A PERMUTATION AUDIT FOR CELL-LEVEL FOUNDATION-MODEL TRANSFERABILITY CLAIMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transferability scores such as LogME are sometimes read as *cell-level* warnings: a low score for a (foundation model, task) cell is taken to predict that this particular cell will go wrong. The usual evidence is a correlation between the score and cell-level error across a grid of models and tasks. Such a correlation can be produced without any cell-level information, by shared task difficulty, by shared model strength, or by the noise in the error estimate itself. We give an audit for such claims: gates that check that the score is defined on the cells and that the outcome has cell-level variance not dominated by seed noise, then three within-task permutation nulls, each aimed at one artefact, all of which must reject in the claimed direction. The noise null adjusts for each cell's measured seed variation instead of dividing by it; we show that a natural ratio standardisation rejects genuine signal when noise grows with difficulty (pass rate $0.98 \rightarrow 0.00$), whereas the adjusted null does not (1.00). On a 13×10 synthetic grid the audit passes 0.6% of no-signal runs and none of five leakage regimes, and detects genuine within-task correlations of 0.4 in every run. Applied to LogME, the H-score and N-LEEP on seven foundation-model pools, 11 of 84 warning-oriented audits pass, 10 of them on two medical-imaging pools where two of the three scores pass on the same outcomes; the audit also identifies where it cannot decide: outcomes dominated by seed noise, targets without seed replicates, and grids too small to have power. The implementation returns a structured record that can accompany a new claim.

1 INTRODUCTION

A transferability score assigns each pair of a pretrained foundation model (FM) and a target task a number computed from frozen embeddings, without fine-tuning (Nguyen et al., 2020; You et al., 2021; Bao et al., 2019; Li et al., 2021; Shao et al., 2022). Scores are usually evaluated by how well they rank models on one task (Agostinelli et al., 2022; Abou Baker & Handmann, 2024). A stronger reading is also possible: that a score works as a *cell-level* warning, i.e. that a low score for a particular (FM, task) cell predicts that a downstream quantity for that cell — say, the error of a learning-curve extrapolation or of a sample-size estimate — will be large. This paper is about how to check such a claim.

The natural evidence is a correlation between the score and the cell-level error across a grid of models and tasks. It is weak evidence. The error of a cell is partly a property of the task (some tasks are hard for every model), partly of the model (some models are strong everywhere), and only partly of the pair. A score that tracks only task difficulty or only model strength correlates with cell error without saying which *cells* are hard. And because the error is itself estimated from a few training seeds, a score that happens to track how noisy a cell is will correlate with the error through the noise alone. These are instances of the familiar hazards of leakage and uncontrolled comparisons in machine-learning evaluation (Kapoor et al., 2024; Lipton & Steinhardt, 2019; Bouthillier et al., 2021); what is missing for this kind of claim is a concrete test.

Contributions.

1. **An audit** (Sec. 3) consisting of three gates and three within-task permutation nulls — one for task-level leakage, one for model-family confounding, one for the noise channel — combined by

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.

conjunction in a pre-declared direction, with verdicts that distinguish `fail` from the cases where the audit cannot decide (`incomplete`, `noise-dominated`, `underpowered`).

2. **A validity study** (Sec. 4) on a synthetic harness that mimics how the real outcomes are measured. It shows which null stops which artefact, gives power by effect size and grid shape, and shows that the common remedy for the noise channel — dividing the error by its noise level — rejects genuine signal when noise grows with difficulty, which covariate adjustment does not.
3. **An estimability gate** (Sec. 3.1): on cells with no more labelled examples than embedding dimensions, the H-score inverts a rank-deficient covariance and LogME's evidence maximisation degenerates, and implementations return plausible-looking values or none. No null on the score–error relation can detect this.
4. **An application** (Sec. 5) to three published scores on seven FM pools and a hand-crafted control pool. Eleven of 84 warning-oriented audits pass, ten of them on two medical-imaging pools where two of the three scores pass on the same outcomes; on language and time-series pools the within-task correlation is mostly negative, so the warning reading fails, and on two other pools half the audits are noise-dominated.

The audit checks a claim; it does not propose a score, and passing it is necessary rather than sufficient for a cell-level claim to hold (Sec. 8).

2 Setup: what a cell-level claim asserts

Objects. An audit concerns a candidate score $\varphi(f,t)$ and an outcome $y(f,t)$ on an $n_{\mathrm{F}} \times n_{\mathrm{T}}$ grid of FMs f and tasks t; cells may be missing. In our application y is the absolute error $|b(f,t)|$ of a baseline predictor of the cell's learning curve, and each $y(f,t)$ is computed from $n_s = 5$ training seeds, whose spread gives a per-cell noise floor $s(f,t) = \sigma_{\mathrm{seed}}(f,t)/\sqrt{n_s}$. A claim has a direction, which must be declared before auditing. The scores we audit measure "easiness", so the warning reading is that $-\varphi$ is large where y is large.

Decomposition. Write the outcome with the two-way decomposition

$$y(f,t) = \mu + \alpha(f) + \beta(t) + \gamma(f,t), \tag{1}$$

with model effects α, task effects β and interaction γ (Snedecor & Cochran, 1989); with missing cells the effects are fitted by alternating least squares. A *cell-level* claim is a claim about γ, the part of the error specific to the pair. A pooled correlation between φ and y mixes all three components.

Three artefacts. Each of the following produces a positive correlation between φ and y without any information about γ.

1. *Task-level leakage*: φ tracks β, i.e. which tasks are hard. This includes the inflation of a raw correlation by shared task structure.
2. *Model-family confounding*: φ tracks α, i.e. which models are strong. Within a task, stronger models then have both higher φ and lower error.
3. *Noise channel*: φ tracks the noise floor s. Since $\mathbb{E}|b|$ grows with s — by $\sqrt{2/\pi}\,s$ when the true error is zero — a score that is high where the error estimate is noisy correlates with $|b|$ through the noise alone.

