

How High Is 0.6? Floors, Ceilings, and Headroom in Interpretability Probing

Pranjal Garg
Independent Researcher

Abstract

Probes are the workhorse of interpretability. If a model’s hidden states predict a variable, the model is said to represent it. But a probe score has no fixed meaning. An R^2 of 0.6 may only reflect what the input already gives away, and the same score can mean different things on different data. We propose reading every probe score against two reference points: a floor, what a declared set of simple inputs already predicts, and a ceiling, what the full input can predict. The gap between them, the headroom, is the range in which a probe can show that a model computes something beyond the simple inputs. We prove that headroom vanishes in two ways: the target stops depending on a hidden variable the model must infer, or the input stops revealing it. We test this on transformers trained for in-context meta-analysis, which must infer the hidden heterogeneity between studies to weight them correctly, and where both reference points are known. Under distribution shift, probe scores fall and prediction error rises 12–15 $\times$, yet the model recovers a similar share of the headroom, indicating that the data lost information, not the representation. We then analyze the real models. The single-cell foundation model scGPT encodes biological variability only partially. We also revisit four influential LLM probing studies, which claim that models represent geography, the state of an Othello board, truth, and the demographics of their users. Against a floor computed from the input text alone, some of these claims hold, while others are largely explained by the text itself.

1 INTRODUCTION

Understanding what information is represented inside a complex system is a central problem across biological and artificial intelligence. In neuroscience, decoding analyses use neural population activity to ask whether variables such as sensory features, cognitive states, or behavioral quantities are represented in the brain; analogous methods are now widely used to study the hidden representations of learned artificial networks (Quiñero and Panzeri, 2009; Naselaris et al., 2011; Belinkov, 2022). In modern AI, where high-dimensional internal states are otherwise difficult to interpret directly, *probes* provide a particularly simple and scalable readout: a classifier or regressor is trained to predict a quantity of interest from a model’s hidden states, and its predictive performance is used to assess how strongly that quantity is represented (Alain and Bengio, 2016; Hewitt and Liang, 2019). Such measurements underlie claims about what models encode, how representations emerge during learning and how information is distributed across layers (Tenney et al., 2019; Saphra and Lopez, 2019). They therefore carry substantial scientific weight, as conclusions about the internal organization of both biological and artificial systems can depend on what a probe score is taken to mean.

Yet this widespread use exposes a basic interpretive problem: a raw probe score has no intrinsic scale of evidential strength. A high R^2 may arise because the target is predictable from readily available input information rather than because the representation captures the computation of interest, and the same numerical score can carry different implications across distributions as this baseline predictability changes (Ravichander et al., 2021; Elazar et al., 2021; Hewitt and Liang, 2019). Interpreting probe scores therefore requires knowing not only what is decodable but also how much decodability is achievable without the computation of interest and how much is achievable from all available information. We study this distinction in a controlled evidence-aggregation task. We train

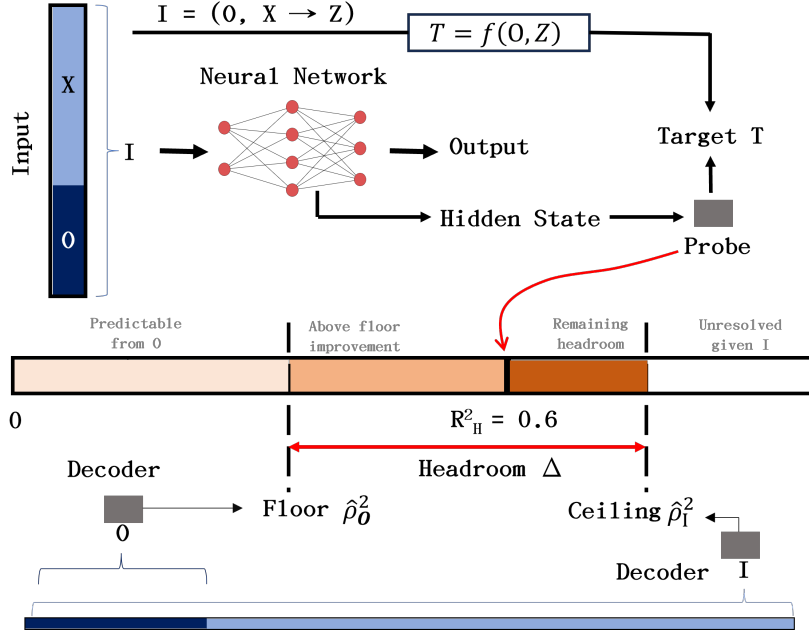


Figure 1: **Schematic of the floor–ceiling normalized probe R^2 .** A probe’s raw decodability (R^2_H) for a target T is normalized between an estimated floor ($\hat{\rho}_O^2$), achievable from restricted information O , and an estimated ceiling ($\hat{\rho}_I^2$), achievable using full information I . The gap between them defines the headroom (Δ).

transformer models in a random-effects meta-analysis setting where the sufficient statistics and the oracle Bayes estimator conditional on the true heterogeneity parameter are known in closed form. Under a distribution shift, the raw decodability of the sufficient statistic falls and the prediction error rises, and the raw scores alone cannot say whether the representation degraded or the data became less informative.

To separate computation from statistical artifact, we **introduce a framework that normalizes probe scores against the bounds implied by the information sets that the scientific question specifies**. We partition the available information into restricted information (O) and additional evidence (X) used to infer a latent variable (Z). We evaluate probe decodability between two population reference points (under the true data distribution): a *floor* (ρ_O^2), which measures the predictability achievable from the baseline alone, and a *ceiling* (ρ_I^2), which measures the maximum predictability achievable given full information (I), which represents all available evidence. We define the gap between these reference points as *headroom* (Δ): the maximum variance a probe could explain by successfully inferring the latent variable Z from X (Figure 1).

The theoretical result demonstrates that this headroom can vanish through two mechanisms: *weak target sensitivity* (where the target ceases to depend on the latent variable) and *weak identification* (where the ev-

idence becomes too noisy to identify the latent variable). When headroom vanishes, raw probe scores can cease to provide evidence about the computation of interest. We test the estimation procedure in the meta-analysis task and a synthetic single-cell model, where the generative model is known and the reference points can be computed exactly by numerical integration. We then apply it to real single-cell data and to published LLM probes, where the reference points can only be estimated from samples. Analyses of a fine-tuned chemistry model and recorded neural activity appear in Appendix J.

Contributions: (1) we introduce a floor, ceiling and headroom normalization for interpretability probes and identify two distinct mechanisms by which this headroom vanishes: weak target sensitivity and weak identification of the latent variable; (2) we show that under distribution shift raw decodability can reflect reduced headroom rather than a degraded representation; (3) we test this estimation procedure in different domains, showing how the choice of floor and ceiling distinguishes different scientific questions; (4) we show that the same raw probe score can imply different amounts of recovered headroom across settings.

2 RELATED WORK

Probing and its baselines. A probe trains a simple readout on hidden states and treats its predic-

tive performance as evidence that a variable is represented (Alain and Bengio, 2016; Belinkov, 2022). Raw probe accuracy underdetermines this conclusion in three ways. Expressive probes can learn the task themselves (Hewitt and Liang, 2019). Random encoders already support high accuracy (Zhang and Bowman, 2018; Wieting and Kiela, 2019). Decodable information need not be used by the model (Ravichander et al., 2021; Elazar et al., 2021; Kumar et al., 2022). Proposed remedies either control probe capacity (Hewitt and Liang, 2019; Voita and Titov, 2020) or estimate information content (Pimentel et al., 2020). Pimentel et al. (2020) note that, by the data-processing inequality, no representation carries more information about a target than its input does.

Closest to our floor is conditional probing (Hewitt et al., 2021). Building on $\mathcal{V}$ -information (Xu et al., 2020), it measures what a representation adds beyond a baseline representation. We differ in three respects: (i) the floor is defined on a declared restricted *input* set O chosen by the scientific question; (ii) we pair it with a ceiling from the full input I , turning the data-processing bound into an estimable reference point against which scores are normalized; (iii) we show that both reference points move with the data regime, so a fixed baseline can misattribute a regime change to the representation; we show this in Section 4.2.1 and Appendix L, where the conditional gain halves under a distribution shift while the headroom-normalized score does not.

Interpreting scientific foundation models. Probing and related readouts are increasingly being used to make scientific claims about learned models, for example, that protein language models encode structure (Rives et al., 2021) and that single-cell foundation models encode regulatory or cell-state information (Cui et al., 2024; Theodoris et al., 2023). Several audits find that such models often fail to beat simple baselines (Kedzierska et al., 2025; Boiarsky et al., 2024; Ahlmann-Eltze et al., 2025). A recent audit of scGPT and Geneformer finds that the apparent biological structure in attention adds nothing beyond trivial gene-level statistics (Kendiukhov, 2026). These studies choose baselines case by case and report raw differences. We formalize the question they share: how much a representation recovers beyond a declared simple baseline, relative to what is predictable from the available inputs, and apply it to scGPT, a fine-tuned TabPFN (Hollmann et al., 2025; Grinsztajn et al., 2026), and neural population activity. The latter two appear in Appendix J.

Distribution shift and causal relevance. Interpretations that are faithful in distribution can fail un-

der systematic shift (Friedman et al., 2023). Causal methods such as nullspace projection and interchange interventions test whether decodable information is used (Ravfogel et al., 2020; Geiger et al., 2024). Our framework is complementary. It does not test causal use; it specifies, before any intervention, how much evidential weight a correlational score can carry in a given regime. It also shows that a drop in raw decodability under distribution shift can reflect reduced headroom rather than a degraded representation.

Controlled testbeds with known sufficient statistics. Meta-trained sequence models approximate Bayes-optimal inference (Mikulik et al., 2020; Müller et al., 2021), and transformers learn statistical estimators in context (Garg et al., 2022; Akyürek et al., 2022; Bai et al., 2023). Probes have been used to compare their internal quantities with classical estimators (Teh et al., 2025). We build on this line with random-effects meta-analysis (Higgins and Thompson, 2002).

3 CONTROLLED META-ANALYSIS SETUP

3.1 Task and model

Controlled random-effects meta-analysis. Each example contains K studies estimating an underlying effect μ . Study-specific effects vary around μ with between-study heterogeneity τ^2 , while each reported estimate x_i is additionally corrupted by study-specific sampling variance σ_i^2 . Marginally, $x_i \mid \mu, \tau^2, \sigma_i^2 \sim \mathcal{N}(\mu, \sigma_i^2 + \tau^2)$. If τ^2 were known, each study would receive precision weight $w_i = \frac{1}{\sigma_i^2 + \tau^2}$, and all information relevant to estimating μ would be summarized by the sufficient statistics $S_1 = \sum_i w_i$ and $S_2 = \sum_i w_i x_i$. Here, S_1 is the total precision of the studies and S_2 is their precision-weighted effect. We therefore use $\log S_1$ as the primary target for representation probing, and separately probe τ^2 to test whether the model represents the hidden heterogeneity needed to construct the weights.

