: name-based clustering undermerges, so 1.51 is a lower bound on the widening; and the bootstrap distribution sits below the point estimate, since resampling creates duplicate models that cannot separate, so the interval is biased low and does not bracket 0.9392.

(3) What does attenuation cost? Correcting each sign for its own flip probability at the subject's standard error moves mean H from 0.9392 to **0.7712**: roughly a fifth of the measured cost is noise and the relation is genuinely cheaper than the raw number says. That number belongs beside the nulls corrected the same way, P-T2-N7 above, and not beside an uncorrected null, for the reason attack (5) gives.

(4) Is the hypothesis class complete? Not as first priced. §5 carried two discrete priors when this corpus is documented to be low rank and continuous (Ruan et al., 2024; Burnell et al., 2023a; Ilić & Gignac, 2024).

What is swept, and what each code pays for. The sweep is over θ, the cutoff turning the real-valued agreement into the binary relation being described: $T(A,B) = \mathbb{1}[\mathrm{agr}(A,B) \geq \theta]$, on the matched relation of Table 1's last row and separately on its de-attenuated counterpart. The grid is 19 quantiles of the observed agreement and $\log_2 19$ bits are charged for the choice, since an advocate of a cheap branch would pick θ. Each branch then pays for its parameters plus its residual, $P(P-1)H(e)$ with e its error rate against T: the vocabulary branch pays $P \log_2 N$ for a label per regime; the low-rank branch pays $2Prb + rb + b$ for a rank-r factorisation of the agreement matrix at b bits per coordinate; the accuracy-factor branch pays $Prb + Wb + b$ for subject coordinates taken from the top-r right singular vectors of the centred model-by-subject accuracy matrix and W logistic weights on $[x_A, x_B, x_A - x_B, x_A \odot x_B]$, the code an advocate of the documented low-rank structure would write. Ranges: $r \leq 10$, $b \in \{2,3,4,6,8\}$, $N \leq P$. Only the observed row pays for nothing, which is P-T2-N11 below.

P-T2-N6 is refuted as written, and the mechanism is not a cheap structural branch. Swept over θ, the cheapest description falls from 95.8% of $P(P-1)$ to **28.8**%. The winner at both ends of the sweep is nevertheless the null rule: at $\theta = 0.2865$ it predicts everything transfers and is wrong 5% of the time, and at $\theta = 0.7546$ it predicts nothing does and is wrong 95% of the time, the same code negated. The same thing happens on the de-attenuated relation at $\theta = 0.9493$. This is Law 3 appearing in data we did not design to show it. Prop. ?? says a rule owing nothing is the null rule, and the minimum-description-length search, given a free threshold, walks straight to it. Priced against what knowing the base rate alone buys, $P(P-1)H(\text{base rate})$, the best structural gain over every branch and threshold on the grid is $+$**4.24**%, attained at $\theta = 0.5154$ where the base rate is 0.500, and +5.19% on the de-attenuated relation. Computed inside this sweep the same gain is 4.2333% rather than Table 7's 4.2389%, and the difference is the $\log_2 19$ this grid charges on both sides, which cancels in the numerator of a fraction and not in its denominator, Table 7's caption gives the arithmetic, and the canonical value quoted everywhere else in this paper is the median rule's, which owes no such charge.

The criterion was mis-specified, and this is the sixth in this paper. P-T2-N6 asked whether any branch costs less than half of $P(P-1)$ and let the code choose the relation it was describing.

A defect in this arm's own design, found while running it. The preregistration set out to measure the growth exponent of the description length in P.

(5) Is the reference row a legal reference? No, and this objection has the largest effect on how the paper reads. Table 1 sets an error-bearing code, the observed 93.9%, against a zero-error one, $\log_2 \mathrm{Bell}(57) = 186.4$ bits or 5.8%, the price of naming a partition and nothing else. The legal reference for the same hypothesis is the two-part code: name the best partition, then pay for its errors, $\log_2 \mathrm{Bell}(P) + P(P-1)H(e^\star)$.

P-T2-D4 is refuted, and refuting it says where this paper's central measurement applies. The gain over a base-rate code was registered to fall monotonically in P, which would have made the mmlu/bbh difference a P effect. Computed on mmlu's own three rungs plus bbh it goes -24.39% at P=4, -3.35% at P=17, $+$**14.52**% at P=24 (bbh) and +0.19% at P=57: not monotone, and on mmlu it rises with P rather than falling. Reviewing the sweep's own arithmetic then found a confound in it, and removing the confound sharpens the answer. The gain is $1 - \log_2 \mathrm{Bell}(P)/P(P-1) - H(e^\star)$ at a balanced base rate, and the middle term is the price of naming a partition, a pure function of P that falls from 32.56% at P=4 to 5.84% at P=57 and has nothing to do with structure. Netting it out leaves the structural content $\Lambda = 1 - H(e^\star)$ at **8.17**%,

9.98%, **25.14**% and **6.02**% for $P = 4$, 17, 24, 57. The structural gain is not a monotone function of P, which is what the criterion was written to decide, and that is all P-T2-D4 refutes. A refuted monotonicity does not license the stronger reading "a property of the corpus and not of P", and §6.2 contradicts it: over the four independent item pools $\text{Spearman}(\Lambda_{\text{out}}, P) = -0.800$, so P and Λ are strongly and negatively associated even though the association is not monotone across these four rungs. The corpus and P are confounded here and §6.2 deals with it; this box establishes only that a monotone P effect is not what makes mmlu and bbh differ. §6.2 then measures the same quantity on eight corpora and finds it ranging from 6.00% to 53.10%. Two consequences we accept. The maximum-entropy finding is a statement about mmlu, where it holds at all three granularities and is stronger than we reported: at P=4 and P=17 the best equivalence code costs more than knowing the base rate, 124.4% and 103.4% of $P(P-1)$. §C's bbh arm replicates the share and not the conclusion. One of these four rungs is excluded downstream, and not by this criterion: §6 shows P=4 cannot resolve a band this narrow, so the 8.17% above is printed as the criterion computed it and is read as a measurement nowhere.