Two further concerns are outside what a within-grid test can settle. A correlation reported on a post-hoc selection of tasks cannot be detected from the reported grid; we instead report how consistently the within-task correlation is positive across tasks. Contamination of benchmarks by pretraining data shifts a benchmark as a whole and needs evidence from outside the grid (Sec. 8).

Verdicts. The audit returns `pass`; `fail` with the name of the blocking null; or a verdict saying why it cannot decide: `not-estimable` (too few cells on which the score is defined), `capacity-exceeded` (too little cell-level variance), `noise-dominated` (the seed-noise floor exceeds half the typical outcome), or `incomplete` (the noise floor is not available). A verdict on a grid below 60 cells is flagged `underpowered`.

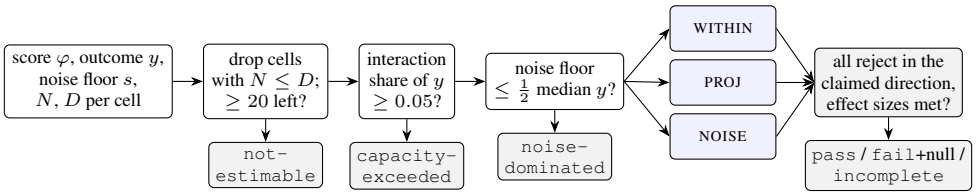


Figure 1: Audit flow: three gates, then three within-task permutation nulls combined by conjunction. NOISE is skipped, and the verdict is `incomplete`, when no noise floor is available. Verdicts on grids below 60 cells are flagged `underpowered`.

3 THE AUDIT

3.1 GATES

Estimability. LogME maximises a Bayesian evidence over two precision parameters using the spectrum of $F^\top F$ for the $N \times D$ embedding matrix F, and the H-score inverts the feature covariance. When $N \leq D$ the covariance the H-score inverts is rank-deficient, and LogME's evidence maximisation degenerates because the labels can be fitted exactly. Implementations still return numbers: on a language task with $N = 45$ labelled examples, the H-score with the usual 10^{-4} ridge is between 0.9772 and 0.9778 for all ten models, against a median of 0.42 on cells with $N > D$, and the standard LogME implementation returns no value for four of them. No null on the relation between score and outcome can detect this, because the score itself is undefined. We therefore remove cells with $N \leq D$ before auditing, for all three scores so that they are audited on the same cells, report how many were removed, and return `not-estimable` if fewer than 20 remain.

Capacity. A cell-level claim needs cell-level variance: we return `capacity-exceeded` if the interaction share $\mathrm{frac}_\gamma = \sum \gamma^2 / \sum (y - \mu)^2$ of (1) is below 0.05. Seed noise also contributes to γ; we report the bound $\sum s^2 / \sum \gamma^2$ on its share alongside.

Noise dominance. The noise null below adjusts linearly for a five-seed estimate of the floor, whose relative sampling error is about 36%; on our medical pools much of the floor's interaction variance is this sampling error, so the adjustment removes only part of a channel that runs through the true noise level. The residual matters only when the channel is strong, i.e. when the floor is comparable to the error. In a sweep where the score tracks the true noise level almost exactly (Table 3), the false-pass rate is at most 0.03 below a floor ratio of 0.4, at most 0.045 between 0.4 and 0.5, and up to 0.18 above 0.8. We return `noise-dominated` if the mean floor exceeds half the median outcome.

3.2 THREE NULLS

All nulls permute y across models *within* each task, $B = 200$ times, and use one-sided p-values $p = (1 + \#\{\text{null} \geq \text{observed}\})/(B + 1)$ (Phipson & Smyth, 2010) in the claimed direction.

WITHIN: within-task association. The statistic is the Pearson correlation of φ and y after subtracting task means from both. Within-task permutation keeps task means fixed, so this test is identical to testing the raw pooled correlation under the same permutations (the pooled covariance is the within-task covariance plus a between-task term, and both that term and the standard deviations are fixed under these permutations); the task effect is removed by the permutation scheme itself. We require $r \geq 0.10$.

PROJ: model-family confounding. Within-task permutation is not a valid null on its own when models differ in overall strength: if φ tracks α, the permuted outcomes lose α while the observed outcome keeps it, and WITHIN rejects without any cell-level signal. This is the known failure of unrestricted permutation in factorial designs with nuisance effects (Anderson & ter Braak, 2003; Winkler et al., 2014). PROJ removes both effects from the statistic rather than from the permuted data: it correlates the interaction components of φ and y, recomputing the interaction of the permuted

Table 1: Rate at which each null rejects (with its effect size) and the audit passes; 13×10 grid, 50 seeds. For R1 with $\rho > 0$ a high pass rate is correct; for every other row a low one is. "n-d": share of runs with verdict `noise-dominated`. Within-task r: mean correlation of φ and y after removing task means. †No-signal rows use 500 seeds.

Regime	within-task r	WITHIN	PROJ	NOISE	n-d	pass
R1, $\rho = 0$†	0.00	0.06	0.01	0.01	0	0.01
R1, $\rho = 0.3$	0.57	1.00	1.00	1.00	0	1.00
R1, $\rho = 0.3$, $\lambda = 0.2$	0.56	1.00	1.00	1.00	0	1.00
R2, task-level leakage†	0.00	0.05	0.01	0.01	0	0.01
R3, model-family confounding	0.43	1.00	**0.00**	0.00	0	0.00
R4a, pure noise channel	0.39	–	–	–	1.00	0.00
R4b, noise channel	0.10	0.32	0.26	**0.02**	0	0.00
R3+R4b	0.33	0.86	0.26	**0.02**	0	0.00

outcome each time. With model effects three times the interaction, its rejection rate at $\alpha = 0.05$ is 0.01–0.05 (Table 5). We require a positive correlation with $r^2 \geq 0.05$.