We define **distribution shift** as an out-of-distribution (OOD) change from a heterogeneity-dominated regime to a sampling-variance-dominated regime. We quantify the regime by the sampling-variance share $r_i = \frac{\sigma_i^2}{\sigma_i^2 + \tau^2}$, whose median increases from 0.31 in distribution (ID) to 0.71 OOD. Thus, the ID regime is predominantly heterogeneity-driven, whereas the OOD regime is predominantly sampling-variance-driven. The distribution shift changes which source of uncertainty determines the precision weights. For full details, see Appendix A.

Model and training. We train causal pre-norm transformers with attention and MLP blocks at five sizes ranging from 3.3k to 794k parameters (xs, s, m, l, xl). The transformer model receives $(x_i, \log \sigma_i^2)$ as input, and the true heterogeneity τ^2 is hidden. It predicts the overall effect μ under a causal mask (Appendix A.3).

Probing and architectural controls. All probe scores and estimated reference points are computed using a bank of linear and nonlinear regressors (ridge, MLP, and gradient boosting) for every layer (Appendix B). We report the standard R^2 for every held-out target. We probe $\log S_1$ rather than S_1 because the logarithm reduces target skew from 3.9 to -0.6 .

3.2 Raw probe behavior under the distribution shift

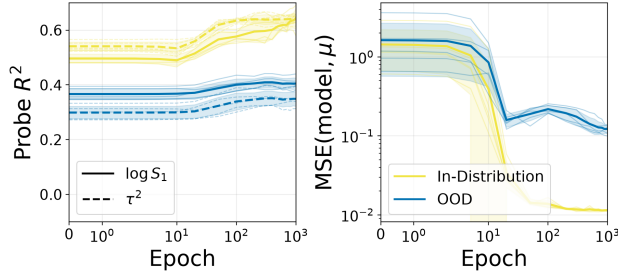


Figure 2: **In-distribution and OOD learning trajectories.** Probe decodability for $\log S_1$ and τ^2 and final-position model MSE across training (xl model).

Table 12 and Figure 2 show that the distribution shift degrades all three measured quantities. For models s through xl, probe R^2 for $\log S_1$ falls (from 0.64 to 0.40 for xl), probe R^2 for τ^2 falls (from 0.64 to 0.34 for xl), and prediction MSE at the final position rises 12–15 $\times$. In distribution, model error is within 3% of the oracle conditional-Bayes risk; out of distribution it is 2.1–2.6 $\times$ that risk. Both targets nonetheless remain partly decodable OOD ($R^2 \approx 0.29$ –0.40 for $\log S_1$ and 0.26–0.34 for τ^2). The xs model is excluded from these comparisons because it does not reliably learn either target in distribution ($R^2 = 0.28 \pm 0.20$ for $\log S_1$).

The raw scores admit two explanations: the representation degraded, or the OOD data carry less information about the targets. A further concern applies to $\log S_1$. When sampling variance dominates, $S_1 = \sum_i (\sigma_i^2 + \tau^2)^{-1} \approx \sum_i \sigma_i^{-2}$ is predictable from the sampling variances alone, so the OOD score of 0.40 might require no inference about τ^2 . If so, the best prediction of $\log S_1$ from the sampling variances alone should rise sharply under the shift. We formalize this quantity as the floor ρ_O^2 in Section 4.1; Section 4.3

shows that it rises only from 0.01 to 0.09. Raw probe scores cannot separate these explanations; Section 4.1 introduces a normalization that does.

4 NORMALIZING PROBE DECODABILITY

4.1 Floor, ceiling and headroom

Let $T \in L^2$ be a non-constant target quantity whose representation we wish to audit. Let O denote restricted information and I the full information available to the system, with $\sigma(O) \subseteq \sigma(I)$. Equivalently, we may write $I = (O, X)$, where X denotes the additional evidence beyond O . The representation under study is denoted by H ; for deterministic models, $H = \Phi(I)$.

We define two population reference points,

$$\rho_O^2 := 1 - \frac{\mathbb{E}[\text{Var}(T | O)]}{\text{Var}(T)}, \quad \rho_I^2 := 1 - \frac{\mathbb{E}[\text{Var}(T | I)]}{\text{Var}(T)},$$

and their difference $\Delta := \rho_I^2 - \rho_O^2 \geq 0$ (Propositions E.1 and E.2).

Here ρ_O^2 is the maximum population R^2 attainable from the restricted information O , whereas ρ_I^2 is the maximum attainable from the full information I . We call ρ_O^2 the *floor*: it is not a lower bound on an arbitrary probe score, but the level of predictability already available without using the additional evidence X . Consequently, a probe score at or below ρ_O^2 does not by itself establish that the representation contains predictive information about T beyond O (Proposition E.2; Remark E.3).

The headroom Δ quantifies how much additional predictability I supplies beyond O . It equals the normalized variance of the Bayes update that X supplies about T beyond O (Proposition F.1), and it vanishes through weak target sensitivity or weak identification (Corollary F.2; Section 4.3).

For a representation H , let R_H^2 denote the held-out probe performance for predicting T from H . When H is measurable with respect to I , no predictor based on H can exceed the population ceiling ρ_I^2 (Proposition E.1). When $\Delta > 0$, we therefore normalize representation decodability as $U_H := \frac{R_H^2 - \rho_O^2}{\rho_I^2 - \rho_O^2}$. Here U_H measures probe performance relative to the headroom available beyond O . Each reference point has three versions. The *population* reference points are defined above. Second is *estimating* the reference points $\hat{\rho}_O^2, \hat{\rho}_I^2$, by fitting a fixed decoder bank to O and to I on held-out data (the *estimation procedure*; Appendix B). When the generative model is known (meta-analysis, synthetic single-cell), we additionally compute them

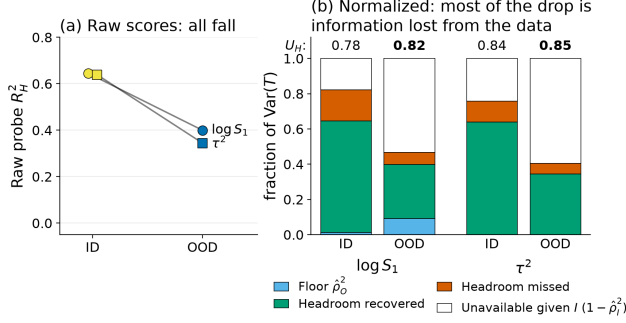


Figure 3: **Raw and normalized probe performance under distribution shift.** Raw probe scores decrease OOD, while normalization by the floor and ceiling shows that most of the decline reflects reduced available information.

exactly by numerical integration. In both of them, estimates closely track exact values (Appendix H, Figure 7). For meta-analysis and single-cell test cases, we estimate the reference points with feature representations $\phi_O(O)$ and $\phi_I(I)$. The feature maps are fixed deterministic functions of their corresponding information sets (Appendix I). Estimated ceilings are conservative, so an underestimated ceiling inflates U_H .

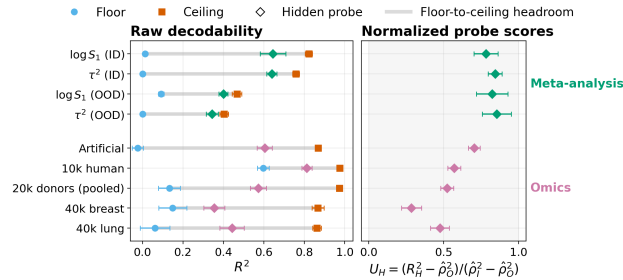


Figure 4: **Raw and headroom-normalized decodability across domains.** Left: R^2 of the floor, the ceiling and the representation probe. Right: the normalized decodability $U_H = (R_H^2 - \hat{\rho}_O^2)/(\hat{\rho}_I^2 - \hat{\rho}_O^2)$, the fraction of the headroom recovered by the probe. Meta-analysis reference points are exact (Table 3). “Artificial” is the synthetic single-cell model and “Omics” the single-cell data of Table 13.

A natural alternative is conditional probing (Hewitt et al., 2021) with baseline O instead of the model’s own non-contextual embedding layer: the gain $R_{O,H}^2 - R_O^2$ from adding H to a probe on O . Measuring this gain from the floor instead (in the population, $R_O^2 = \rho_O^2$) and normalizing it by the headroom gives $U_{H|O} := (R_{O,H}^2 - \rho_O^2)/(\rho_I^2 - \rho_O^2)$. In the population $U_{H|O} \leq U_{H|I} \leq 1$, with equality on the left when H retains everything in O relevant to T (Appendix L).

When an estimated floor is well below chance for classi-

fication or well below zero for regression, the headroom normalization is not applicable, and we report the raw probe performance instead. Floors only slightly below chance or zero, as expected from estimation noise when the population floor is near zero (Figure 7), are kept unclipped (Appendix B).

Evaluation settings. We specify the contrast (O, I, T) in settings with progressively weaker access to ground truth: the controlled meta-analysis and a synthetic single-cell model, where the reference points are exact (Appendix H), then real single-cell representations and published LLM probes. The contrast is chosen from the scientific question, independently of the probe scores, and normalized results are interpreted relative to it. Where no unique contrast exists, we evaluate nested baselines $O_1 \subset O_2 \subset \dots \subset O_k$. Appendix J applies this to a fine-tuned chemistry model and recorded neural activity.

4.2 Normalized probes across domains

4.2.1 Meta-analysis

Using the controlled random-effects model defined above, we take the sufficient statistic $\log S_1$ as the probe target, $T = \log \sum_{i=1}^K \frac{1}{\sigma_i^2 + \tau^2}$. The scientific question is whether the transformer’s hidden representation contains information about the precision-weighting computation beyond what the reported sampling variances alone predict. We therefore define the restricted information as $O = \sigma_{1:K}^2$ and the full information as $I = (\sigma_{1:K}^2, x_{1:K})$, where the observed effects x_i provide the additional evidence from which the latent heterogeneity τ^2 can be inferred. We also probe τ^2 directly. Because τ^2 is independent of O , its floor is zero, and its ceiling measures how much heterogeneity the reported effects reveal. The distribution shift changes the relative contribution of sampling variance and heterogeneity, so it tests whether raw decodability tracks the information available in each regime.

Figure 4 and Table 13 place each raw probe score between a floor and a ceiling set by a domain-specific contrast. In the **meta-analysis**, the contrast asks whether the hidden state carries heterogeneity inferred from the reported effects beyond what the sampling variances predict. In distribution, the probe recovers 78% of the headroom for $\log S_1$ and 84% for τ^2 . Under the shift, raw decodability falls for both targets ($0.64 \rightarrow 0.40$ for $\log S_1$, $0.64 \rightarrow 0.34$ for τ^2 ; Table 12), and so do the exact ceilings ($0.82 \rightarrow 0.47$ and $0.76 \rightarrow 0.40$; Table 3). The fraction of headroom recovered does not decrease ($U_H = 0.78 \rightarrow 0.82$ for $\log S_1$, $0.84 \rightarrow 0.85$ for τ^2 ; Figure 3). The lower raw scores therefore reflect less recoverable information in

the data, not a less informative representation (Section 4.3).