P-T2-N9 holds, and it moves this paper's structural gain from mmlu to bbh. On mmlu's matched relation the best partition errs on $e^\star = 0.3565$ of pairs, so the correct equivalence reference is 3186.1 bits, **99.8**% of $P(P-1)$, against the 5.8% printed, and 99.9% on the de-attenuated relation. An equivalence hypothesis on this relation buys 0.2%, not 94%, and the conclusion is no longer "the relation costs sixteen times what a label would" but "a label buys essentially nothing". Inside bbh the same computation gives 471.9 bits, 85.5%, so an equivalence prior over 24 dimensions buys **14.52**% against 0.19% over 57 subjects (both figures against the Bell reference, which is the one an equivalence hypothesis owes), and that is the largest structural gain among the four corpora of §6.2 that are independent item pools and can resolve their own Λ. It is not the largest in the paper: the two leaderboard-suite rows read 27.55% on the Bell reference and up to 33.10% on the vocabulary reference, above bbh on both. They are excluded here on two grounds frozen before they were computed (each is nested in other rows of the same table, and each fails Prop. 5), and stating the superlative without its population was itself the failure this programme's census counts (§8, and the ledger in Anonymous, 2026a). We do not quote the ratio of the two: 0.2% is the difference of two numbers printed to four figures and a ratio built on it would not survive its own rounding, the clause P-T2-D3 registered against us. Nothing here supports the reading that no structure exists; the measurements support gains of single-digit to mid-teen percentages of $P(P-1)$ rather than the collapse to $\Theta(P)$ that Prop. ?? requires.

Two disclosures the same reading forces. P-T2-N10: 96.3% in Table 1 and 95.8% here were two heuristic partition searches for one quantity at one θ, so both were upper bounds and neither was the vocabulary branch's price. Re-searched with twenty restarts and two agglomerative linkages the answer is 95.8%, so the canonical value is the smaller one and every vocabulary figure is written as a bound. P-T2-N11: $L = \sum_{A \neq B} H(\text{agr}(A, B))$ is a code length conditional on the $P(P-1)$ per-pair Bernoulli parameters, and only the observed row receives that subsidy while every structured branch pays for its parameters explicitly. The shares in Table 1 and Table 5 are comparable with each other and are not absolute description lengths. P-T2A-M1 prices that same observation consistently. Multiplying $\frac{1}{2}\log_2 n_{\text{pair}} = 4.95$ bits by the 3192 parameters gives 15,792 bits, but that sum pairs a penalty calibrated for 952 observations with a likelihood coding one and is not the saturated code's price. Priced consistently the conclusion inverts: a per-pair parameter buys $n_{\text{pair}} D = 57.3$ bits for its 4.95, so mdl keeps every one, and the same arithmetic rejects the permutation null's parameters at 0.82× (Table 7, Panel B).

C The same accounting inside one benchmark

§5 priced the relation across a suite's dimensions; the same accounting runs inside one benchmark, and a second corpus is the only way to tell a fact about transfer relations from a fact about mmlu. A suite with P labelled dimensions carries a transfer claim for every ordered pair of them, "a system that handles this dimension handles that one" being exactly the reading a dimension table invites, so the obligation is $P(P-1)q$ inside the artifact, with q counting the distinct things one might claim transfers (a score level, a ranking of systems, an ablation's direction).

For a ten-dimension suite that is ninety ordered pairs before q, and the companion's audit of one such suite reports no test of any (Anonymous, 2026b): the dimensions are reported side by side and the reader supplies the generalisation. An internal replication is the cheapest extension available and gives a second real data point, so it is run below.

The 552 ordered pairs inside bbh. Criteria are frozen in prereg/E_INTERNAL_PREREG.md with two disclosed addenda. big-bench Hard (Suzgun et al., 2023) is one benchmark with 24 labelled dimensions, hence 552 ordered internal transfer claims, and 895 public models on the second Open llm Leaderboard (Fourrier et al., 2024) carry all 24 at 146 to 250 questions each. An equivalence relation on 24 dimensions would cost 58.6 bits, 10.6% of $P(P-1)$.

Table 5. The transfer relation inside one benchmark, priced the way §5 prices it across mmlu. Leave-two-out matches a model pair on the mean of the other 22 dimensions, so the control removes the general factor without putting either compared dimension into the matching variable. internal_transfer.py.

model pairs compared	agr	mean H	share (%)	D
unrestricted	0.8151	0.6286	62.9	6.20%
matched within the benchmark	0.4730	0.9523	95.2	4.56%
matched leave-two-out	0.5224	**0.9525**	**95.2**	—
matched on ability outside	0.6834	0.8541	85.4	4.66%

reference: an equivalence relation priced with its own errors costs 471.9 bits, **85.5**%, so a label buys **14.52**% here against the Bell reference and 20.79% against the vocabulary reference, the largest of the four resolvable, non-nested corpora on either, and not the largest in the paper (§6.2). P-T2-D4 shows it is not monotone in P; §6.2 shows P is confounded with it. The 58.6 bits / 10.6% figure is the zero-error code

The result replicates, and §C's claim moves from untested to tested. At matched ability the internal relation costs 95.2% of $P(P-1)$, so a benchmark's own labelled dimensions do not license transfer between themselves any more cheaply than mmlu subjects do. P-T2-I2 holds at 85.4% against its registered floor of 80%.

The corrected reference row makes this the paper's least favourable measurement, not its most. §B's P-T2-N9 replaces the zero-error equivalence figure of 10.6% with the two-part code that pays for its own errors, 471.9 bits or 85.5%, at a best-partition error of 0.2138. An equivalence prior over these 24 dimensions buys 14.5% of $P(P-1)$ against the Bell reference, where over mmlu's 57 subjects it buys 0.2% against the same reference. That is a real difference and it runs against the paper: the internal relation has structure a label can find, an order of magnitude more of it than the external one, and the sentence above about "no more cheaply" is true of the observed cost and false of the structural gain. What survives is the order rather than the constant: 14.5% is a constant-factor saving on a quadratic count, which is Def. ??, and not the collapse to $\Theta(P)$ that Prop. ?? needs, and it is that collapse, not a constant factor, that would put a programme on (??)'s cheap row.

P-T2-I3 failed, it was registered to catch a defect in §5, and it caught one. The criterion asked whether two ways of controlling for ability agree. They do not: matching within the benchmark gives 0.9523 and matching outside it 0.8541, a gap of 0.0982 against a widest bootstrap width of 0.0310. Recomputing §5 with ability measured outside mmlu gives **40.2**% on the 659 models that carry all five outside tasks, a drop of 53.7 points. The registered expectation for that recomputation was "near 84%" and is refuted.
The refutation is informative because the two controls do different things, which we had not thought through when registering. The within-mmlu mean removes the general factor and mechanically forces an advantage on A to be paid for elsewhere. The outside tasks impose no such dependence and do not remove the factor: models matched to two points on them still agree on 0.9120 of subject pairs, against 0.9648 with no matching at all. It gives **93.2**%, 0.7 points from the published figure, and the same control on bbh gives 95.2% against its within-benchmark 95.2%. The mechanical dependence is real and worth half a point, the out-of-benchmark control is simply not a control for a within-benchmark general factor, and §5's measurement survives a test designed to break it. P-T2-I6's registered floor of 80% holds.