NOISE: the noise channel. NOISE tests the partial correlation of the interaction components of φ and y given the interaction component of the noise floor s, with the Freedman–Lane scheme (Freedman & Lane, 1983; Winkler et al., 2014): y is regressed on s, and the residuals are permuted within tasks and added back to the fitted values. We require a positive partial correlation with $r^2 \geq 0.05$. The obvious alternative, correlating φ with y/s, is a ratio variable and inherits the spurious-correlation problems of ratios (Kronmal, 1993): when noise grows with difficulty, dividing by s removes genuine signal along with the noise (Table 2).

Decision rule. A claim passes if the three gates pass and all three nulls reject with their effect sizes. Each null tests a condition that a genuine cell-level signal must meet, so all three are required rather than corrected for: this is an intersection–union test (Berger, 1982), whose false-pass probability is at most that of its least permissive null under any dependence between them, provided each null is valid for the artefact it targets — which for NOISE requires the noise-dominance gate. If any null that can be run fails, the verdict is `fail` with that null named; if WITHIN and PROJ pass but no noise floor is available, it is `incomplete`. Alongside the verdict we report, for description only, the share of tasks whose within-task correlation is positive. One audit with $B = 200$ takes under a second on a single CPU.

Output. Each audit returns a structured record naming the verdict, the blocking null if any, the gate statistics and the effect size and p-value of every null that was run (example in Appendix A).

4 VALIDITY ON A SYNTHETIC HARNESS

Harness. The harness mimics how the real outcomes are measured. Each cell has a true error $b^*(f,t)$ and a seed-noise level $\sigma(f,t)$; five seeds are drawn, $y = |\text{their mean}|$, and the audited noise floor is their sample standard deviation over $\sqrt{5}$. With model, task and interaction effects $\alpha, \beta, \gamma \sim \mathcal{N}(0,1)$: *R1* (genuine signal) $b^* = 1.5 + 0.3\alpha + 0.3\beta + \rho\gamma$, $\sigma = 0.6\, e^{\lambda\gamma}$, $\varphi = \gamma + 0.3\varepsilon$, where $\lambda > 0$ makes the noise grow with difficulty; *R2* φ tracks β only; *R3* φ tracks α only; *R4a* pure noise channel ($b^* = 0$, φ tracks σ); *R4b* noise channel at the real-data floor ratio (b^* has its own interaction, σ and φ share an independent factor); *R3+R4b* both. The mean noise floor is 17% of the median outcome in R1, as on the real medical pools (20–25%). Each setting has 50 seeds on a 13×10 grid unless stated; generators are in Appendix B.

Which null stops which artefact. Table 1 shows why all three nulls are needed. Task-level leakage (R2) is removed by within-task permutation itself. Model-family confounding (R3) passes WITHIN in every run — the unrealistic-null problem of Sec. 3.2 — and only PROJ stops it. The noise channel (R4b, R3+R4b) passes WITHIN and PROJ in about a quarter of runs and only NOISE stops it. When

Table 2: Genuine signal and the noise null. Left: rate of passing the noise test for the covariate-adjusted NOISE and for the ratio alternative (correlating φ with y/s), R1 with $\rho = 0.3$ as the noise grows with difficulty (λ), and on the two noise-channel regimes. Right: audit pass rate under R1 by signal strength, with the mean within-task correlation r.

	ratio y/s	NOISE
R1, $\lambda = 0$	0.98	1.00
R1, $\lambda = 0.1$	0.54	1.00
R1, $\lambda = 0.2$	0.10	1.00
R1, $\lambda = 0.3$	0.00	1.00
R4a	0.02	0.60
R4b	0.00	0.02

ρ	0.05	0.10	0.15	0.20	0.30
within-task r	0.12	0.23	0.33	0.42	0.56
pass, $\lambda = 0$	0.30	0.70	0.94	1.00	1.00
pass, $\lambda = 0.2$	0.22	0.58	0.90	1.00	1.00

Table 3: False-pass rate of the three nulls under a noise channel in which the score tracks the true noise level, by floor ratio (mean floor over median outcome) and by how strongly noise varies (λ). The gate rejects ratios above 0.5.

Floor ratio	runs	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.7$
≤ 0.2	375	0.029	0.020	0.019
0.2–0.3	473	0.032	0.013	0.024
0.3–0.4	547	0.026	0.016	0.030
0.4–0.5	529	0.036	0.028	0.045
0.5–0.6	451	0.044	0.039	0.018
0.6–0.8	786	0.053	0.068	0.097
> 0.8	439	0.048	0.103	0.176

the outcome is mostly noise (R4a, floor 1.6 times the median outcome) the gate catches it before any null is run. Without signal the audit passes 0.6% of 500 runs, and under task-level leakage 0.8%.

Why not divide by the noise level. Standardising the error by its noise level is the natural remedy for the noise channel, and it works when noise is unrelated to difficulty (pass rate 0.98). But on real benchmarks noise usually grows with difficulty, and then the ratio removes the signal it is supposed to protect: at $\lambda = 0.2$, where the score correlates with the noise floor about as strongly as on our medical pools, the ratio test passes genuine signal in 10% of runs, and at $\lambda = 0.3$ in none (Table 2). Covariate adjustment keeps it in every run. Its one weakness is an outcome dominated by noise (R4a), where a five-seed floor is a worse proxy for σ than the score is; this is what the noise-dominance gate is for. Table 3 calibrates it on the hardest noise channel we could construct: a signed true error centred at zero, a score that tracks the true noise level almost exactly, and the nulls run without the gate (setup in Appendix B).

Power and grid size. On the 13×10 grid the audit detects within-task correlations of 0.33 in 90% or more of runs and of 0.42 in all runs; noise that grows with difficulty costs at most 0.12 of power (Table 2). Smaller grids cost power at weak signals: at a within-task correlation of 0.23, the pass rate is 0.14–0.36 on grids of 25–40 cells, 0.38–0.56 on grids of 64–88 cells and 0.76 on 13×8; at a correlation of 0.56 it is 0.72–0.92 on the small grids and 1.00 from 64 cells (Table 4). False passes under model-family confounding reach up to 0.08 on grids of 25–40 cells and stay at or below 0.02 elsewhere. We therefore flag verdicts on grids below 60 cells as `underpowered`: a pass there is informative, a fail much less so.