Conditional probing alone would suggest the opposite. With a linear decoder throughout, the conditional gain roughly halves under the shift ($0.58 \rightarrow 0.31$ for $\log S_1$; $0.64 \rightarrow 0.34$ for τ^2), which alone suggests a degraded representation. The headroom (estimated with linear decoder) also halves ($0.62 \rightarrow 0.33$; $0.66 \rightarrow 0.36$), and once the gain is divided by it, $U_{H|O}$ stays flat ($0.95 \rightarrow 0.95$; $0.97 \rightarrow 0.95$), corresponding to U_H (Appendix L).

4.2.2 Genomics and single-cell data

Single-cell RNA sequencing has the same structure as meta-analysis. Each example is a gene g measured in K cells. The normalized count of a cell varies for two reasons: technical noise, which decreases as the depth of the cell’s sequencing d_{gi} increases, and the biological variation from cell-to-cell, summarized by the overdispersion of the gene b_g . Under a Gamma–Poisson model the normalized count has variance $\lambda_g(d_{gi}^{-1} + b_g)$, where λ_g is the gene’s mean expression. Technical noise d_{gi}^{-1} therefore plays the role of the sampling variance σ_i^2 , and overdispersion b_g the role of heterogeneity τ^2 . The target is the analogue of $\log S_1$, the gene’s effective precision $T_g = \log \sum_{i=1}^K (d_{gi}^{-1} + b_g)^{-1}$, which depends on the hidden overdispersion. The restricted information is the sequencing depths alone, $O_g = d_{g,1:K}$, and the full information adds the observed counts, $I_g = (y_{g,1:K}, d_{g,1:K})$, from which overdispersion can be inferred. We ask whether frozen scGPT (Cui et al., 2024) encodes T_g beyond what depth alone predicts.

We first ask this on synthetic genes ($K = 40$; Appendix G) and then on four real scRNA-seq datasets (10x Genomics, 2021, 2024, 2022). In real data, the true b_g is unknown, so we split the cells of each dataset into half at random. Overdispersion is estimated from the first half and defines the target, while the depths, counts and scGPT inputs come from the second half, so the ceiling cannot simply recompute its own target. K is then the number of cells in the second half.

For **single-cell data**, raw scores are misleading because the floor differs across datasets. In the 10k human dataset, depth alone already reaches $R^2 = 0.60$, so scGPT’s raw score of 0.81 recovers 57% of the headroom. In the 20k donors dataset the raw score is much lower (0.57), but the floor is only 0.13, so scGPT recovers a similar 52%. Across the four datasets, raw scores range from 0.35 to 0.81, while the fraction of headroom recovered ranges from 0.28 to 0.57; on synthetic genes it is 0.70. scGPT therefore encodes biological overdispersion beyond sequencing depth, but only partially. scGPT probes use a 64-component PCA fit-

ted on training rows only; all floors, ceilings and other probes use the full feature vectors (Appendix I).

4.2.3 Published LLM probes

We compute floors for four published LLM probes, using descriptions of the input. T is the published target, I is the input the model received, and O is a surface description of it: n -grams of a name, statement, or conversation, or move summaries that never simulate Othello flips. The floor is what the decoder bank learns from O on the probe’s own training split, so it needs only the released inputs, labels, and splits. For Othello, the ceiling is exactly 1, since the board follows from the moves; elsewhere we set it to 1, which can only lower U_H , while a richer floor readout could lower it further. We therefore read U_H here as an approximate value and compare the raw scores with the floor directly where the headroom is small. For categorical targets, we use accuracy in place of R^2 . Figure 5 summarizes the results. See full results in Appendix K.

Space. Gurnee and Tegmark (2024) report that Llama-2 represents the coordinates of named places. A floor computed from a description of the place name, its character 2–4-gram counts, reaches $R^2 = 0.61$ for world places and 0.46 for US places, yet the published probes recover 63–77% of the headroom above it. For New York City this floor is 0.33 (95% CI 0.29–0.37), and the published scores (0.22 for 7B, 0.36 for 70B) lie within or below it, so they cannot separate a city-level map from borough names and numbers in the strings.

Othello. Li et al. (2022) report that Othello-GPT represents the board. A floor that knows which player placed each disc, but never simulates flips, reaches 0.80 accuracy (chance 0.46). The linear black/white/empty probe (0.74) falls below it ($U_H = -0.30$), so its accuracy is accounted for by disc placement. A non-linear probe recovers 45% of the headroom with our decoders. In the mine/theirs/empty encoding (Nanda et al., 2023), a linear probe reaches 0.99 and recovers 92% of the headroom above the floor from the same features (0.81 for this target), indicating that the effect of flips is represented linearly in the relative encoding.

Truth. Marks and Tegmark (2023) report that LLMs represent whether a statement is true along a linear direction, and that a probe trained on one dataset of statements still works on another. We compute the floor from a description of each statement, its character and word n -grams. Within a single dataset, this floor is near chance for most datasets but reaches 0.97–0.99 on **larger_than** and **smaller_than** datasets, whose truth labels follow from the wording of the numbers. Probes trained and tested on those two datasets therefore show little beyond the text. Transfer be-

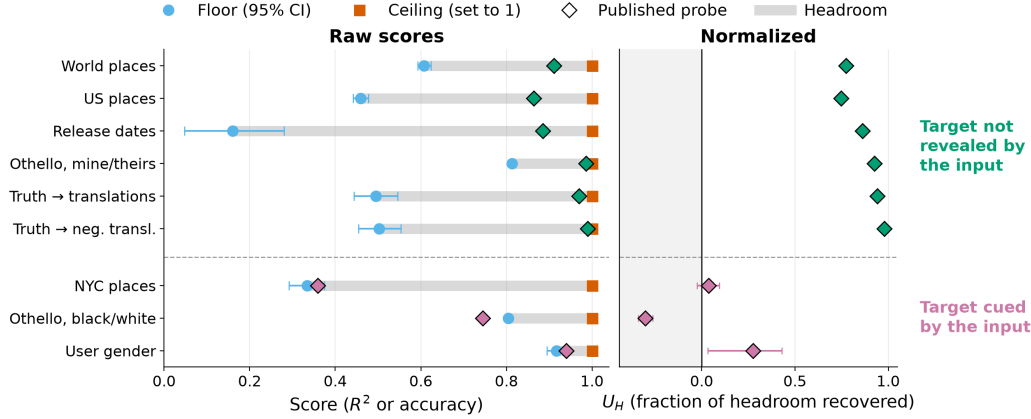


Figure 5: **Probe normalization puts published LLM probe scores in perspective.** Left: floor computed from a description of the input (n -grams; disc placement without flips for Othello), with 95% CI, the probe score, and the ceiling set to 1 (R^2 for places and dates, accuracy otherwise). Right: normalized probe score, U_H , with error bars from the floor’s CI.

tween datasets is more informative, because wording learned on one dataset does not carry over to another. A floor trained on the two number datasets and tested on Spanish–English translations is at chance (0.49), while the published probes reach 0.97 ($U_H = 0.94$). Trained on *cities* and *neg_cities* and tested on negated translations, the floor is again at chance (0.50), while the probes reach 0.96–0.99 (U_H up to 0.98).

User attributes. Chen et al. (2024) report probes that read a user’s age, gender, education and socioeconomic status from Llama-2-Chat activations. The floor, an n -gram classifier on the conversation, reaches 0.92–0.98, and every published reading probe (0.94–0.98) lies within its 95% interval; every control probe falls below its point estimate. The floor is as high from the user’s turns alone, and falls to 0.76–0.90 when trained on GPT-3.5 conversations and tested on Llama-2-Chat ones. However, the role-played conversations were written to reveal the attribute, so these accuracies can neither establish nor rule out that the model internally represents the user’s attributes.

Summary. Probe normalization puts these published scores in perspective (Figure 5). Probe scores reach well above the floor when the target is not revealed by the surface form: coordinates at world scale, the effect of flips on the board, and truth transferred across datasets. When the target is written or cued in the input, such as borough names and numbers for New York City places, disc placement for the absolute board encoding, or attributes the role-played user writes into the conversation, the floor alone accounts for much of the score.

4.3 Mechanisms for vanishing headroom

The headroom Δ is the predictability that the full information I adds beyond the restricted baseline O . It can vanish in two distinct ways (Corollary F.2). Let $T = f(O, Z)$ for a latent variable Z , and let $I = (O, X)$, where X is the additional evidence about Z . **(1) Weak target sensitivity:** T barely depends on Z . If f is Lipschitz in Z with constant $L(O)$, the headroom is bounded by $\mathbb{E}[L(O)^2 \text{Var}(Z | O)] / \text{Var}(T)$, so knowing Z adds little even when X reveals it perfectly. The floor rises toward the ceiling. **(2) Weak identification:** X reveals little about Z beyond O . If f varies by at most M over Z , the headroom is bounded by $M^2 \text{MI}(X; Z | O) / (2 \text{Var}(T))$, where MI is the conditional mutual information in nats, so T may depend strongly on Z , but the data cannot pin Z down. The ceiling falls toward the floor.

Which mechanism drives the meta-analysis shift? The distribution shift (Section 3) pushes in both directions. When sampling variance dominates, $\log S_1 = \log \sum_i (\sigma_i^2 + \tau^2)^{-1}$ depends less on τ^2 (weaker target sensitivity), and the per-study Fisher information about τ^2 , which scales as $(\sigma_i^2 + \tau^2)^{-2}$, drops, so the reported effects reveal less about τ^2 (weaker identification; Remark E.7). The exact reference points show that weak identification dominates (Table 3). The headroom for $\log S_1$ shrinks from 0.81 to 0.37. The floor rises only from 0.01 to 0.09, while the ceiling falls from 0.82 to 0.47, so most of the loss comes from the ceiling. For τ^2 itself the target is the latent variable, so target sensitivity cannot change; its floor stays at zero and its ceiling falls from 0.76 to 0.40, a pure identification loss (Figures 3 and 4).

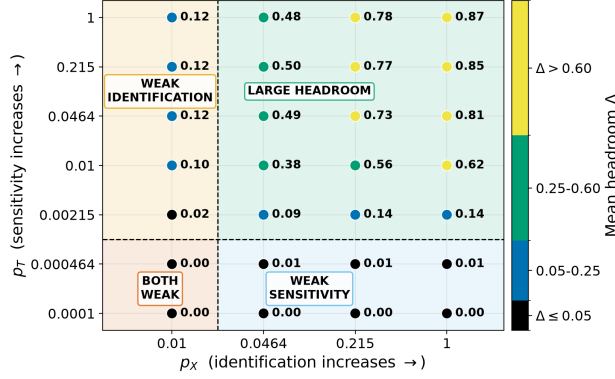


Figure 6: **Headroom as evidence quality and target sensitivity vary, in the synthetic single-cell model (with binomial thinning).** Evidence quality p_X controls how well the data identify the latent variable; target sensitivity p_T controls how strongly the target depends on it. Low p_X gives weak identification, low p_T gives weak target sensitivity, and low values of both give both.

Separating the mechanisms in a controlled grid.