Two defects of ours in this arm, both found by running it. (i) P-T2-I1 was written one-sided, requiring the observation to fall below the null. That silently assumed the null sits above the observation, which holds on mmlu and fails here: under out-of-benchmark matching the null is at 0.3544 and the observation at 0.8541, and the criterion as written returned "no structure detected" for a separation of more than fifty interval widths.

D The four nulls, in full

§B reports five checks on the structural term and states their outcome. The nulls themselves are here: what each destroys, what it leaves alone, and how far the observation sits from the value the null produces. A null is worth reporting only if it could have come out the other way; Figure 2, drawn against the break-even of Prop. 4 so a reader can see which do, is in the main text (§5) for the two nulls that matter most.

Table 6. Four nulls, 200 draws each, every one pushed through the identical pipeline including the matching filter, and corrected for attenuation on both sides of the comparison. The mean agreement is not evidence and the per-pair entropy is: an additive null with no transfer structure at all reproduces the observed mean agreement to the fourth decimal. The two H-perm rows fit the additive model to the real accuracies, permute the residuals within each model across subjects and add the fit back, so both main effects and the realised residual magnitude survive and only the model-by-dimension association dies; they need no noise model, which the binomial rows do. Separation is quoted in null standard deviations. transfer_noise.py, noise_nulls.py. The whole tabular is generated from results/noise/NULLS.json, and the summary line now carries all four separations: it quoted three, omitting H-flat's, the smallest of them, so nothing was overstated, but a reader counting null rows finds four.

	raw		de-attenuated	
	agr	mean H	agr	mean H
observed, matched at 0.02	0.5143	**0.9392**	0.5198	**0.7712**
H-flat null	0.5105	0.9897 [0.9889, 0.9905]	0.5194	0.9606 [0.9576, 0.9635]
H-additive null	**0.5142**	0.9892 [0.9885, 0.9898]	0.5263	0.9591 [0.9564, 0.9616]
H-perm, in model	0.4958	0.9957 [0.9954, 0.9960]	0.4925	0.9894 [0.9878, 0.9904]
H-perm, in tertile	0.4974	0.9955 [0.9952, 0.9959]	0.4951	0.9891 [0.9879, 0.9902]
de-attenuated separation, in null SD: 133.6, 140.1, 337.6, 399.7 over all four nulls				

Figure 2 (the two nulls that matter most) is given in the main text (§5) and is not repeated here.

E One relation, priced two ways

The two codings of §5 are internally consistent and disagree about what the relation is, which is the point of pricing it twice rather than choosing (Figure 4). The table is the arithmetic; the figure is the comparison the arithmetic exists to make.

Table 7. One relation, priced two ways, each internally consistent. Panel A codes the $P(P-1)$ binary values of the relation: one observation per parameter, so a per-pair parameter is unpayable there and the saturated code is inadmissible rather than expensive. Panel B codes all n_{pair} model-pair comparisons behind each dimension pair, where the parameter is payable and the question is whether it pays. Panel B is the adjudication and Panel A is why the share was never it. One convention governs every row: θ is the median, a rule and not a search, so no selection charge is owed; every admissible code fits exactly one real parameter, worth $\frac{1}{2}\log_2 3192 = 5.82$ bits, identical across the three and therefore omitted from the levels; and the gain is the difference of two rows the table prints. **4.24**% is the canonical value and the one every section below prints. §B's 19-quantile sweep prints 4.2333% for the same gain and both are exact: a gain is a fraction of the reference code, so the sweep's $\log_2 19 = 4.2479$ bits for choosing θ, charged on both sides, cancel in the numerator and survive in the denominator, 135.307/3196.248 against Panel A's 135.307/3192.000. The median rule owes no such charge, so Panel A's is the value to quote; structure_math.py asserts both against the scripts that print them, together with the two readings a reader may reach for and neither section uses. Three rows of this programme's defect ledger were opened by this very table (free parameters on one side only, an n=952 penalty against an n=1 likelihood returning 588.7%, and two base rates in one column), and they are rows 6–8 of the ten kept in criterion_defects.py and printed in the companion (Anonymous, 2026a). refine_arms.py, P-T2A-M1; noise_nulls.py, decompose.py.

Panel A. One outcome per pair, 3192 of them at the median $\theta = 0.5154$

code for the matched relation	bits	of $P(P-1)$	what it pays for
observed, saturated	—	inadmissible	one free parameter per observation: not a code, and the 93.9% share-of-maximum-entropy statistic is its residual entropy
base rate only	3192.0	100.0%	$P(P-1)H(0.500)$: one number, how often transfer happens
equivalence relation, two-part	3186.1	99.8%	$\log_2 \text{Bell}(57) = 186.4$ for the partition + $P(P-1)H(e^\star)$, $e^\star = 0.3565$; buys **0.19**%
best structured code found	**3056.7**	95.8%	$P \log_2 N = 57$ bits for a label per dimension + the same residual; buys +**4.24**%

Panel B. All $n_{\text{pair}} = 952$ comparisons per pair: a parameter costs 4.95 bits and pays for itself iff $\mathbf{D} > \mathbf{0.52}$%

what is coded	D	buys	payoff against 4.95 bits
observed, matched, raw	**6.02**%	57.3 bits	**11.58**×: mdl keeps them
observed, matched, de-attenuated	22.77%	216.7 bits	43.81×
H-additive null, raw	1.02%	9.7 bits	1.97×
H-additive null, de-attenuated	3.89%	37.0 bits	7.48×
H-perm null, raw	**0.42**%	4.0 bits	**0.82**×: mdl rejects them
H-perm null, de-attenuated	1.04%	9.9 bits	2.00×

F The forward ladder, rung by rung

§6.1 runs the refinement limit forwards on a pinned item budget and reports the exponents (Figure 5). The rungs are here, and so is the single-axis form of the trade the exponents describe.