Stress tests. Three further checks, with 500 runs each (Table 5). With model effects three times the interaction, the audit passes 1–4% of runs, including grids with 30% of cells missing. When models differ in their interaction variance and the score's variance tracks it — a violation of within-task exchangeability — the pass rate rises to 4–8%; we check the real passes for this in Sec. 5. And the harness uses a true error far from zero, whereas real residuals are centred at zero, where the absolute value interacts most with noise; the noise-dominance sweep of Table 3 uses zero-centred errors.

Table 4: Audit pass rate by grid shape, 50 seeds, $\lambda = 0$. R1 columns: genuine signal with within-task correlation about 0.23, 0.56 and 0.73 ($\rho = 0.1, 0.3, 0.5$). R3 and R4b: false-pass rates.

Grid	cells	R1 ρ=0.1	R1 ρ=0.3	R1 ρ=0.5	R3	R4b
5×5	25	0.14	0.72	0.96	0.08	0.02
3×10	30	0.18	0.76	0.98	0.00	0.00
4×8	32	0.28	0.82	1.00	0.06	0.00
5×8	40	0.36	0.92	1.00	0.06	0.00
8×8	64	0.38	1.00	1.00	0.02	0.00
4×16	64	0.56	1.00	1.00	0.00	0.00
11×8	88	0.56	1.00	1.00	0.00	0.00
13×8	104	0.76	1.00	1.00	0.02	0.04
13×10	130	0.70	1.00	1.00	0.00	0.00
20×20	400	0.96	1.00	1.00	0.00	0.00

Table 5: Calibration, 500 runs per row: rate of rejecting with the effect size, and audit pass rate. Strong model effect: model effects three times the interaction, score tracks the model effect. Heteroscedastic models: the interaction variance of both outcome and score depends on the model.

Setting	grid	missing	WITHIN	PROJ	NOISE	pass
R1, no signal	13×10	0	0.060	0.012	0.012	0.006
R2	13×10	0	0.046	0.008	0.008	0.008
Strong model effect	13×8	0	1.000	0.010	0.014	0.010
	13×8	30%	1.000	0.022	0.026	0.016
	5×8	0	0.982	0.050	0.032	0.030
	5×8	30%	0.957	0.051	0.055	0.040
Heteroscedastic models	13×8	0	0.502	0.068	0.080	0.062
	13×8	30%	0.448	0.106	0.098	0.084
	5×8	0	0.376	0.094	0.090	0.064
	5×8	30%	0.314	0.072	0.058	0.036

5 APPLICATION: THREE SCORES ON SEVEN FOUNDATION-MODEL POOLS

Pools, outcome and scores. Each pool is a grid of FMs and binary tasks from one domain (Appendix C): two medical-imaging pools with 13 FMs, language (11 FMs), molecules (4), audio (3), natural images (4) and time series (5), plus a control pool of four hand-crafted audio feature sets on the audio tasks. For every cell a baseline regression predicts the linear-probe learning curve from cheap geometric features of the embeddings; the outcome y is the absolute residual $|b|$ at 25, 50 and 100 labels and at the asymptote. The probe AUC at each label budget is the mean over five seeds, whose spread gives the noise floor; the asymptote comes from one full-data run, so NOISE cannot be run there. We audit LogME (You et al., 2021), the H-score (Bao et al., 2019) and N-LEEP (Li et al., 2021) (LEEP (Nguyen et al., 2020) needs a source classification head, which the frozen encoders do not have; SFDA and NCTI scores were not available for these pools, but the audit applies to any score unchanged). The claim audited is the warning orientation, $-\varphi$; the reversed orientation is audited only descriptively. This gives 84 warning-oriented FM audits and 12 on the control.

Where the claim holds. Eleven of 84 warning-oriented audits pass (Tables 6 and 7), ten of them on the two medical pools. There two of the three scores pass consistently across targets (the H-score fails PROJ on Medical B): on Medical B, LogME and N-LEEP both pass at 25, 50 and 100 labels; on Medical A, LogME passes at all three budgets and the H-score at 100 labels. Their interaction correlations are 0.27–0.44, and they barely change when the noise floor is partialled out (0.24–0.44). On these pools the scores do correlate with the noise floor within tasks, which is exactly the pattern that defeats a ratio test, but the association with the error survives adjustment for it (Figure 2). The within-task correlation is positive in 75–88% of tasks for nine of these passes; the exception is LogME at 25 labels on Medical A, positive in only 3 of 8 tasks and the most noise-exposed pass (floor ratio 0.36, noise share of the interaction variance up to 0.52), which we regard as the weakest. The eleventh pass, the H-score at 100 labels on the natural-image pool, is on a 30-cell grid and flagged `underpowered`. The control passes none.

Table 6: Verdicts of the 12 warning-oriented audits per pool (3 scores $\times$ 4 targets). n: cells per audit after the estimability gate; "removed": share of cells with $N \leq D$; floor: median ratio of mean noise floor to median outcome. Failures are split by blocking null (W: WITHIN, P: PROJ, N: NOISE). u: grid below 60 cells (`underpowered`).

Pool	n	removed	floor	pass	incompl.	noise-dom.	fail (W/P/N)
Medical B	104	11%	0.20	**6**	2	–	0/4/0
Medical A	103	28%	0.25	**4**	1	–	3/2/2
Language	73	12%	–	–	–	–	12/0/0
Molecules	64	0%	0.51	–	1	6	4/1/0
Natural images (u)	24–30	3%	0.25	1	2	–	9/0/0
Time series (u)	26	64%	0.29	–	–	–	12/0/0
Audio (u)	30	0%	0.92	–	–	6	0/6/0
FM total				11	6	12	40/13/2
audio, hand-crafted (u)	40	0%	0.52	–	–	6	1/5/0

Table 7: The 11 passing warning-oriented audits: correlations of the three nulls (all $p \leq 0.04$) and the share of tasks with a positive within-task correlation.