In the meta-analysis both mechanisms move together. The synthetic single-cell model (Section 4.2.2; Appendix G) lets us vary them independently. The latent variable is the gene’s biological overdispersion, $Z = b_g$; the restricted information is sequencing depth, $O_g = d_{g,1:K}$; the additional evidence is the observed counts, $X_g = y_{g,1:K}$; and the target is T_g . Reducing sequencing depth by binomial thinning, retaining a fraction p of counts, lowers the effective depth from d_{gi} to pd_{gi} . In real data this affects both the evidence and the target at once. In simulation we apply it separately (Appendix G.1): thinning the counts used as evidence (p_X) degrades inference of b_g , giving weak identification, whereas thinning the depths entering the target (p_T) lets technical variance dominate overdispersion, giving weak target sensitivity; the restricted information is never thinned. Figure 6 shows the resulting headroom over the (p_X, p_T) grid, with estimated reference points. Headroom vanishes when either parameter is small and stays large only when both are large, so a loss of headroom can be traced to its cause.

5 CONCLUSION

A probe score of 0.6 has no fixed meaning. What it says about a representation depends on two quantities: how much of that predictability a declared baseline already provides, and how much the available data support in total. We made both explicit. **The floor ρ_O^2 is what a declared baseline O already predicts,**

the ceiling ρ_I^2 is what the full input I supports, and the headroom $\Delta = \rho_I^2 - \rho_O^2$ between them is the range in which a probe score can show information beyond O . Headroom vanishes in two ways: weak target sensitivity, when the target stops depending on the latent variable, and weak identification, when the evidence stops revealing it.

In the controlled meta-analysis, where the reference points are exact, normalization reverses the reading of a distribution shift. Raw decodability of $\log S_1$ and τ^2 falls and prediction error rises, which alone suggests a degraded representation. The reference points show instead that the data lost information: the ceiling falls from 0.82 to 0.47 for $\log S_1$ and from 0.76 to 0.40 for τ^2 , while the floor for $\log S_1$ rises only from 0.01 to 0.09, so weak identification dominates. The representation recovers a similar share of the smaller headroom ($U_H = 0.78 \rightarrow 0.82$ for $\log S_1$; $0.84 \rightarrow 0.85$ for τ^2).

The estimation recovers the exact reference points in two fully specified generative models and extends to settings where only samples are available. In frozen scGPT, similar raw scores correspond to different fractions of recovered headroom across datasets (0.28–0.57), so biological overdispersion is encoded only partially. For published LLM probes, a floor computed from the input text alone separates claims the text cannot explain (world-scale coordinates, the effect of flips in the relative Othello encoding, truth transferred across datasets) from claims it largely explains (New York City locations, the absolute Othello board, user attributes in role-played conversations).

The framework complements causal interventions: it bounds how much a correlational score can show before any intervention is run. **We recommend that interpretability studies report the floor, the ceiling and the fraction of headroom recovered alongside raw scores, state the contrast (O, I, T) before looking at probe scores, and, where no unique baseline exists, report a nested sequence $O_1 \subset \dots \subset O_k$ so that readers can see how the choice of baseline affects the conclusions.**

Limitations. Normalized scores are defined relative to a stated contrast (O, I, T) . Different contrasts answer different scientific questions, so we report nested baselines (Appendix J) rather than a single choice. On observational data the reference points are estimated, and an underestimated ceiling or floor inflates U_H . For published LLM probes the ceiling cannot be estimated and is set to 1, so U_H there is approximate. Our controlled tasks are kept simple so that the reference points can be computed exactly, and the transformers trained on them are small.

AI use statement

In this work, we used generative AI tools to aid and polish writing, for retrieval and discovery of related work, for research execution, and to draft sections of the paper. As part of research execution, we used coding agents to assist with implementing and debugging analysis code and with producing scientific figures. We did not use generative AI tools for generating synthetic datasets or for proving mathematical claims. We have reviewed all AI-assisted work. AI-assisted code and analyses were inspected and their outputs were checked against the corresponding experiments and results; AI-assisted text and figures were reviewed for consistency with the underlying methods, analyses, and findings. We take responsibility for the final content of this work, including text, claims, code, figures, and other artifacts produced with the aid of generative AI.

References

- 10x Genomics (2021). 10k PBMCs from Human Female, 3' v3.1, Chromium X. 10x Genomics.
- 10x Genomics (2022). 40k mixture of cells dissociated from 4 fixed tumor tissues, multiplexed samples, 4 probe barcodes (next gen). 10x Genomics Dataset.
- 10x Genomics (2024). 20k pbmc from human pbmc multiplex sample. 10x Genomics Dataset.
- Ahlmann-Eltze, C., Huber, W., and Anders, S. (2025). Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8):1657–1661.
- Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. (2022). What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*.
- Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.
- Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. (2023). Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, page 57125–57211. Neural Information Processing Systems Foundation, Inc. (NeurIPS).
- Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.
- Boiarsky, R., Singh, N. M., Buendia, A., Amini, A. P., Getz, G., and Sontag, D. (2024). Deeper evaluation of a single-cell foundation model. *Nature Machine Intelligence*, 6(12):1443–1446.
- Chen, Y., Wu, A., DePodesta, T., Yeh, C., Li, K., Marin, N. C., Patel, O., Riecke, J., Raval, S., Seow, O., et al. (2024). Designing a dashboard for transparency and control of conversational ai. *arXiv preprint arXiv:2406.07882*.
- Chowdhury, R. H., Glaser, J. I., and Miller, L. E. (2022). Data from: Area 2 of primary somatosensory cortex encodes kinematics of the whole arm. Dryad Dataset. Originally published 2020; dataset updated May 20, 2022.
- Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. (2024). scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480.
- de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., and Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *The Journal of Neuroscience*, 37(27):6539–6557.
- Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. (2021). Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.
- Friedman, D., Lampinen, A., Dixon, L., Chen, D., and Ghandeharioun, A. (2023). Interpretability illusions in the generalization of simplified models. *arXiv preprint arXiv:2312.03656*.
- Garg, S., Tsipras, D., Liang, P., and Valiant, G. (2022). What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, page 30583–30598. Neural Information Processing Systems Foundation, Inc. (NeurIPS).
- Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. (2024). Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR.
- Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., and Kording, K. P. (2020). Machine learning for neural decoding. *eneuro*, 7(4):ENEURO–0506.
- Grinsztajn, L., Flöge, K., Key, O., Birkel, F., Jund, P., Roof, B., Manium, M., Hoo, S. B., Bühler, M., Garg, A., et al. (2026). Tabpfn-3: Technical report. *arXiv preprint arXiv:2605.13986*.
- Gurnee, W. and Tegmark, M. (2024). Language models represent space and time. In *International Conference on Learning Representations*, volume 2024, pages 2483–2503.
- Hebart, M. N. and Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *Neuroimage*, 180:4–18.

- Hewitt, J., Ethayarajh, K., Liang, P., and Manning, C. D. (2021). Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639.
- Hewitt, J. and Liang, P. (2019). Designing and interpreting probes with control tasks. 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)*, pages 2733–2743.
- Higgins, J. P. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, 21(11):1539–1558.
- Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326.
- Hsu, A., Borst, A., and Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation in Neural Systems*, 15(2):91–109.
- Kedzierska, K. Z., Crawford, L., Amini, A. P., and Lu, A. X. (2025). Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1).
- Kendiukhov, I. (2026). Systematic evaluation of single-cell foundation model interpretability reveals attention captures co-expression rather than unique regulatory signal.
- Kriegeskorte, N. and Douglas, P. K. (2019). Interpreting encoding and decoding models. *Current opinion in neurobiology*, 55:167–179.
- Kumar, A., Tan, C., and Sharma, A. (2022). Probing classifiers are unreliable for concept removal and detection. *Advances in Neural Information Processing Systems*, 35:17994–18008.
- Lage-Castellanos, A., Valente, G., Formisano, E., and De Martino, F. (2019). Methods for computing the maximum performance of computational models of fmri responses. *PLOS Computational Biology*, 15(3):e1006397.
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. (2022). Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.
- Marks, S. and Tegmark, M. (2023). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.
- Mikulik, V., Delétang, G., McGrath, T., Genewein, T., Martic, M., Legg, S., and Ortega, P. (2020). Meta-trained agents implement bayes-optimal agents. *Advances in neural information processing systems*, 33:18691–18703.
- Müller, S., Hollmann, N., Arango, S. P., Grabocka, J., and Hutter, F. (2021). Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*.
- Nanda, N., Lee, A., and Wattenberg, M. (2023). Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30.
- Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410.
- Pimentel, T., Valvoda, J., Maudslay, R. H., Zmigrod, R., Williams, A., and Cotterell, R. (2020). Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.
- Quian Quiroga, R. and Panzeri, S. (2009). Extracting information from neuronal populations: information theory and decoding approaches. *Nature Reviews Neuroscience*, 10(3):173–185.
- Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7237–7256. Association for Computational Linguistics.
- Ravichander, A., Belinkov, Y., and Hovy, E. (2021). Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377.
- Ritchie, J. B., Kaplan, D. M., and Klein, C. (2019). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British journal for the philosophy of science*.
- Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15).
- Saphra, N. and Lopez, A. (2019). Understanding learning dynamics of language models with svcca.

- In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3257–3267.
- Schoppe, O., Harper, N. S., Willmore, B. D. B., King, A. J., and Schnupp, J. W. H. (2016). Measuring the performance of neural models. *Frontiers in Computational Neuroscience*, 10.
- Snoek, L., Miletić, S., and Scholte, H. S. (2019). How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage*, 184:741–760.
- Teh, A., Jabbour, M., and Polyanskiy, Y. (2025). Solving empirical bayes via transformers. *arXiv preprint arXiv:2502.09844*.
- Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601.
- Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., and Ellinor, P. T. (2023). Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624.
- Voita, E. and Titov, I. (2020). Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.
- Wieting, J. and Kiela, D. (2019). No training required: Exploring random encoders for sentence classification. *arXiv preprint arXiv:1901.10444*.
- Xu, Y., Zhao, S., Song, J., Stewart, R., and Ermon, S. (2020). A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*.
- Zhang, K. and Bowman, S. R. (2018). Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pages 359–361.

Supplementary Materials

A CONTROLLED RANDOM-EFFECTS META-ANALYSIS

Generative model. Each example is a meta-analysis containing K studies that attempt to estimate the same underlying effect. We denote this effect by

$$\mu \sim \mathcal{N}(0, 1).$$

The transformer models are trained to predict μ . The individual studies need not have exactly the same true effect. Their study-specific effects are

$$\theta_i \sim \mathcal{N}(\mu, \tau^2),$$

where

$$\tau^2 \sim \text{U}(0, 0.5)$$

is the between-study heterogeneity. Large τ^2 means that the true effects θ_i vary substantially from study to study even before sampling noise is introduced; small τ^2 means that the studies are close to sharing a common true effect. Each study also has finite-sample measurement uncertainty. We draw

$$n_i \sim \text{U}\{5, \dots, 99\}, \quad s_i^2 \sim \text{U}(0.1, 10),$$

and define its sampling variance as

$$\sigma_i^2 = \frac{s_i^2}{n_i}.$$

Thus studies with larger samples or smaller within-study variance are measured more precisely. Finally, the reported study estimate is

$$x_i \sim \mathcal{N}(\theta_i, \sigma_i^2).$$

The difference between x_i and the global effect μ therefore arises from two distinct sources:

$$\underbrace{\theta_i - \mu}_{\text{between-study heterogeneity}} \quad \text{and} \quad \underbrace{x_i - \theta_i}_{\text{sampling noise}}.$$

Marginalizing over the unobserved study-specific effect θ_i gives

$$x_i \mid \mu, \tau^2, \sigma_i^2 \sim \mathcal{N}(\mu, \sigma_i^2 + \tau^2).$$

The total uncertainty of study i is therefore the sum of its sampling variance and the common between-study heterogeneity. The transformer receives $(x_i, \log \sigma_i^2)$ for each study as an input. It does *not* observe θ_i , τ^2 , or μ . We hypothesize that it infers whatever information about τ^2 is useful for combining the reported studies.