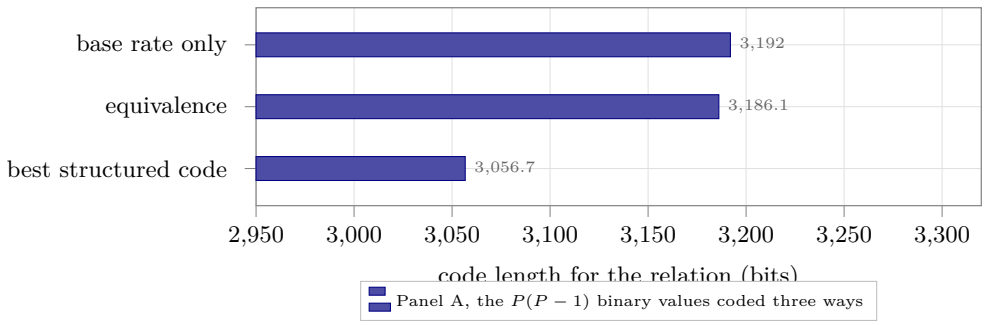


Figure 4. What each code buys, in the currency the proposition is stated in. Coding the relation as an equivalence buys 0.19% of the base-rate code's length, essentially nothing, while the best structured code found buys 4.24%. A relation that a partition explained would show the opposite ordering, and the flatness of the middle bar is the measurement behind §5's claim that the transfer relation this suite induces is not one.

Table 8. The limit run forwards on a pinned item budget: refinement multiplies the pairs and divides the evidence. 250 models, the same subsample at every m, the $m = 1$ matched-ability control held throughout. The partition search uses the identical candidate budget at every m and always evaluates the parent-subject grouping, so a fall cannot be a search artefact. The last column is the arm's answer, and it is a power law rather than a ratio: on the between-parent pairs, whose truth is the $m = 1$ truth by construction, $\Lambda \propto m^{-1.223}$ against an analytic no-structure baseline of m^{-1}, derived in the claimbox below. The two se columns differ and the difference is predicted, not excused: se is the binomial error at each part's own realised accuracy, $p(1-p)$ is concave, and the parts' accuracies spread as m grows, so mean $p(1-p)$ falls from 0.2028 to 0.1941 and se grows as $\sqrt{m\,\overline{p(1-p)}_m/\overline{p(1-p)}_1}$, which predicts 1.00, 1.41, 1.98, 2.77 against the measured 1.00, 1.41, 1.97, 2.73, a match to 0.04 where $\sqrt{m}$ (1.00, 1.41, 2.00, 2.83) misses by 0.10. refine_arms.py, P-T2A-M4. The whole tabular is generated from results/refine/REFINE.json.

m	P	$\overline{\text{agr}}$	D (%)	$0.5 - e^\star$	se growth	Λ, between-parent (%)
1	57	0.5112	8.08	0.1479	1.00×	**6.40**
2	114	0.4874	3.81	0.0956	1.41×	**2.66**
4	228	0.4615	2.85	0.0619	1.97×	**1.10**
8	456	0.4371	2.59	0.0428	2.73×	**0.51**

fitted exponents: $0.5 - e^\star \propto m^{-0.599}$ against the noise-only $m^{-0.5}$, and $\Lambda \propto m^{-1.223}$ against m^{-1}; twice the margin exponent is -1.199 against Λ's -1.223, the quadratic relation to 0.02

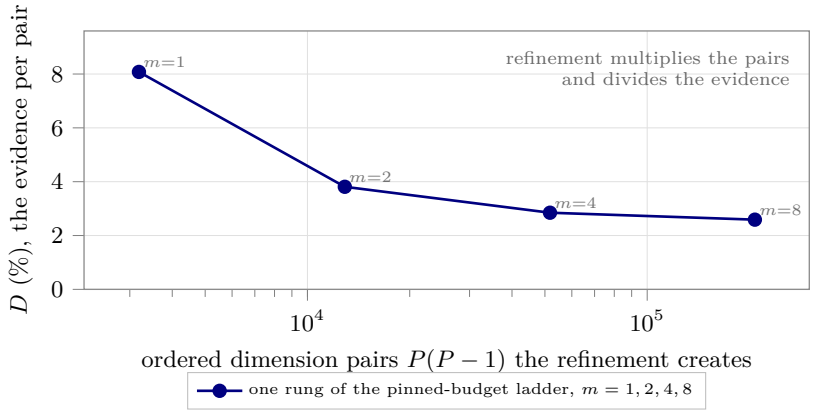


Figure 5. The limit run forwards, on one axis. Splitting each dimension m ways on a pinned item budget multiplies the ordered pairs by about m^2 and drives the structural term per pair down with them: a 65× increase in pairs buys a 92% fall in D. The item budget is identical at every point; nothing is being measured worse, there is simply less evidence behind each of many more quantities. This is the same inequality as Figure 3(a) with the truth held fixed instead of read off a published taxonomy, and the two agree.

G Notation, including the letters a companion reuses

Every symbol this paper uses, with the place it is defined. The second half of the section exists because this family reuses letters across papers and no paper said so: a reader holding two of them at once has had no way to discover it except by being confused first. The collisions below are computed from this table rather than asserted, so a symbol added to any paper of the family cannot silently collide with another.

Table 9. The symbols of this paper.

symbol	what it denotes	defined
P	the number of labelled dimensions a suite reports. $K = P(P-1)$ is the number of ordered transfer claims reporting them side by side makes	§1
$\mathrm{agr}(A, B)$, θ, T	the agreement statistic on an ordered pair, the threshold that binarises it, and the resulting transfer fact	Protocol ??
$H(\cdot)$	binary entropy, in bits. In Law 1 the entropy of the hypothesis class, which is also a test count, because one test yields at most one bit	§3
D	pair heterogeneity: the Jensen gap $H(\overline{\mathrm{agr}}) - \overline{H(\mathrm{agr})}$, in bits. It decides whether a per-pair parameter pays for itself, and it is not what a dimension table claims	Prop. 3
$e^\star$	the error rate of the best partition a restarted search finds, on a relation over K ordered pairs, so $e^\star \in \{k/K\}$	Prop. 5
$\Lambda = 1 - H(e^\star)$	label-predictability: what a dimension table claims. Λ_{in} scores the partition on the relation it was searched on, Λ_{out} on a held-out half	§5
N, n_i, $M(N)$	the induced partition's class count, class sizes, and the within-class ordered pairs it predicts will transfer	Def. ??
q_Λ, α	Λ's own lattice quantum, and the fraction of the reported value one step must stay within for the value to be decidable	Prop. 5
n_{pair}, $D^\star(n)$	comparisons behind one pair, and the mdl threshold family $\log_2 n/2n$, a family and not a constant, which is why a single quoted threshold belongs to a single corpus	Prop. 4
m	the number of disjoint parts each subject's items are split into, when the refinement limit is run forwards with the truth held fixed by construction	§6.1

The same letter, a different object. Each line below names a symbol of this paper that a companion uses for something else. None of these is a notational slip to be repaired by renaming: the objects are each named after their own literature's convention, and renaming one to suit another paper would cost more than it saves. What was missing is the list.