Pool, target	score	WITHIN	PROJ	NOISE	tasks +
Medical A, 25 labels	LogME	0.20	0.27	0.26	0.38
Medical A, 50 labels	LogME	0.27	0.36	0.33	0.75
Medical A, 100 labels	LogME	0.29	0.39	0.39	0.75
Medical A, 100 labels	H-score	0.34	0.44	0.44	0.75
Medical B, 25 labels	LogME	0.33	0.31	0.24	0.75
Medical B, 25 labels	N-LEEP	0.50	0.42	0.34	0.75
Medical B, 50 labels	LogME	0.43	0.36	0.32	0.88
Medical B, 50 labels	N-LEEP	0.52	0.40	0.36	0.88
Medical B, 100 labels	LogME	0.39	0.32	0.33	0.88
Medical B, 100 labels	N-LEEP	0.44	0.35	0.36	0.88
Natural images, 100 labels (u)	H-score	0.53	0.56	0.58	0.71

Where it fails, and where the audit cannot decide. On the language and time-series pools every warning-oriented audit fails the first null; the within-task correlation is negative in 21 of the 24, i.e. the scores are *lower* where the error is lower, the opposite of the warning reading. On the audio pool and the control the within-task correlations are large (0.4–0.8) but vanish once model effects are removed (PROJ correlations -0.1 to 0.3, none significant): the scores track which models are strong, the model-family confounding of regime R3. Half the audits on the molecule and audio pools are `noise-dominated`: their noise floor is about half the median error or more, and five seeds are not enough to audit them. The six `incomplete` verdicts are all at the asymptotic target, which has one run per cell. The noise floor could explain at most 11–52% of the interaction variance of the medical outcomes, depending on the target, which is why the noise null matters there.

Robustness. The passes do not depend on the thresholds: with noise-dominance thresholds of 0.4 or 0.6, or $B = 1000$ permutations, the same 11 audits pass; with looser effect sizes ($r \geq 0.05$, $r^2 \geq 0.025$) 17 pass, with stricter ones ($r \geq 0.20$, $r^2 \geq 0.10$) 9; the control passes none in any setting. Models do differ in how variable their errors are (the per-model spread of the outcome's interaction component varies 3–7-fold on the medical pools, and co-varies with the score's), which is the situation where the harness showed inflated false passes; standardising the interaction components per model keeps all 11 passes ($r = 0.25$–0.66, $p \leq 0.018$). At the default thresholds no audit passes in the reversed orientation (two `incomplete` on the language pool).

The estimability gate matters in both directions. Without it, only 3 warning-oriented audits pass: nine medical passes turn into failures, because degenerate cells add noise that masks the signal, and one LogME audit on the natural-image pool becomes a pass that disappears when the two cells with $N \leq D$ are removed.

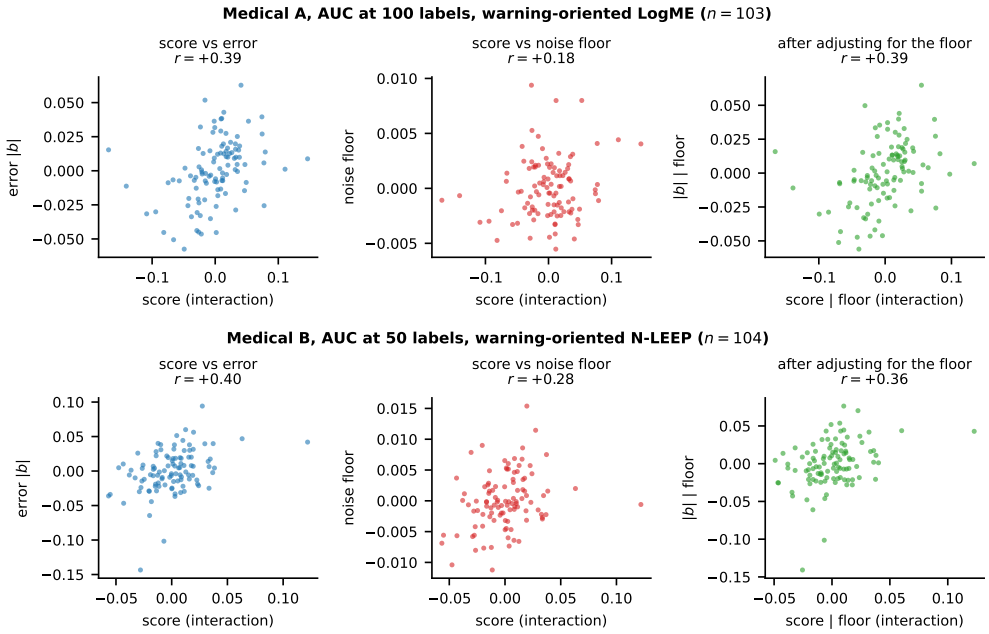


Figure 2: What NOISE sees on two passing audits. All quantities are interaction components (model and task effects removed). The warning-oriented score predicts the error (left) and also the noise floor (middle), but its association with the error is essentially unchanged once the floor is regressed out (right). The adjustment is partial because a five-seed floor is noisy; these passes are protected by their floor ratios (0.15–0.36), where the sweep of Table 3 finds a residual channel producing false passes in at most 3% of runs.

6 REPORTING A CELL-LEVEL CLAIM

The application suggests what an evaluation of a cell-level warning should report so that a reader can audit it.

1. **The claimed direction**, fixed before looking at the data, and the effect size of every null, not only p-values.
2. **The grid**: numbers of models and tasks per outcome and which cells are missing. Below about 60 cells the audit has little power at moderate effects (Sec. 4).
3. **Estimability**: N and D for every cell and the share of cells on which the score is not defined.
4. **The noise floor**: the number of seeds behind each outcome and the per-cell seed standard deviation. Without it the noise channel cannot be tested; with a floor above about half the typical error, more seeds are needed before it can be.
5. **The capacity of the outcome**: the interaction share frac_γ and the bound on its noise share.
6. **The structured verdict** of Appendix A, including the blocking null or the reason the audit could not decide.