Precision weighting. A study with lower total variance $\sigma_i^2 + \tau^2$ should contribute more strongly to an estimate of μ . Its corresponding precision is the inverse of that variance,

$$w_i = \frac{1}{\sigma_i^2 + \tau^2}.$$

Thus a large w_i denotes a comparatively informative study, while a small w_i denotes a noisy or unreliable one. If τ^2 were known, the K studies could be summarized by two quantities:

$$S_1 = \sum_{i=1}^K w_i, \quad S_2 = \sum_{i=1}^K w_i x_i.$$

The first, S_1 , is the *total precision*: it measures how much effective information the collection of studies provides about μ . The second, S_2 , is the precision-weighted sum of the observed effects. For fixed τ^2 , the likelihood for μ can be written as

$$p(x_{1:K} \mid \mu, \tau^2, \sigma_{1:K}^2) \propto \exp \left[-\frac{1}{2} (S_1 \mu^2 - 2S_2 \mu) \right].$$

Consequently, once τ^2 is known, the individual study observations contain no additional information about μ beyond (S_1, S_2) . Therefore, (S_1, S_2) are sufficient statistics for estimating μ conditional on τ^2 . This is why $\log S_1$ is useful as a representation-level audit target: constructing it requires the model to combine the observed sampling variances with information about the otherwise hidden heterogeneity τ^2 .

Inverse-variance estimate. The standard inverse-variance-weighted estimator is

$$\hat{\mu}_{\text{IVW}} = \frac{S_2}{S_1}.$$

This denotes a weighted average of the reported study effects, where more precise studies receive larger weights. For example, if two studies have identical reported effects but one has much smaller total variance, the latter contributes more strongly to $\hat{\mu}_{\text{IVW}}$. The heterogeneity τ^2 plays an important role: increasing heterogeneity makes the study weights more similar because the common τ^2 term begins to dominate their individual sampling variances.

Bayesian estimate. The generative model places the prior $\mu \sim \mathcal{N}(0, 1)$, whose precision is 1. Combining this prior with the conditional likelihood gives

$$\mu \mid x_{1:K}, \sigma_{1:K}^2, \tau^2 \sim \mathcal{N} \left(\frac{S_2}{S_1 + 1}, \frac{1}{S_1 + 1} \right).$$

The corresponding posterior mean is

$$\hat{\mu}_{\text{B}}(\tau^2) = \frac{S_2}{S_1 + 1}.$$

As the prior on μ is centered at zero, the Bayesian estimate is the inverse-variance estimate multiplied by $S_1/(S_1 + 1) < 1$. When the accumulated study precision S_1 is small, the prior has appreciable influence and the estimate is pulled toward zero. As S_1 becomes large, $\frac{S_1}{S_1 + 1} \rightarrow 1$, and the Bayesian and classical estimators become nearly identical.

Their corresponding conditional mean-squared errors are

$$\text{MSE}_{\text{B}} = \frac{1}{S_1 + 1}, \quad \text{MSE}_{\text{IVW}} = \frac{1}{S_1}.$$

The difference is therefore largest when little total precision has accumulated, for example at short sequence prefixes, and becomes negligible when many precise studies are available.

The expression $\frac{S_2}{S_1 + 1}$ requires the true τ^2 , because both S_1 and S_2 use $w_i = (\sigma_i^2 + \tau^2)^{-1}$. As the transformer is never given τ^2 , $S_2/(S_1 + 1)$ is *not* the Bayes-optimal predictor directly available from the model’s observed inputs. The actual Bayes-optimal predictor must instead average over uncertainty about the unknown heterogeneity: $\hat{\mu}_{\text{Bayes}} = \mathbb{E}[\mu \mid x_{1:K}, \sigma_{1:K}^2]$. Equivalently, $\hat{\mu}_{\text{Bayes}} = \int \frac{S_2(t)}{S_1(t) + 1} p(t \mid x_{1:K}, \sigma_{1:K}^2) dt$, where t denotes a candidate value of the heterogeneity, $S_1(t) = \sum_{i=1}^K \frac{1}{\sigma_i^2 + t}$, and $S_2(t) = \sum_{i=1}^K \frac{x_i}{\sigma_i^2 + t}$.

The conditional quantity $\frac{S_2}{S_1 + 1}$ is nevertheless useful because the simulator knows the true τ^2 . We therefore use it only as an *oracle conditional-Bayes reference*: it tells us how well prediction could perform if the hidden heterogeneity were known. It is not supplied to the transformer and is not the model’s available Bayes-optimal solution. See Appendix C for the full derivation.

A.1 Distribution shift

The total variance of study i is $\sigma_i^2 + \tau^2$. We summarize which source of uncertainty dominates using the sampling-variance share

$$r_i = \frac{\sigma_i^2}{\sigma_i^2 + \tau^2}.$$

Model size	d	h	L	Parameters	Peak learning rate	Epoch budget
xs	16	2	1	3,345	4×10^{-4}	3,300
s	32	4	2	25,537	4×10^{-4}	3,300
m	64	4	2	100,225	5×10^{-5}	3,000
l	64	4	4	200,193	5×10^{-5}	3,500
xl	128	8	4	793,601	5×10^{-5}	3,500

Table 1: Transformer model configurations. d is the residual-stream width, h the number of attention heads (head width d/h), and L the number of transformer blocks; the block MLP has hidden width $4d$. Parameter counts include the input embedding, all blocks, and the scalar readout.

This quantity lies between zero and one. When $r_i \approx 0$, most uncertainty comes from between-study heterogeneity. When $r_i \approx 1$, most uncertainty comes from ordinary study-level sampling variance.

In the in-distribution regime, the median value is approximately $r_{\text{ID}} = 0.31$, so heterogeneity is usually the dominant contribution to total variance. In the OOD regime, which draws $s_i^2 \sim \text{U}(10, 20)$, $n_i \sim \text{U}\{2, \dots, 19\}$ and $\tau^2 \sim \text{U}(0, 1.5)$ with all other draws as above, the median increases to $r_{\text{OOD}} = 0.71$, so sampling variance becomes the dominant contribution.

The distribution shift therefore changes the relative source of uncertainty governing the correct inverse-variance weights. A model that has learned a regime-specific shortcut, accurate only when τ^2 dominates, can consequently fail when σ_i^2 becomes dominant, whereas a model implementing the appropriate weighting computation should adapt its weights to the changed balance.

A.2 Heterogeneity baseline

The simulator knows τ^2 ; the transformer does not. A classical baseline for estimating heterogeneity from the observed study estimates is the DerSimonian–Laird estimator,

$$\hat{\tau}_{\text{DL}}^2 = \max \left\{ 0, \frac{Q - (K - 1)}{C} \right\},$$

where Q measures excess dispersion of the study estimates around their fixed-effect weighted mean and C is a normalization determined by the fixed-effect study weights.

Since $\hat{\tau}_{\text{DL}}^2$ is computable from the model inputs, a probe that predicts τ^2 may simply reproduce information already available through this standard estimator. The full-information features used for the estimated reference points therefore include $\hat{\tau}_{\text{DL}}^2$ (Appendix I).

A.3 Training and checkpoints

Models are causal pre-norm transformers at five sizes, from 3.3k to 794k parameters (Table 1), trained with mean-squared error at every prefix. Checkpoints are retained at epochs $\{0, 1, 2, 5, 10, 20, 50, 100, \dots\}$ for training-trajectory analyses. Training stops when validation MSE, averaged over all prefixes, is within 2% of the oracle conditional-Bayes risk averaged the same way (Appendix D), or after 300 epochs without validation improvement; the oracle risk uses the true τ^2 and is not attainable from the model’s inputs. Of the 25 random-effects runs, 24 stopped on patience. Model predictions are closer to the oracle conditional posterior mean than to the oracle IVW estimator. Behavioral comparisons with the IVW and conditional-Bayes estimators are performed on held-out meta-analyses.

B DECODER BANK

All reported regression-probe scores are computed using one canonical bank (Table 2) of candidate decoders. For the controlled meta-analysis, all probes and reference points are evaluated at the final position, using the final-token representation. Decoder type and hyperparameters are selected using the selection split based on R^2 , after which the selected decoder is evaluated once on the evaluation split. Every probe uses disjoint training,

selection, and evaluation rows. Categorical decoders are chosen by McFadden R^2 , but reported and normalized into U_H , by accuracy. We report standard R^2 on this held-out split and retain negative values rather than clipping them to zero.

$$\mathcal{M} = \{\text{GBM, shallow GBM, MLP, RidgeCV}\}.$$

Decoder	Fixed specification
GBM	HistGradientBoostingRegressor with 400 maximum iterations, learning rate 0.06, unrestricted depth, early stopping, and a 0.15 internal validation fraction.
Shallow GBM	HistGradientBoostingRegressor with 600 maximum iterations, learning rate 0.04, maximum depth 3, early stopping, and a 0.15 internal validation fraction.
MLP	Standardization followed by an MLP with hidden widths (256, 128), L_2 penalty 10^{-4} , 600 maximum iterations, early stopping, and patience 25.
Linear	Standardization followed by RidgeCV with $\alpha \in \text{logspace}(-6, 4, 21)$.

Table 2: The decoder bank. These definitions are fixed across domains, representations, targets, and random seeds.