- D, pair heterogeneity here; in Anonymous (2026d) it is an item's distractor set.
- Λ, label-predictability here; in Anonymous (2026d) it is a lever: a family of maps onto records of length L.
- α, a decidability level here; in Anonymous (2026d) it is the exponent in $k^* \propto L^\alpha$.
- θ, an agreement threshold here; in Anonymous (2026e) it is the statistic a criterion is written on.
- m, a split count here; in Anonymous (2026c) it is the cardinality an answer depends on; in Anonymous (2026e) it is the number of executions averaged.
- q, a lattice quantum here; in Anonymous (2026e) it is the lattice spacing of a statistic.

What is deliberately shared. k is a per-query access budget in every paper of the family that has one; V is verbatim containment in all four papers that count it; ι is the credit rate in all three that use it; n is a count of items throughout. Those four are shared because they are the same object, and a paper that imports one imports the definition with it.

H Every criterion this paper names, with its freeze date

A preregistration's claim is a claim about time, so it is the one claim in this paper that must not be transcribed: the commit column below is read from git log --reverse when the paper is compiled, not copied from the file's own front matter. Amendments are counted and never overwrite. Naming is not grading: a criterion this paper cites for someone else's verdict appears here beside one it grades itself, and the claim ledger (claims_all.tex) is what separates the two. The five criteria carrying the prefix P-T2A-* are this paper's own; the P-T2-* rows keep the prefix of the preregistration that froze them, because renaming an identifier breaks the link between a graded criterion and its own frozen text, and the paper says so in one sentence rather than re-labelling the record.

Table 10. The preregistration record for every arm this paper names, generated from git at build time.

experiment	frozen at	criteria	amend.	words
E_DECOMP_DISCLOSURE	5991954f, 2026-08-17 09:55	P-T2-A2, P-T2-D4	0	913
E_INTERNAL_ADDENDUM2	e9a8316a, 2026-08-17 07:50	P-T2-I6	0	480
E_INTERNAL	335c4d3f, 2026-08-17 07:32	P-T2-I1, P-T2-I2, P-T2-I3, P-T2-I4	0	604
E_NOISE_ADDENDUM	82943e59, 2026-08-17 06:33	P-T2-N6	0	412
E_NOISE_ADDENDUM3	94bfd50a, 2026-08-17 09:06	P-T2-N7, P-T2-N8, P-T2-N9, P-T2-N10, P-T2-N11	0	964
E_NOISE	798705cc, 2026-08-17 06:06	P-T2-N1, P-T2-N2, P-T2-N3, P-T2-N4	0	714
E_RECOMPUTE_DISCLOSURE	0c03c4b1, 2026-08-17 03:37	P-T2-D1, P-T2-D2, P-T2-D3	0	584
E_STRUCTURE	d0817fdf, 2026-08-17 15:22	P-T2A-M1, P-T2A-M2, P-T2A-M3, P-T2A-M4, P-T2A-M5, P-T2A-M6	0	2272
8 arms		26 criteria	0	6,943

I The derived matrices

Every number in this paper is a function of two public leaderboard snapshots, and a reader who has to rebuild those snapshots before checking anything is a reader who will not check anything. The three matrices below are the whole input: model × dimension accuracies and their standard errors, plus the per-dimension item counts the lattice quantum $1/n$ is computed from. They are what the released bundle ships, and the hashes are recomputed here at build time rather than copied from SHA256SUMS, so a matrix that changed after the bundle was sealed changes this table too.

Table 11. The derived matrices, their shapes and their hashes, read from the files at build time. outside marks the held-out benchmarks a corpus's own dimensions are scored against; n is the per-dimension item count.

file	models	dims	arrays	sha256 (16)
bbh_v2.npz	895	24	acc, n, names, outside, se, tasks	2413abcb818c7d15
mmlu_v1.npz	690	57	acc, n, names, se, subjects	d0d14dacc47a2c09
mmlu_v1_with_outside.npz	659	57	acc, names, outside, se, subjects	7b720cc56b5b7fbb

J Every corpus's held-out repetitions and cluster bootstrap, in full

Table 3 reports one Λ_{out} per corpus. Each of those numbers is the mean of 20 held-out repetitions, and each row also carries a 60-draw bootstrap clustered on declared base family. The spread is not decoration here: it carries two facts the point estimates cannot, and both of them cut the same way the body's own reading does.

First, at small P the repetitions stop varying. 3 of 8 corpora return a standard deviation of zero to machine precision over 20 repetitions (musr at $P = 3$, open llm v2 suite at $P = 6$, open llm v1 suite at $P = 6$), because at those dimension counts the half-split has too few admissible configurations for the repetitions to land anywhere else. A zero standard deviation read as a precise measurement is the error this paper is about; it is the resampling side of the same coarse lattice that Prop. 5 prices, and it is reported here rather than in the body because it changes no verdict, both suite rows being already excluded for nesting and musr already failing the lattice test.

Second, the two leaderboard suites return the identical value. They are different model populations over different benchmarks and they agree to four decimals, which is what a quantity pinned by its lattice rather than by its data looks like. Neither row is a unit of the exact test, for the reason the body gives, and this is the independent symptom of it.

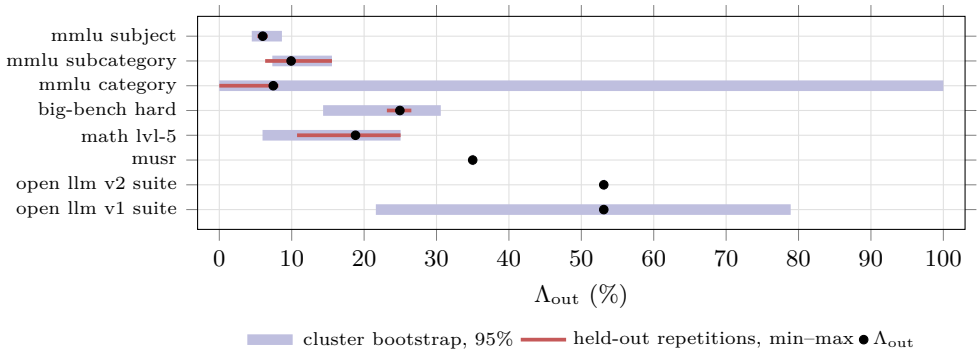


Figure 6. Which uncertainty is the big one, per corpus. The narrow bar is the range of the 20 held-out repetitions and the wide one is the 60-draw bootstrap clustered on declared base family; the dot is the number Table 3 prints. The clustering is what the uncertainty is made of: on big-bench hard the bootstrap is 4.8× the repetition range, and on mmlu category it is the whole admissible interval $[0, 100]$, the same rung the lattice test already calls not decidable, failing a second time by a wholly independent route. The 3 rows that carry no bar at all are the ones whose repetitions do not vary; that is the degenerate case the paragraph above reports. Every coordinate is read from results/refine/CORPORA.json at build time.