7 RELATED WORK

Evaluating transferability scores. Transferability scores are usually evaluated by rank correlation with fine-tuned accuracy over models on one task (Nguyen et al., 2020; You et al., 2021; Li et al., 2021; Shao et al., 2022; Wang et al., 2023). Agostinelli et al. (2022) show that such evaluations are unstable under small changes of the setup, and Abou Baker & Handmann (2024) that the best score depends on the dataset. We address a different reading of the scores — as cell-level warnings across a grid of models and tasks — and ask what evidence supports it.

Permutation tests and nuisance effects. Permutation tests for classifier performance (Ojala & Garriga, 2010), their exact p-values (Phipson & Smyth, 2010) and their construction (Noreen, 1989; Welch, 1990; Good, 2005) are standard. In factorial designs, unrestricted permutation is invalid when nuisance effects are present, and one must permute within blocks or permute residuals of a reduced model (Anderson & ter Braak, 2003; Freedman & Lane, 1983; Winkler et al., 2014); PROJ and NOISE are instances of these two ideas for (FM, task) grids. Ratio variables induce spurious correlation (Kronmal, 1993), which is why NOISE adjusts for the noise floor instead of dividing by it.

Leakage, contamination and reporting standards. REFORMS (Kapoor et al., 2024) gives a checklist for ML-based science including leakage, and work on variance and reporting in ML benchmarks argues for stronger evidence standards (Lipton & Steinhardt, 2019; Sculley et al., 2018; Bouthillier et al., 2021; Dodge et al., 2019). Leakage specific to foundation models has been documented for time-series benchmarks (Meyer et al., 2025) and multiple-choice language benchmarks (Hidayat et al., 2025; Sainz et al., 2023); pipeline stress tests perturb data to probe robustness (Zhu et al., 2025). These concern train–test overlap and contamination; our audit targets artefacts that arise within a benchmark grid when a cell-level claim is supported by a pooled correlation.

8 Discussion and Limitations

What a verdict means. A `pass` means a claim survived task-level leakage, model-family confounding and the noise channel on a grid large enough to detect them; it is necessary, not sufficient, for the claim to hold on new models or tasks. A `fail` names the first condition the evidence does not meet: an association in the claimed direction, one that survives removing model effects, or one that survives adjusting for seed noise. The other verdicts say that the data cannot settle the question.

Pretraining contamination and benchmark construction. A benchmark can be close to some models' pretraining data, or built in a way that favours some model families. Such effects shift the benchmark as a whole; the nulls rearrange cells within a grid, so these effects leave no signature for them to detect. Identifying them needs evidence from outside the grid, such as held-out benchmarks or contamination tests (Sainz et al., 2023).

Limitations. (i) NOISE adjusts linearly for a five-seed estimate of the noise floor; it under-adjusts when the outcome is dominated by noise, which the noise-dominance gate guards against with a threshold whose false-pass rate we map in Appendix B. (ii) Power is low for weak effects on grids below about 60 cells. (iii) Within-task permutation assumes that, after the effects each null removes, models are exchangeable within a task. When models differ in the variance of their errors and the score's variance tracks it, false passes rise to 4–8% in the harness; standardising per model is a simple check, which the real passes survive, and strongly clustered model families may call for permutation within families. (iv) Only one hand-crafted control pool was available, and it is small. (v) The application uses a single outcome construction (the error of a learning-curve predictor); other outcomes may behave differently.

Conclusion. A pooled correlation between a transferability score and cell-level error does not show that the score knows which cells are hard. The audit tests three alternative explanations and says when the data cannot decide. On three published scores it finds cell-level warnings on two medical-imaging pools that survive adjustment for model strength and seed noise; evaluations of such warnings should report the claimed direction, per-cell seed variation and effect sizes so that the audit can be run (Sec. 6).

AI Use Statement

In this work, we used generative AI tools to help develop and refine the methodology (the design of the null set, including the covariate-adjusted noise null, the noise-dominance gate and the verdict set); to formulate mathematical claims (the equivalence of the raw and task-demeaned tests under within-task permutation, checked numerically); to design and give feedback on the experiments; to generate synthetic data sets (the harness of Sec. 4); to implement methods (the audit library and all

analysis scripts); to interpret results; and to probe the audit for failure modes (whether a constant noise floor suffices, whether effect-size thresholds are needed, and whether a noise null removes genuine signal). We have not used generative AI tools to prove mathematical claims, to translate text, or to clean or reformat data sets, and qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to draft and edit text, to generate figures and tables, and to search for and verify references. We have reviewed all AI-assisted work: every reference was checked against its published source; every number in the paper is regenerated by the released scripts from the stored result tables. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

The audit uses public benchmarks, including public, de-identified clinical imaging datasets, and embeddings precomputed from frozen models; no new data involving human subjects were collected. The audit is meant to prevent unsupported claims that a score can flag where a model will fail, which matters most in clinical deployment. A `pass` is not evidence that a model is safe to deploy, and a `fail` is not evidence that it is unsafe.

Reproducibility statement

The supplementary material contains the audit library; the code that generates the synthetic harness and every table of Sec. 4 and Appendix B; the real-data audit with its robustness and gate analyses and the records of all audits; the code for the estimability example and Figure 2; and the stored outcome, score and probe-run tables they read. A README maps each table and figure to the code that produces it. Definitions and pseudocode are in Sec. 3 and Appendix A; generators in Appendix B; pools, outcomes and score implementations in Appendix C.

References

Nermeen Abou Baker and Uwe Handmann. One size does not fit all in evaluating model selection scores for image classification. *Scientific Reports*, 14(1):30239, 2024. doi: 10.1038/s41598-024-81752-w.