C BAYES POSTERIOR MEAN CONDITIONAL ON HETEROGENEITY

Consider a prefix containing the first k studies. Conditional on the overall effect μ , the heterogeneity τ^2 , and the sampling variances $\sigma_{1:k}^2$, the reported effects are independent and satisfy $x_i \mid \mu, \tau^2, \sigma_i^2 \sim \mathcal{N}(\mu, \sigma_i^2 + \tau^2)$. Define the precision weights and their cumulative sufficient statistics by

$$w_i = \frac{1}{\sigma_i^2 + \tau^2}, \quad S_1^{(k)} = \sum_{i=1}^k w_i, \quad S_2^{(k)} = \sum_{i=1}^k w_i x_i.$$

Ignoring factors that do not depend on μ , the likelihood is

$$p(x_{1:k} \mid \mu, \tau^2, \sigma_{1:k}^2) \propto \exp \left[-\frac{1}{2} \sum_{i=1}^k w_i (x_i - \mu)^2 \right] \propto \exp \left[-\frac{1}{2} \left(S_1^{(k)} \mu^2 - 2S_2^{(k)} \mu \right) \right].$$

Combining this likelihood with the prior $\mu \sim \mathcal{N}(0, 1)$ gives

$$\begin{aligned} p(\mu \mid x_{1:k}, \tau^2, \sigma_{1:k}^2) &\propto p(x_{1:k} \mid \mu, \tau^2, \sigma_{1:k}^2) p(\mu) \\ &\propto \exp \left[-\frac{1}{2} \left(S_1^{(k)} \mu^2 - 2S_2^{(k)} \mu \right) \right] \exp \left(-\frac{1}{2} \mu^2 \right) \\ &= \exp \left[-\frac{1}{2} \left((S_1^{(k)} + 1) \mu^2 - 2S_2^{(k)} \mu \right) \right] \\ &\propto \exp \left[-\frac{1}{2} \left(S_1^{(k)} + 1 \right) \left(\mu - \frac{S_2^{(k)}}{S_1^{(k)} + 1} \right)^2 \right]. \end{aligned}$$

Therefore,

$$\mu \mid x_{1:k}, \tau^2, \sigma_{1:k}^2 \sim \mathcal{N} \left(\frac{S_2^{(k)}}{S_1^{(k)} + 1}, \frac{1}{S_1^{(k)} + 1} \right).$$

Under squared-error loss, the posterior mean uniquely minimizes the conditional expected loss. Hence, when τ^2 is known, the Bayes estimator is

$$\hat{\mu}_B^{(k)}(\tau^2) = \arg \min_a \mathbb{E}[(a - \mu)^2 \mid x_{1:k}, \tau^2, \sigma_{1:k}^2] = \mathbb{E}[\mu \mid x_{1:k}, \tau^2, \sigma_{1:k}^2] = \frac{S_2^{(k)}}{S_1^{(k)} + 1}.$$

In the actual prediction task, τ^2 is not observed. The fully Bayes-optimal estimator therefore averages this conditional posterior mean over the posterior distribution of τ^2 :

$$\hat{\mu}_{\text{Bayes}}^{(k)} = \mathbb{E}[\mu \mid x_{1:k}, \sigma_{1:k}^2] = \int_0^{0.5} \frac{S_2^{(k)}(t)}{S_1^{(k)}(t) + 1} p(t \mid x_{1:k}, \sigma_{1:k}^2) dt,$$

where $S_1^{(k)}(t) = \sum_{i=1}^k (\sigma_i^2 + t)^{-1}$ and $S_2^{(k)}(t) = \sum_{i=1}^k x_i (\sigma_i^2 + t)^{-1}$. Thus, $S_2^{(k)}/(S_1^{(k)} + 1)$ is an oracle Bayes reference that assumes access to the true heterogeneity, whereas the model must approximate the marginal posterior mean by inferring τ^2 from the observed studies. We do not compute $\hat{\mu}_{\text{Bayes}}$; all behavioral references in this paper are oracle quantities.

D EXPECTED MEAN-SQUARED ERROR OF THE IVW AND BAYES ESTIMATORS

We derive the expected mean-squared error of the oracle IVW and oracle conditional-Bayes estimators (both using the true τ^2) derived in Appendix C. At a prefix containing the first k studies, define $w_i = 1/(\sigma_i^2 + \tau^2)$, $S_1^{(k)} = \sum_{i=1}^k w_i$, and $S_2^{(k)} = \sum_{i=1}^k w_i x_i$. Conditional on μ , τ^2 , and the reported sampling variances $\sigma_{1:k}^2$, each observation satisfies $x_i \sim \mathcal{N}(\mu, 1/w_i)$, so $S_2^{(k)} \sim \mathcal{N}(S_1^{(k)} \mu, S_1^{(k)})$. Hence, the weighted sum can be written as

$$S_2^{(k)} = S_1^{(k)} \mu + \varepsilon_k, \quad \varepsilon_k \mid S_1^{(k)} \sim \mathcal{N}(0, S_1^{(k)}).$$

The inverse-variance-weighted estimator is

$$\hat{\mu}_{\text{IVW}}^{(k)} := \frac{S_2^{(k)}}{S_1^{(k)}} = \mu + \frac{\varepsilon_k}{S_1^{(k)}}.$$

Its estimation error is therefore $\varepsilon_k/S_1^{(k)}$, and its conditional mean-squared error is

$$\begin{aligned} \mathbb{E} \left[\left(\hat{\mu}_{\text{IVW}}^{(k)} - \mu \right)^2 \mid S_1^{(k)} \right] &= \mathbb{E} \left[\left(\frac{\varepsilon_k}{S_1^{(k)}} \right)^2 \mid S_1^{(k)} \right] \\ &= \frac{1}{(S_1^{(k)})^2} \mathbb{E} [\varepsilon_k^2 \mid S_1^{(k)}] \\ &= \frac{\text{Var}(\varepsilon_k \mid S_1^{(k)})}{(S_1^{(k)})^2} \\ &= \frac{S_1^{(k)}}{(S_1^{(k)})^2} = \frac{1}{S_1^{(k)}}. \end{aligned}$$

For the prior $\mu \sim \mathcal{N}(0, 1)$, the Bayes posterior mean (from Appendix C) is

$$\hat{\mu}_{\text{B}}^{(k)} = \frac{S_2^{(k)}}{S_1^{(k)} + 1}.$$

Substituting $S_2^{(k)} = S_1^{(k)} \mu + \varepsilon_k$ gives

$$\hat{\mu}_{\text{B}}^{(k)} - \mu = \frac{\varepsilon_k - \mu}{S_1^{(k)} + 1}.$$

The variables ε_k and μ are independent, with variances $S_1^{(k)}$ and 1, respectively. Averaging over both the observation noise and the prior on μ therefore gives

$$\mathbb{E} \left[\left(\hat{\mu}_{\text{B}}^{(k)} - \mu \right)^2 \mid S_1^{(k)} \right] = \frac{\text{Var}(\varepsilon_k) + \text{Var}(\mu)}{(S_1^{(k)} + 1)^2} = \frac{S_1^{(k)} + 1}{(S_1^{(k)} + 1)^2} = \frac{1}{S_1^{(k)} + 1}.$$

Thus, at every prefix,

$$\text{MSE}_B^{(k)} = \frac{1}{S_1^{(k)} + 1} < \frac{1}{S_1^{(k)}} = \text{MSE}_{\text{IVW}}^{(k)}.$$

The absolute improvement is

$$\text{MSE}_{\text{IVW}}^{(k)} - \text{MSE}_B^{(k)} = \frac{1}{S_1^{(k)}(S_1^{(k)} + 1)},$$

while the ratio of the Bayes risk to the IVW risk is

$$\frac{\text{MSE}_B^{(k)}}{\text{MSE}_{\text{IVW}}^{(k)}} = \frac{S_1^{(k)}}{S_1^{(k)} + 1}.$$

The gain is therefore largest when only a few studies have been observed and $S_1^{(k)}$ is small, and it vanishes as the accumulated precision grows. To match the model’s all-position training objective, we average these conditional risks over every prefix and every generated meta-analysis:

$$\overline{\text{MSE}}_{\text{IVW}} = \mathbb{E}_{\tau^2, \sigma_{1:K}^2, k} \left[\frac{1}{S_1^{(k)}} \right], \quad \overline{\text{MSE}}_B = \mathbb{E}_{\tau^2, \sigma_{1:K}^2, k} \left[\frac{1}{S_1^{(k)} + 1} \right].$$

Evaluating these expectations on the generated evaluation data gives oracle risks $\overline{\text{MSE}}_{\text{IVW}} = 0.048$ and $\overline{\text{MSE}}_B = 0.041$. The conditional-Bayes risk lower-bounds the risk attainable without τ^2 ; the IVW risk does not, because the IVW estimator ignores the prior on μ .

E FLOOR–CEILING DERIVATIONS AND REFERENCE-POINT COMPUTATION

This appendix derives the floor–ceiling normalization used in Section 4.1. We first specialize the argument to the probe target $\log S_1$, then state the general headroom identity and the two mechanisms by which headroom vanishes, and finally describe the exact computation of the reference points in Appendix H.

Task-specific definitions and bounds. One meta-analysis contains the reported effects and sampling variances

$$X := x_{1:K}, \quad O := \sigma_{1:K}^2, \quad I := (O, X).$$

Thus, O contains the directly reported variances, X contains the additional evidence carried by the reported effects, and I contains all observed information. The latent variables (μ, τ^2) are drawn independently of O , with

$$\tau^2 \sim \text{U}(0, \tau_{\max}), \quad \mu \sim \mathcal{N}(0, 1),$$

with $\tau_{\max} = 0.5$ in distribution and $\tau_{\max} = 1.5$ out of distribution (Appendix A). At prefix k , the probe target is

$$T_k := \log S_1^{(k)} = \log \sum_{i \leq k} \frac{1}{\sigma_i^2 + \tau^2}.$$

The target is therefore a deterministic function of (O, τ^2) . Conditional on O , the reported effects X are relevant to T_k only through what they reveal about the latent heterogeneity τ^2 . Throughout, we write

$$\text{MMSE}(T \mid I) := \mathbb{E}[\text{Var}(T \mid I)].$$

Proposition E.1 (Probe ceiling). *Let $H = \Phi(I)$ be any representation constructed from the observed inputs, whether produced by a trained or untrained architecture, and let h be any probe. The population R^2 of $h(H)$ for a target T satisfies*

$$R^2 \leq \rho_I^2 := 1 - \frac{\mathbb{E}[\text{Var}(T \mid I)]}{\text{Var}(T)} = 1 - \frac{\text{MMSE}(T \mid I)}{\text{Var}(T)}.$$

Naturally, $\rho_I^2 \leq 1$.

Proof. Because $H = \Phi(I)$, the prediction $h(H)$ cannot use information beyond I . The minimum mean-squared error attainable from I is achieved by $\mathbb{E}[T \mid I]$. Therefore,

$$\text{MSE}(h(H), T) \geq \text{MMSE}(T \mid I),$$

and substitution into the definition of population R^2 gives the result. $\square$

Proposition E.2 (Floor and headroom). *Define*

$$\rho_O^2 := 1 - \frac{\mathbb{E}[\text{Var}(T \mid O)]}{\text{Var}(T)} = 1 - \frac{\text{MMSE}(T \mid O)}{\text{Var}(T)}.$$

This is the population R^2 attained by the optimal predictor using only the reported sampling variances O and the prior on τ^2 . In particular, it performs no inference about heterogeneity from the reported effects X . Every O -measurable predictor satisfies

$$R^2 \leq \rho_O^2.$$

For any predictor based on the complete observations,

$$R^2 \leq \rho_O^2 + \Delta, \quad \Delta := \rho_I^2 - \rho_O^2 \geq 0.$$

Since $O \subseteq I$, conditioning on I cannot increase the Bayes MMSE, so $\rho_O^2 \leq \rho_I^2 \leq 1$ and therefore $\Delta \geq 0$.

Proof. The conditional expectation $\mathbb{E}[T \mid O]$ minimizes MSE among all O -measurable predictors, giving the first inequality. The second follows from

$$\rho_O^2 + \Delta = \rho_I^2$$

and Proposition E.1. $\square$

Remark E.3 (Evidential meaning of the floor and headroom). The quantity ρ_O^2 is called a floor because it is a floor for evidential interpretation, not because every probe must attain it. A probe must exceed this threshold before its score can provide evidence of information derived from X rather than O alone.