Table 12. Every corpus, its held-out repetitions and its cluster bootstrap. Λ_{out} is the held-out mean; sd, min and max are over the repetitions; first is the first split alone, which is what a single held-out run would have reported. The bootstrap columns are clustered on declared base family. Generated from results/refine/CORPORA.json.

corpus	P	Λ_{out}	sd	min	max	first	Λ 95% CI	clus.
mmlu subject	57	0.0600	0.0042	0.0525	0.0668	0.0543	[0.0448, 0.0865]	426
mmlu subcategory	17	0.0992	0.0232	0.0633	0.1556	0.0918	[0.0732, 0.1556]	426
mmlu category	4	0.0745	0.0223	0.0000	0.0817	0.0817	[0.0000, 1.0000]	426
big-bench hard	24	0.2494	0.0076	0.2315	0.2652	0.2446	[0.1434, 0.3059]	492
math lvl-5	7	0.1881	0.0459	0.1074	0.2504	0.1704	[0.0597, 0.2504]	495
musr	3	0.3500	0.0000	0.3500	0.3500	0.3500	[0.3500, 0.3500]	495
open llm v2 suite	6	0.5310	0.0000	0.5310	0.5310	0.5310	[0.5310, 0.5310]	492
open llm v1 suite	6	0.5310	0.0000	0.5310	0.5310	0.5310	[0.2162, 0.7892]	405

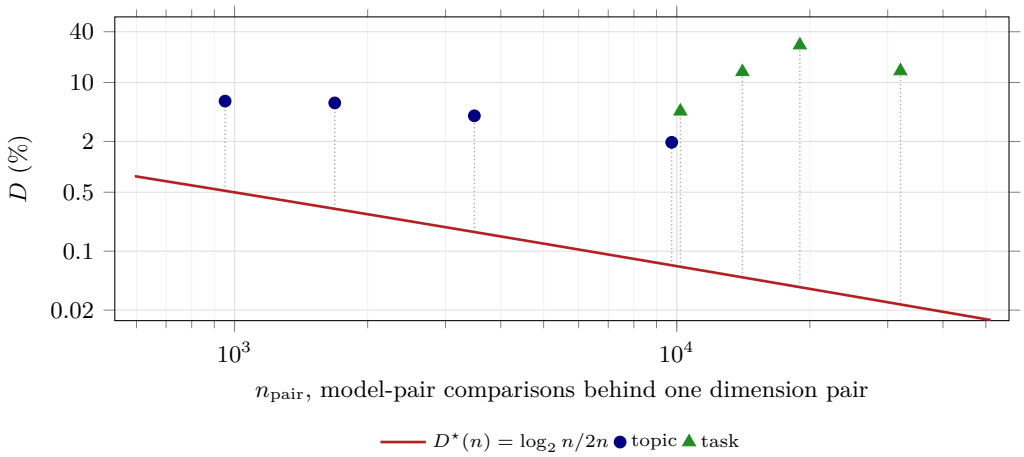


Figure 7. All eight corpora on the break-even curve, where the payoff is a distance. Prop. 4 prices a per-pair parameter at $\log_2 n/2n$, so a corpus pays for one exactly when it sits above the line, and the dotted drop is the factor Table 3's mdl column reports. Every one of the eight is above it, from 11.6× on mmlu subject to 741× on musr. The spread is not mostly a fact about D: D itself spans 14× across these corpora while the threshold under them falls 22×, so a corpus with more comparisons behind each pair buys a cheaper parameter as much as a stronger relation. Recomputed from results/refine/CORPORA.json.

K Every null, drawn twice, with the observation against each

Four nulls, 200 draws each, every one pushed through the identical pipeline over 690 models and 57 dimensions. Each is drawn twice: once on the raw agreement and once de-attenuated for the measurement error in each cell. The second pass is the one that could have ended the paper, since de-attenuation moves the observation towards the nulls if the apparent structure is measurement error and away from them if it is not, so reporting only the pass that came out well would be the sixth failure type §5 names, committed inside the table that diagnoses it. Observed mean H is 0.9392 raw and 0.7712 de-attenuated; the null means are all above both.

Table 13. The four nulls. outside is whether the observation falls outside the null's 95% interval, and sep. is how many null standard deviations separate the two on the de-attenuated pass. Generated from results/noise/NULLS.json.

null	pass	mean H	95% CI	outside	sep. (sd)
H-flat	raw	0.9897	[0.9889, 0.9905]	yes	—
	de-attenuated	0.9606	[0.9576, 0.9635]	yes	133.6
H-additive	raw	0.9892	[0.9885, 0.9898]	yes	—
	de-attenuated	0.9591	[0.9564, 0.9616]	yes	140.1
H-perm (residuals within model)	raw	0.9957	[0.9954, 0.9960]	yes	—
	de-attenuated	0.9894	[0.9878, 0.9904]	yes	337.6
H-perm (within question-count tertile)	raw	0.9955	[0.9952, 0.9959]	yes	—
	de-attenuated	0.9891	[0.9879, 0.9902]	yes	399.7

The dependence null, which is the one that widened the interval. Clustering the bootstrap on declared base family rather than on the model collapses 690 models to 426 clusters and widens the interval by 1.51×, from [0.9155, 0.9321] to [0.9073, 0.9324]. The widened interval is the one this paper reports, and it still excludes the equivalence share of 0.0584, which is the only reason the widening is reported as a robustness result rather than as a correction. A recovery check accompanies both passes: recovering a planted structure of zero returns 0.000 raw and 0.000 de-attenuated, and the reported standard error is 1.0029 times the binomial one, so neither pass is inflating what it measures.