Andrea Agostinelli, Michal Pándy, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. How stable are transferability metrics evaluations? In *Computer Vision – ECCV 2022*, volume 13694 of *Lecture Notes in Computer Science*, pp. 303–321. Springer, 2022. doi: 10.1007/978-3-031-19830-4_18.

Marti J. Anderson and Cajo J. F. ter Braak. Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation*, 73(2):85–113, 2003. doi: 10.1080/00949650215733.

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12449–12460, 2020.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2309–2313. IEEE, 2019. doi: 10.1109/ICIP.2019.8803726.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4): 295–300, 1982. doi: 10.2307/1267823.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 3, 2021.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction, 2020. URL https://arxiv.org/abs/2010.09885.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1224.

David Freedman and David Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983. doi: 10.1080/07350015.1983.10509354.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Proceedings of Interspeech 2021*, pp. 571–575, 2021. doi: 10.21437/Interspeech.2021-698.

Phillip I. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Series in Statistics. Springer, New York, 3rd edition, 2005. doi: 10.1007/b138696.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, 2024.

Naila Shafirni Hidayat, Muhammad Dehan Al Kautsar, Alfan Farizki Wicaksono, and Fajri Koto. Simulating training data leakage in multiple-choice benchmarks for LLM evaluation. In *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP)*, pp. 21–39. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.eval4nlp-1.3.

Sayash Kapoor, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russell A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M. Stewart, Gilles Vandewiele, and Arvind Narayanan. REFORMS: Consensus-based recommendations for machine-learning-based science. *Science Advances*, 10 (18):eadk3452, 2024. doi: 10.1126/sciadv.adk3452.

Richard A. Kronmal. Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 156(3):379–392, 1993. doi: 10.2307/2983064.

Yandong Li, Xuhui Jia, Ruoxin Sang, Yukun Zhu, Bradley Green, Liqiang Wang, and Boqing Gong. Ranking neural checkpoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2662–2672, 2021. doi: 10.1109/CVPR46437.2021.00269.

Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. *ACM Queue*, 17(1):45–77, 2019. doi: 10.1145/3317287.3328534.

Marcel Meyer, Sascha Kaltenpoth, Kevin Zalipski, and Oliver Müller. Rethinking evaluation in the era of time series foundation models: (Un)known information leakage challenges. *arXiv preprint arXiv:2510.13654*, 2025.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305. PMLR, 2020.

Eric W. Noreen. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley, New York, 1989. ISBN 978-0-471-61136-3.

Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(62):1833–1863, 2010.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, Brisbane, Australia, 2015. ACM. doi: 10.1145/2733373.2806390.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.722.

D. Sculley, Jasper Snoek, Alexander B. Wiltschko, and Ali Rahimi. Winner's curse? On pace, progress, and empirical rigor. In *6th International Conference on Learning Representations (ICLR 2018), Workshop Track*, 2018.

Wenqi Shao, Xun Zhao, Yixiao Ge, Zhaoyang Zhang, Lei Yang, Xiaogang Wang, Ying Shan, and Ping Luo. Not all models are equal: Predicting model transferability in a self-challenging Fisher space. In *Computer Vision – ECCV 2022*, volume 13694 of *Lecture Notes in Computer Science*, pp. 286–302. Springer, 2022. doi: 10.1007/978-3-031-19830-4_17.

George W. Snedecor and William G. Cochran. *Statistical Methods*. Iowa State University Press, Ames, IA, 8th edition, 1989. ISBN 0-8138-1561-6.

Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5526–5535, 2023. doi: 10.1109/ICCV51070.2023.00511.

William J. Welch. Construction of permutation tests. *Journal of the American Statistical Association*, 85(411):693–698, 1990. doi: 10.1080/01621459.1990.10474929.

Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014. doi: 10.1016/j.neuroimage.2014.01.060.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143. PMLR, 2021.

Jiongli Zhu, Geyang Xu, Felipe Lorenzi, Boris Glavic, and Babak Salimi. Stress-testing ML pipelines with adversarial data corruption. *Proceedings of the VLDB Endowment*, 18(11):4668–4681, 2025. doi: 10.14778/3749646.3749721.

A AUDIT SPECIFICATION

Input. Outcome y, score φ (in the claimed orientation), model and task indices, per-cell noise floor $s = \sigma_{\text{seed}}/\sqrt{n_s}$ (optional), and per-cell N and D (optional). **Output.** A structured record; for warning-oriented LogME at 100 labels on Medical A (Sec. 5):


```
{"n_in": 143, "n_removed": 40, "n": 103, "underpowered": false,
 "frac_gamma": 0.113, "frac_gamma_noise_adj": 0.096, "floor_ratio": 0.166,
 "verdict": "pass", "blocking": null, "task_consistency": 0.75,
 "WITHIN": {"r": 0.285, "p": 0.005, "ok": true},
 "PROJ":   {"r": 0.388, "p": 0.005, "ok": true},
 "NOISE":  {"r": 0.386, "p": 0.005, "ok": true}}
```

Pseudocode. `interaction(v)` removes model and task effects by alternating demeaning (exact for complete grids, least squares with missing cells); `perm(v)` permutes v across models within each task; $p = (1 + \#\{\text{null} \geq \text{obs}\})/(B + 1)$ with $B = 200$.

```
def WITHIN(y, phi):   # task-level leakage
    stat = lambda yy: corr(demean_task(phi), demean_task(yy))
    return stat(y), [stat(perm(y)) for _ in range(B)]       # r >= 0.10

def PROJ(y, phi):     # model-family confounding
    stat = lambda yy: corr(interaction(phi), interaction(yy))
    return stat(y), [stat(perm(y)) for _ in range(B)]       # r >= sqrt(.05)

def NOISE(y, phi, s): # noise channel, Freedman-Lane
    yg, pg, sg = interaction(y), interaction(phi), interaction(s)
    fit = ols_fit(yg, sg); res = yg - fit
    stat = lambda yy: partial_corr(yy, pg, given=sg)
    return stat(yg), [stat(interaction(fit + perm(res))) for _ in range(B)]

def audit(y, phi, s, N, D):
    keep = N > D                                    # estimability gate
    if keep.sum() < 20:           return "not-estimable"
    if frac_gamma(y[keep]) < 0.05: return "capacity-exceeded"
    if mean(s) / median(y) > 0.5:  return "noise-dominated"
    for null in (WITHIN, PROJ, NOISE):              # NOISE skipped if s missing
        if not rejects(null):     return ("fail", null)
    return "pass" if s is not None else "incomplete"
```

The supplementary material contains the full implementation and the records of all 192 real-data audits. The conjunction's false-pass probability is at most that of its least permissive null: if A_i are the pass events of the nulls, $\Pr(\bigcap_i A_i) \leq \min_i \Pr(A_i)$ for any dependence between them.