The quantity Δ is the maximum amount by which a predictor can outperform the optimal predictor that ignores the information about τ^2 contained in X . Consequently, only

$$\max\{0, R^2 - \rho_O^2\},$$

which is at most Δ , can provide evidence that the representation encodes heterogeneity inferred from the reported effects. Since T is measurable with respect to (O, τ^2) , the only relevance of X to T given O is through what X reveals about τ^2 .

Proposition E.4 (Regime dependence of the headroom). *When sampling variance dominates, T_k is almost entirely predictable from the observed variances alone, whereas when heterogeneity dominates, predicting T_k increasingly requires information about the latent τ^2 . Let*

$$w_i := \frac{1}{\sigma_i^2 + \tau^2}, \quad S_1^{(k)} := \sum_{i \leq k} w_i, \quad T_k := \log S_1^{(k)}.$$

(i) Sampling-variance-dominated regime ($\sigma_i^2 \gg \tau^2$). *For every $\tau^2 \geq 0$,*

$$\left| \frac{\partial T_k}{\partial \tau^2} \right| \leq (\min_i \sigma_i^2)^{-1}.$$

It follows that

$$\mathbb{E}[\text{Var}(T_k \mid O)] \leq \frac{\tau_{\max}^2}{12} \mathbb{E}[(\min_i \sigma_i^2)^{-2}].$$

Therefore, along any sequence of regimes for which

$$\frac{\tau_{\max}^2 \mathbb{E}[(\min_i \sigma_i^2)^{-2}]}{\text{Var}(T_k)} \rightarrow 0,$$

we have

$$\rho_O^2 \rightarrow 1, \quad \Delta \rightarrow 0, \quad \rho_I^2 \rightarrow 1.$$

(ii) Heterogeneity-dominated regime ($\tau^2 \gg \sigma_i^2$). At the final position K ,

$$\frac{K}{\tau^2 + \max_i \sigma_i^2} \leq S_1^{(K)} \leq \frac{K}{\tau^2}.$$

Hence,

$$|T_K - (\log K - \log \tau^2)| \leq \log \left(1 + \frac{\max_i \sigma_i^2}{\tau^2} \right).$$

Consider a sequence of regimes in which

$$\frac{\max_i \sigma_i^2}{\tau^2} \rightarrow 0$$

in probability, and suppose that the squared upper bounds

$$\left\{ \log^2 \left(1 + \frac{\max_i \sigma_i^2}{\tau^2} \right) \right\}$$

are uniformly integrable. Then

$$T_K - (\log K - \log \tau^2) \rightarrow 0$$

in L^2 . If $\text{Var}(T_K)$ remains bounded away from zero, it follows that

$$\rho_O^2 \rightarrow 0.$$

Thus, in this limit, nearly all predictable variance in T_K requires information about the latent heterogeneity.

Proof. For part (i), by definition,

$$T_k = \log \sum_{i \leq k} w_i, \quad w_i = \frac{1}{\sigma_i^2 + \tau^2}.$$

Therefore,

$$\left| \frac{\partial T_k}{\partial \tau^2} \right| = \left| \frac{\partial \log \sum_{i \leq k} w_i}{\partial \tau^2} \right|.$$

Since

$$\frac{\partial w_i}{\partial \tau^2} = -w_i^2 \quad \text{and} \quad \frac{\partial \log x}{\partial x} = \frac{1}{x},$$

we have

$$\begin{aligned} \left| \frac{\partial T_k}{\partial \tau^2} \right| &= \frac{\sum_{i \leq k} w_i^2}{\sum_{i \leq k} w_i} \\ &= \sum_{i \leq k} \frac{w_i}{\sum_{j \leq k} w_j} w_i \\ &\leq \max_i w_i \\ &\leq (\min_i \sigma_i^2)^{-1}. \end{aligned}$$

The second expression is a weighted average of the w_i , and therefore cannot exceed their maximum. Moreover, since $\tau^2 \geq 0$,

$$w_i = \frac{1}{\sigma_i^2 + \tau^2} \leq \frac{1}{\sigma_i^2}.$$

Once O is fixed, the map $\tau^2 \mapsto T_k$ is therefore Lipschitz with constant

$$L(O) = (\min_i \sigma_i^2)^{-1}.$$

For a function f and random variable V , if $|f'(V)| \leq L$, then

$$\text{Var}(f(V)) \leq L^2 \text{Var}(V).$$

Therefore,

$$\text{Var}(T_k \mid O) \leq (\min_i \sigma_i^2)^{-2} \text{Var}(\tau^2 \mid O).$$

Since τ^2 is independent of O and

$$\tau^2 \sim \text{U}(0, \tau_{\max}),$$

we have

$$\text{Var}(\tau^2 \mid O) = \text{Var}(\tau^2) = \frac{\tau_{\max}^2}{12}.$$

Hence,

$$\text{Var}(T_k \mid O) \leq \frac{\tau_{\max}^2}{12} (\min_i \sigma_i^2)^{-2}.$$

Taking expectations gives

$$\mathbb{E}[\text{Var}(T_k \mid O)] \leq \frac{\tau_{\max}^2}{12} \mathbb{E}[(\min_i \sigma_i^2)^{-2}].$$

By definition,

$$\rho_O^2 = 1 - \frac{\mathbb{E}[\text{Var}(T_k \mid O)]}{\text{Var}(T_k)}.$$

Therefore,

$$0 \leq 1 - \rho_O^2 \leq \frac{1}{12} \frac{\tau_{\max}^2 \mathbb{E}[(\min_i \sigma_i^2)^{-2}]}{\text{Var}(T_k)}.$$

Under the stated condition,

$$\frac{\tau_{\max}^2 \mathbb{E}[(\min_i \sigma_i^2)^{-2}]}{\text{Var}(T_k)} \rightarrow 0,$$

so

$$\rho_O^2 \rightarrow 1.$$

Since

$$0 \leq \Delta \leq 1 - \rho_O^2$$

and

$$\rho_O^2 \leq \rho_I^2 \leq 1,$$

it follows that

$$\Delta \rightarrow 0, \quad \rho_I^2 \rightarrow 1.$$

For part (ii), at the final position,

$$S_1^{(K)} = \sum_{i=1}^K \frac{1}{\sigma_i^2 + \tau^2}.$$

For every i ,

$$\tau^2 \leq \sigma_i^2 + \tau^2 \leq \tau^2 + \max_j \sigma_j^2.$$

Taking reciprocals,

$$\frac{1}{\tau^2 + \max_j \sigma_j^2} \leq \frac{1}{\sigma_i^2 + \tau^2} \leq \frac{1}{\tau^2}.$$

Summing over i gives

$$\frac{K}{\tau^2 + \max_i \sigma_i^2} \leq S_1^{(K)} \leq \frac{K}{\tau^2}.$$

Since $\log$ is increasing and $T_K = \log S_1^{(K)}$,

$$\log K - \log(\tau^2 + \max_i \sigma_i^2) \leq T_K \leq \log K - \log \tau^2.$$

Now,

$$\log(\tau^2 + \max_i \sigma_i^2) = \log \tau^2 + \log\left(1 + \frac{\max_i \sigma_i^2}{\tau^2}\right),$$

and therefore

$$\log K - \log \tau^2 - \log\left(1 + \frac{\max_i \sigma_i^2}{\tau^2}\right) \leq T_K \leq \log K - \log \tau^2.$$

Hence,

$$|T_K - (\log K - \log \tau^2)| \leq \log\left(1 + \frac{\max_i \sigma_i^2}{\tau^2}\right).$$

If

$$\frac{\max_i \sigma_i^2}{\tau^2} \rightarrow 0$$

in probability, then, by continuity of $\log(1+x)$,

$$\log\left(1 + \frac{\max_i \sigma_i^2}{\tau^2}\right) \rightarrow 0$$

in probability. The preceding bound therefore gives

$$T_K - (\log K - \log \tau^2) \rightarrow 0$$

in probability.

Convergence in probability alone does not guarantee convergence of the mean squared error. The additional uniform-integrability assumption on

$$\left\{\log^2\left(1 + \frac{\max_i \sigma_i^2}{\tau^2}\right)\right\}$$

rules out rare extreme values that would contribute a non-vanishing amount to the expected squared error. Since

$$(T_K - (\log K - \log \tau^2))^2 \leq \log^2\left(1 + \frac{\max_i \sigma_i^2}{\tau^2}\right),$$

the two conditions together imply

$$\mathbb{E}\left[(T_K - (\log K - \log \tau^2))^2\right] \rightarrow 0.$$

Equivalently,

$$T_K - (\log K - \log \tau^2) \rightarrow 0$$

in L^2 .

Let

$$g(\tau^2) := \log K - \log \tau^2.$$

Then

$$T_K - g(\tau^2) \rightarrow 0$$

in L^2 . Since τ^2 is independent of O ,

$$\mathbb{E}[g(\tau^2) \mid O] = \mathbb{E}[g(\tau^2)].$$

Therefore,

$$\text{Var}(\mathbb{E}[T_K \mid O]) = \text{Var}(\mathbb{E}[T_K - g(\tau^2) \mid O]).$$

By conditional Jensen's inequality,

$$\text{Var}(\mathbb{E}[T_K \mid O]) \leq \mathbb{E}[(T_K - g(\tau^2))^2] \rightarrow 0.$$

Finally, by the law of total variance and Proposition E.2,

$$\rho_O^2 = 1 - \frac{\mathbb{E}[\text{Var}(T_K \mid O)]}{\text{Var}(T_K)} = \frac{\text{Var}(\mathbb{E}[T_K \mid O])}{\text{Var}(T_K)}.$$

If $\text{Var}(T_K)$ remains bounded away from zero, it follows that

$$\rho_O^2 \rightarrow 0.$$

Thus, in this limit, nearly all predictable variance in T_K requires information about the latent heterogeneity. $\square$

Corollary E.5 (Distribution shifts can produce a probe–performance dissociation). *Under a distribution shift from a regime with substantial heterogeneity-dependent headroom to a sampling-variance-dominated regime,*

$$\rho_O^2 \rightarrow 1, \quad \Delta \rightarrow 0.$$

Consequently, probe R^2 for T can rise because O alone becomes more predictive, while the maximum amount by which any predictor can benefit from information about τ^2 contained in X shrinks to zero. Whether a particular network’s probe score rises, and whether its prediction error increases, must be determined experimentally.

Proposition E.6 (Ceilings for τ^2). *Any probe for τ^2 satisfies*

$$R^2 \leq \rho_I^2(\tau^2) := 1 - \frac{\text{MMSE}(\tau^2 \mid I)}{\text{Var}(\tau^2)}.$$

Proof. This is Proposition E.1 applied with target τ^2 . $\square$

Remark E.7 (Why identification of τ^2 deteriorates under the shift). If μ were known, one study would contribute Fisher information

$$\frac{1}{2(\sigma_i^2 + \tau^2)^2}$$

about τ^2 . Thus, scaling the total variance $\sigma_i^2 + \tau^2$ by a factor c would reduce this per-study information by a factor of c^2 . Since typical total variance is substantially larger under the OOD regime, this scaling provides an intuitive explanation for the reduced identifiability of τ^2 .