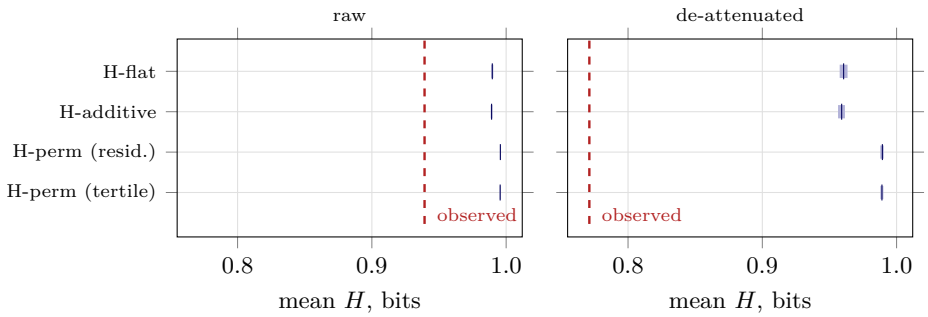


Figure 8. The pass that could have ended the paper, beside the one that did not. Each bar is a null's 95% interval on mean H and the dashed line is the observation the same pipeline returns; both panels share one axis, so the two passes can be compared and not merely reported. De-attenuation moves the observation away from the nulls, not towards them, which is the direction that only holds if the structure is not measurement error: the separation runs from 134 to 400 null standard deviations on the de-attenuated pass. Nulls are ordered by that separation. Recomputed from results/noise/NULLS.json.

L What is on disk, with its two preconditions

Every arm this paper reports, counted from results/ at build time. Two of the scripts refuse to report at all unless they first reproduce an earlier script's stored output: transfer_noise.py will not print until it reproduces the stored unrestricted agreement to four decimals, and noise_nulls.py will not print until it reproduces both of transfer_noise.py's matched arrays. A stored array whose producing script is absent is a number no reader can regenerate, so the precondition is what makes that class of failure loud rather than silent.

Table 14. The experimental record, counted from the artifacts at build time.

arm	files	records	arm	files	records
anchor	1	9	granularity	1	9
bits	3	156	internal	1	15
bits2	1	9	loose	0	0
covar	1	11	noise	2	39
decomp	1	13	recompute	3	17
defects	1	32	refine	3	47
exposed	1	6	reflex	1	9
external	1	9	scaling	1	9
grain	1	7	transfer	3	16
total	26	413			

M The scorecard, every graded row

This paper's claims are graded here rather than in its sections, so that the sections can be about what was found and this appendix can be complete. Criteria were frozen in prereg/ before the corresponding data existed, and one generator writes all three tables from one source, so the index below cannot disagree with the rows it counts.

Table 15. The scorecard in eight lines, counted from the scorecard. Every graded row of Table 17 and Table 18 falls in exactly one class below. Where a family's status names two classes at once, the row is counted in the more serious one, which is the only direction of error a summary of one's own scorecard may have. It is an index, not a substitute, and the rows are all still there: 24 claim families in Table 17, 26 sub-rows in Table 18. What the unflattering rows are for. 23 of the 49 rows carry a positive grade and 23 an unflattering one, and the second number is not a defect count: each of those rows narrows a claim made above it, names a number of ours that moved, or marks a question this apparatus cannot decide. None of them is an experiment left undone: work nobody has run is not graded here at all, and is bounded where it is stated.

grade	rows	what the grade means in this paper
withdrawn / superseded	0	asserted, then taken back or re-assigned; the sequence is in the post-mortems
not supported / failed	2	a frozen criterion was missed, on our own data
self-correction	14	our own arithmetic or our own criterion was wrong and is repaired here
not decidable / not evaluable	0	the apparatus or the band cannot separate the outcomes: not a null
acknowledged limit	7	a boundary of the result, stated where the result is
formal	5	proved from the definitions; no measurement can move it
measured	11	an artifact on disk, and a reader may dispute the artifact
argued / reported	7	no grade is claimed: an argument, or a number printed without a criterion
other	3	status words outside the eight classes above
total graded rows	49	every numbered row of both tables

Table 17. Every claim, its grade, and its evidence. formal means arithmetic on the definitions, checkable without data; measured means it comes from public model results a reader can recompute; self-correction rows are errors this paper found in its own earlier draft or that referees found in it. This paper's positive claim is the one weakest graded row, stated here rather than discovered by a reader. Rows are numbered by position in this table and sub-rows inherit their family's number.

Claim	Status	Evidence
1. One test yields one bit, so identification costs the description length and a prior buys the bits it removes	formal	Prop. 2
2. Transfer is a property of a (pair, quantity), so a comparison is well posed only once the quantity is fixed on both sides	formal	Prop. 1
3. A rule owing no tests differs from the null on no pair: cost is content, and no rule gets a discount	formal	Prop. 3 of Anonymous (2026a)
4. Identification is an obligation exactly where no text restricts scope; only a restrictive declaration reduces it	formal	Prop. ??, conditional on (A1)
5. On 690 public models over 57 subjects the structural term is 6.02% of $P(P-1)$, and 22.77% de-attenuated	measured	§5; Table 1, the positive result
6. The share cannot adjudicate a branch: on the same footing the saturated code costs 588.7% against 99.8%	self-correction	Table 7: the framework claim, not the measurement
7. The refinement limit is a fall with a measured exponent: at a pinned item budget $\Lambda \propto m^{-1.223}$ against an analytic no-structure baseline of m^{-1}, from 6.40% to 0.51% as m goes $1 \to 8$	measured	Tab. 8, M4
8. Λ is a corpus property, not a function of P: out of sample, over twenty half-splits, it runs 6.00% to 53.10% over eight public corpora, so P-T2-D4's registered monotonicity fails under both readings	refuted; below: 1 not supported	Tab. 3
9. Inside one benchmark the relation costs 95.2% of $P(P-1)$ over 552 ordered pairs of 24 dimensions	measured	§C
10. De-attenuation rescales an observed deviation and cannot manufacture resolution: at $f = 0.257$ it maps 0.5142 to 0.529 against 0.5263	measured	§B
11. The binomial nulls treat item noise as independent across systems, which inflates them in our favour; the sign is disclosed and is unmeasurable from aggregates	limit	§8(iii)
12. $q = 1$ by measurement format, so every count here is a floor	limit	§8(v)
13. D and Λ are two measurements this paper called by one adjective, as a referee noticed: they coincide at P=57 by coincidence and rank at Spearman +0.563 across eight corpora	self-correction	§6
14. Λ out of sample is below Λ in sample at every rung but only by 0.02 to 0.72 points over twenty splits, so the search was not overfitting, and our registered reason (that the gap grows with P) is refuted	measured	Tab. 2, M3

Table 17, continued.