B Synthetic harness

Generators. With $\alpha \sim \mathcal{N}(0,1)^{n_F}$, $\beta \sim \mathcal{N}(0,1)^{n_T}$, $\gamma, \varepsilon, z \sim \mathcal{N}(0,1)^{n_F \times n_T}$ and $c = 1.5 + 0.3\alpha + 0.3\beta$: R1: $b^* = c + \rho\gamma$, $\sigma = 0.6e^{\lambda\gamma}$, $\varphi = \gamma + 0.3\varepsilon$; R2: $b^* = c + 0.5z$, $\sigma = 0.6$, $\varphi = \beta + 0.3\varepsilon$; R3: as R2 with $\varphi = \alpha + 0.3\varepsilon$; R4a: $b^* = 0$, $\sigma \sim U(0.2, 1)$, $\varphi = \sigma + 0.1\varepsilon$; R4b: $b^* = c + 0.5\gamma$, $\sigma = 0.6e^z$, $\varphi = z + 0.3\varepsilon$; R3+R4b: as R4b with $\varphi = \alpha + z + 0.3\varepsilon$. Each cell draws five seeds from $\mathcal{N}(b^*, \sigma^2)$; y is the absolute value of their mean and the noise floor is their sample standard deviation over $\sqrt{5}$. Defaults: 50 seeds per setting, $B = 200$; the full run takes about four minutes on ten CPU cores.

Noise-dominance sweep. To calibrate the gate we simulate a noise channel that is as hard as we could make it: the true error is signed and centred at zero, $b^* = 0.3\alpha + 0.3\beta + \gamma$ (independent of the score), the seed noise is $\sigma = \sigma_0 e^{\lambda z}$, and the score tracks the true noise level almost exactly, $\varphi = z + 0.1\varepsilon$, so that the five-seed floor is a worse proxy for σ than the score is. We vary $\sigma_0 \in \{0.3, \ldots, 1.6\}$ and $\lambda \in \{0.2, 0.4, 0.7\}$ (200 runs each, 13×8 grid), run the three nulls without the gate, and bin the false-pass rate by the observed floor ratio (Table 3 in Sec. 4).

C Real-data details

Pools. All tasks are binary. *Medical* (two pools with different task sets, Medical A and Medical B): 13 general-purpose and medical vision FMs (CLIP and OpenCLIP, SigLIP, DINOv2, BiomedCLIP, PubMedCLIP, RAD-DINO, RETFound) on public clinical imaging datasets covering retinal fundus, dermatology, histopathology, chest radiography, optical coherence tomography and brain MRI; 8 tasks per pool remain after the estimability gate. *Language*: 11 text encoders and decoder language models on 8 text-classification tasks. *Molecules*: ChemBERTa SMILES encoders (Chithrananda et al., 2020) (two pretraining corpora, two poolings) on 16 MoleculeNet-derived tasks (BACE, BBBP, ClinTox, Tox21). *Audio*: wav2vec 2.0 (Baevski et al., 2020) and AST (Gong et al., 2021) (two poolings) on 10 pairwise category tasks of ESC-50 (Piczak, 2015); the control uses four hand-crafted feature sets (temporal, spectral, mel and combined; D from 40 to 224) on the same tasks. *Natural images*: 4 vision FMs on 8 classification datasets. *Time series*: Chronos-family (Ansari et al., 2024), MOMENT (Goswami et al., 2024) and Time-MoE encoders on UCR classification tasks.

Outcome and noise floor. For each cell a linear probe is trained at several label budgets with five seeds; a baseline regression predicts the learning curve from cheap geometric features of the embeddings, and the outcome is the absolute residual at 25, 50 and 100 labels (where the observed value is the five-seed mean) and at the asymptote (one full-data run). The noise floor is the standard deviation of the five seed AUCs over $\sqrt{5}$. On the language pool the seed runs are incomplete, so no floor is used there.

Scores. LogME with the standard implementation, the H-score with a 10^{-4} ridge on the feature covariance, and N-LEEP with a Gaussian mixture on PCA-reduced embeddings, all computed on the labelled pool of each cell. Share of cells removed by the estimability gate: Molecules and Audio 0%, Natural images 3.4%, Medical B 11.1%, Language 12.0%, Medical A 28.0%, Time series 64.1%, control 0%.

Noise share and capacity. The bound $\sum s^2 / \sum \gamma^2$ on the share of the outcome's interaction variance due to seed noise is 0.52, 0.30 and 0.15 on Medical A at 25, 50 and 100 labels, and 0.42, 0.22 and 0.11 on Medical B. It exceeds 0.8 on the audio and molecule pools, and also on the natural-image pool at 25 and 50 labels (1.23, 1.47) and the time-series pool at 25 labels (1.29), which the floor-ratio gate does not flag because model and task effects inflate their median outcome; all of these fail WITHIN. The floor-ratio gate screens for the noise channel; it does not measure how much of the interaction is noise. Subtracting the bound from the interaction variance leaves every one of the 11 passes with a noise-adjusted interaction share of at least 0.096. The capacity gate did not bind on any real grid (no `capacity-exceeded` verdict).