Claim	Status	Evidence
15. Six things a benchmark author can do without running a transfer test, one a single command on the submitted artifact and one a dimension count computable before any data exists	argued from the laws	Tab. 4; Prop. 5
16. Panel A's own gain did not equal the difference of its own rows: a base-rate code priced at the observed mean agreement was set beside a structured code priced at the thresholded relation's, with a $\log_2 19$ charge on one side only	self-correction	Tab. 7A
17. Row 9 of the programme's footing ledger, and the one of the nine we found ourselves: mmlu's +4.24% was quoted beside bbh's 14.5%, one against a vocabulary reference and one against a Bell reference	self-correction	§5
18. Λ's lattice resolution in closed form, and the dimension count it implies: $q_\Lambda \simeq \frac{4}{\ln 2}(0.5 - e^\star)/P(P-1)$, so reporting Λ at tolerance α needs $P(P-1) \geq 2/(\alpha\sqrt{\Lambda \ln 2/2})$; it reproduces all three of this paper's own decidability verdicts	formal, and the only result here usable before data	Prop. 5; Fig. 3
20. The low-rank line's newest member ships a redundancy tool, and it answers a different question from Prop. 4	argued	factor count is set by a variance criterion; MDL asks whether a named parameter earns $\frac{1}{2}\log_2 n$ bits (Maimon et al., 2026)
21. An interior granularity optimum is measured from the representation side, on twelve models	reported	and it has low agreement with the human taxonomy, which is the premise §5 prices (Li et al., 2026)
22. This paper is a measurement and its companion is a position and an accounting	acknowledged limit	no number in §5 or §6 depends on a test the companion counted (Anonymous, 2026a)
19. Fifteen published numbers of ours have moved, twelve found by referees and three by our own audit, five of them in this paper's favour	self-correction	the fifteen are rows of this scorecard, graded self-correction and counted in Table 15
83. This paper's two measurement tables are generated from their artifacts	code-verifiable	make_t2a_tables.py emits both from refine/REFINE.json and noise/NULLS.json; the residual of hand-typed three-decimal values falls from 79 to 64

One row per claim family. Each status cell names the grades of that family's sub-rows, so a failure is counted here even though its row is elsewhere: all 26 sub-rows are in Table 18 in full, of which 9 self-correction, 6 acknowledged limit, 1 not supported. Nothing is dropped, and nothing is summarised into a kinder word than it had.

Table 18. Supporting refinements, moved from Table 17 by the stated rule. The direction is deliberate: a selection rule for a scorecard should be biased against its authors, so unflattering rows stay in the main table and supporting detail moves here. claims_all.tex is the source both tables are generated from.

Claim	Status	Evidence
The 26 rows below were moved here from Table 17 by the rule in papers/split_claims.py: a lettered sub-row whose status is supportive moves; every top-level claim and every row carrying a withdrawal, failure or self-correction stays. Nothing is dropped.		
2a. …and this paper broke it twice: a zero-error code against an error-bearing one, and a sum against one of its parts	self-correction	§5, §B; both found by referees
3a. …and it appears in the data unlooked for: the cheapest code at 28.8% is the null rule wearing a threshold	measured	§B
4a. …and (A1) is about the granularity of scope statements, not their existence: measurable, unmeasured	acknowledged limit	§8(0)
5a. …against 3.89% for a no-interaction null and 1.04% for a residual permutation, i.e. 5.86× and 21.86× on D, at 138.3 and 346.5 null SD on D and not on mean H	measured	Table 6
5b. …and the 4.9× effect of controlling ability is 93.4% base rate, not structure	self-correction	§5: the warning is kept, the description-length claim withdrawn
5c. …and the general factor itself is not our finding	reported	(Ilić & Gignac, 2024; Burnell et al., 2023a; Ruan et al., 2024)
6a. …because $\sum H(\text{agr})$ is a code length conditional on $P(P-1)$ free per-pair parameters	acknowledged limit	§B, P-T2-N11
6b. …and the 588.7% that replaced the share paired an n=952 penalty with an n=1 likelihood, which a referee found: withdrawn	self-correction	Tab. 7, M1
6c. The statistic that does adjudicate is in bits: a per-pair parameter costs 4.95 and needs $D > 0.52\%$; the observation pays 11.58×, the residual-permutation null 0.82×, and all eight corpora pay	measured	Tab. 7B
7a. …and backwards on mmlu's own taxonomy it falls too, 9.98% → 6.02% from 17 dimensions to 57	measured	Tab. 2, M2
7b. …so refinement multiplies the claims and divides the evidence; the quadratic surface is not escaped by relabelling	argued from the boundary	§2
7c. The "flat" boundary this paper published is withdrawn, a referee having asked: the P=4 rung's one lattice step moves Λ by 10.70 points against a band of 3.96, so it is not decidable by this programme's own rule	self-correction	§6, M2
7d. …and the sweep behind it first carried a pure-P confound: the naming cost falls 32.56% → 5.84%	self-correction	§B
7e. …while the forward arm's own ground-truth argument was void: split siblings are negatively correlated at $-1/(m-1)$, so its answer is read off the 98.6% of pairs that are not siblings	self-correction	§6.1
8a. …so the maximum-entropy reading is about mmlu; the second corpus replicates the share only	acknowledged limit	§8(i)
8b. The registered hypothesis about what sets it, task labels over topic labels, is ordered on every corpus but worth $p = 0.167$ once units are counted rather than rows, and it fails its interval test even on the one P-matched pair	not supported	§6.2, M5
8c. …and four of the eight corpora cannot answer it at all: one lattice step moves Λ by 4.71 to 26.83 points there	acknowledged limit	Tab. 3
9a. …and a registered check caught a defect in §5's control: a gap of 0.0982 against 0.0310	falsified	§C, P-T2-I3

Table 18, continued.

Claim	Status	Evidence
9b. …repaired by leave-two-out matching: 93.2% against the published 93.7%	measured	§C, P-T2-I6
9c. …and one criterion was written one-sided, returning "no structure" for a fifty-width gap	self-correction	§C warnbox; P-T2-I1 regraded two-sided
10a. …so P-T2-N7's expectation was refuted while the criterion passed, and the repair helped us	self-correction	§B
14a. …and the single split an earlier version published misreported that gap by a factor of four	acknowledged limit	§8(i″)
20a. …and their verdict is read at a fixed task count, where ours pins the item budget and refines	argued	"redundant" and "cannot be shown to pay" can both hold with all structure intact
21a. …and it is corroboration only in the weak sense the difference allows	acknowledged limit	their objective carries no item budget, so an optimum in it does not imply Λ's decay in ours
21b. …and which capabilities win depends on the readout	reported	difference-in-means against logistic regression; a geometry is a (corpus, readout) property (Anonymous, 2026f, row 205a)
83a. …and the noise summary had quoted three separations for four null rows	self-correction	H-flat's 133.6 null SD was omitted; it is the smallest, so nothing was overstated, and all four are printed now