F THE GENERAL HEADROOM IDENTITY

The task-specific argument is an instance of a more general principle. Let Z be a latent variable, let $T = f(O, Z)$ be a square-integrable target, let X denote additional evidence that may reveal Z , and let $I := (O, X)$. Define

$$\rho_O^2 := 1 - \frac{\mathbb{E}[\text{Var}(T \mid O)]}{\text{Var}(T)}, \quad \rho_I^2 := 1 - \frac{\mathbb{E}[\text{Var}(T \mid I)]}{\text{Var}(T)}, \quad \Delta := \rho_I^2 - \rho_O^2,$$

and define the Bayes update (change in the Bayesian prediction of T after seeing the additional evidence X)

$$B(I, O) := \mathbb{E}[T \mid I] - \mathbb{E}[T \mid O].$$

Proposition F.1 (Headroom identity). *The headroom satisfies*

$$\Delta \text{Var}(T) = \mathbb{E}[\text{Var}(B(I, O) \mid O)].$$

Thus, the headroom is the conditional variance of the Bayes update about T supplied by X beyond the information already available in O .

Proof. By definition,

$$\Delta \text{Var}(T) = \mathbb{E}[\text{Var}(T \mid O)] - \mathbb{E}[\text{Var}(T \mid I)].$$

Conditional total variance gives

$$\text{Var}(T \mid O) = \mathbb{E}[\text{Var}(T \mid I) \mid O] + \text{Var}(\mathbb{E}[T \mid I] \mid O).$$

Taking expectations yields

$$\mathbb{E}[\text{Var}(T \mid O)] = \mathbb{E}[\text{Var}(T \mid I)] + \mathbb{E}[\text{Var}(\mathbb{E}[T \mid I] \mid O)].$$

Hence,

$$\Delta \text{Var}(T) = \mathbb{E}[\text{Var}(\mathbb{E}[T | I] | O)].$$

Since $\mathbb{E}[T | O]$ is constant conditional on O , subtracting it does not change the conditional variance, giving the result. $\square$

Corollary F.2 (Two mechanisms for vanishing headroom). *For each fixed O , suppose that $z \mapsto f(O, z)$ is Lipschitz continuous with Lipschitz constant $L(O)$,*

$$|f(O, z) - f(O, z')| \leq L(O)|z - z'|,$$

and has oscillation bounded by M ,

$$\text{osc}_z f(O, \cdot) := \sup_z f(O, z) - \inf_z f(O, z) \leq M.$$

Then the headroom can vanish through either of two mechanisms:

$$\Delta \leq \frac{\mathbb{E}[L(O)^2 \text{Var}(Z | O)]}{\text{Var}(T)} \quad \text{weak target sensitivity}$$

when the target $T = f(O, Z)$ becomes insensitive to the latent Z , and

$$\Delta \leq \frac{M^2 \text{MI}(X; Z | O)}{2 \text{Var}(T)} \quad \text{weak identification},$$

when the additional evidence X contains little information about Z beyond O , where MI denotes conditional mutual information in nats. The first inequality describes weak target sensitivity, in which the target becomes insensitive to the latent variable. The second describes weak identification, in which the additional evidence becomes uninformative about the latent variable given O .

Proof. (i) *Weak target sensitivity:* By the conditional law of total variance (proof of Proposition F.1),

$$\text{Var}(\mathbb{E}[T | I] | O) \leq \text{Var}(T | O),$$

since the variance of the conditional mean cannot exceed the variance of T . Since $T = f(O, Z)$, holding O fixed gives $\text{Var}(T | O) = \text{Var}(f(O, Z) | O)$. By the Lipschitz assumption,

$$\text{Var}(f(O, Z) | O) \leq L(O)^2 \text{Var}(Z | O).$$

Therefore,

$$\text{Var}(\mathbb{E}[T | I] | O) \leq L(O)^2 \text{Var}(Z | O).$$

Taking expectations and applying Proposition F.1 gives the first inequality.

$$\Delta \leq \frac{\mathbb{E}[L(O)^2 \text{Var}(Z | O)]}{\text{Var}(T)}.$$

Intuitively, if T barely changes when Z changes, learning Z cannot help very much.

(ii) *Weak identification:* Let $\pi_{Z|I}$ and $\pi_{Z|O}$ denote the posterior beliefs of Z given I and O . The oscillation bound gives

$$|B(I, O)| \leq M \text{TV}(\pi_{Z|I}, \pi_{Z|O}).$$

Therefore,

$$\mathbb{E}[\text{Var}(B(I, O) | O)] \leq \mathbb{E}[B(I, O)^2] \leq M^2 \mathbb{E}[\text{TV}^2(\pi_{Z|I}, \pi_{Z|O})].$$

Pinsker's inequality gives

$$\text{TV}^2(\pi_{Z|I}, \pi_{Z|O}) \leq \frac{1}{2} \text{KL}(\pi_{Z|I} \| \pi_{Z|O}),$$

whose expectation is $\text{MI}(X; Z | O)$. Applying Proposition F.1 completes the proof. $\square$

Draw	Meaning
$\log \lambda_g \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2)$	mean expression
$b_g \sim \text{U}(0, b_{\max})$	overdispersion (latent Z)
$\log d_{gi} \sim \mathcal{N}(\mu_d, \sigma_d^2)$	sequencing depth of cell i
$\theta_{gi} \sim \text{Gamma}(\text{shape} = \lambda_g/b_g, \text{scale} = b_g)$	true expression rate
$y_{gi} \mid \theta_{gi}, d_{gi} \sim \text{Poisson}(d_{gi}\theta_{gi})$	observed count

G SYNTHETIC SINGLE-CELL GENERATIVE MODELING

To test whether the estimation procedure generalizes beyond the controlled meta-analysis setting, we construct a second completely defined generative model based on single-cell RNA sequencing. Because the data-generating process is known, the reference points can be computed exactly and compared with their estimates (Appendix H).

Generative model. Each example is one gene g observed in $K = 40$ cells, indexed by $i = 1, \dots, K$. The draws are

with $\mu_\lambda = -4.5$, $\sigma_\lambda = 1$, $b_{\max} = 0.05$, $\mu_d = 6.9$ and $\sigma_d = 0.7$. The draws of λ_g , b_g and $d_{g,1:K}$ are mutually independent, and the depths are independent across genes and cells. (In the simulation the lower end of the prior on b_g is 10^{-6} rather than 0, which keeps the Gamma shape finite.) Each seed contains 1500 genes, and all results use five seeds. At these values the median depth is $e^{6.9} \approx 990$ and the median expected count $d_{gi}\lambda_g$ is $e^{2.4} \approx 11$.

The Gamma draw gives $\mathbb{E}[\theta_{gi} \mid \lambda_g, b_g] = \lambda_g$ and $\text{Var}(\theta_{gi} \mid \lambda_g, b_g) = \lambda_g b_g$, so true expression rate θ_{gi} differs from mean expression λ_g because of biological cell-to-cell variation (b_g), while observed count y_{gi} adds technical counting noise from the finite depth d_{gi} . Averaging the Poisson likelihood over θ_{gi} gives a negative-binomial count,

$$y_{gi} \mid \lambda_g, b_g, d_{gi} \sim \text{NB}\left(\text{mean} = d_{gi}\lambda_g, \text{size} = \frac{\lambda_g}{b_g}\right),$$

with $\text{Var}(y_{gi} \mid \lambda_g, b_g, d_{gi}) = d_{gi}\lambda_g + b_g d_{gi}^2 \lambda_g$. The normalized count $r_{gi} = y_{gi}/d_{gi}$ therefore has

$$\text{Var}(r_{gi} \mid \lambda_g, b_g, d_{gi}) = \lambda_g \left(\frac{1}{d_{gi}} + b_g \right),$$

the sum of a technical component $1/d_{gi}$ and a biological component b_g . This is the variance structure of the meta-analysis task,

$$\sigma_i^2 \leftrightarrow \frac{1}{d_{gi}}, \quad \tau^2 \leftrightarrow b_g.$$

Target and information sets. For known biological overdispersion b_g , cell i has precision $w_{gi} = (d_{gi}^{-1} + b_g)^{-1}$. The probe target is

$$T_g = \log \sum_{i=1}^K \frac{1}{d_{gi}^{-1} + b_g},$$

the analogue of $\log S_1$ with d_{gi}^{-1} in place of σ_i^2 and b_g in place of τ^2 . The restricted information is the depths alone, $O_g = d_{g,1:K}$; the additional evidence is the observed counts, $X_g = y_{g,1:K}$; and the full information is $I_g = (y_{g,1:K}, d_{g,1:K})$. Because b_g is independent of $d_{g,1:K}$, the floor ρ_O^2 comes only from how T_g varies with the depths, averaged over the prior on b_g ; inferring b_g requires the counts. The floor ρ_O^2 and ceiling ρ_I^2 are the population reference points of Section 4.1 (Propositions E.1 and E.2).

Binomial thinning. Retaining a fraction $p \in (0, 1]$ of the counts gives $y'_{gi} \sim \text{Bin}(y_{gi}, p)$ and $d'_{gi} = p d_{gi}$. Because binomial thinning of $\text{Poisson}(d_{gi}\theta_{gi})$ is $\text{Poisson}(p d_{gi}\theta_{gi})$, the thinned data follow the same model with depths $p d_{gi}$ (so $\log d'_{gi} \sim \mathcal{N}(\mu_d + \log p, \sigma_d^2)$).

G.1 Separating the two mechanisms of vanishing headroom

(Figure 6). Sequencing depth affects headroom in two ways at once: shallower data give fewer counts from which to infer the dispersion b_g (weaker identification), and also make technical noise $1/d_{gi}$ larger relative to b_g in the target, so that T_g depends less on b_g (weaker sensitivity). In real data a single depth p moves both. In the synthetic model we vary them separately with two depth fractions:

- the *evidence fraction* p_X thins only the counts the ceiling sees: the count-based features of ϕ_I are computed from thinned counts $y'_{gi} \sim \text{Bin}(y_{gi}, p_X)$ with depths $p_X d_{gi}$;
- the *target fraction* p_T scales only the depths in the target, $T_g = \log \sum_i ((p_T d_{gi})^{-1} + b_g)^{-1}$.

The floor features ϕ_O always use the full depths $d_{g,1:K}$; since p_T is a known constant, they carry the same information as $p_T d_{g,1:K}$. Both fractions run on a log grid, $p_X \in \{1, \dots, 10^{-2}\}$ (4 values) and $p_T \in \{1, \dots, 10^{-4}\}$ (7 values), in steps of $10^{-2/3}$. For scale, at full depth the median technical variance $1/d_{gi} \approx 0.001$ is small next to $b_g \leq 0.05$, whereas at $p = 0.01$ it is ≈ 0.1 and dominates b_g . Each grid value is the mean over five seeds of the estimated headroom $\hat{\rho}_I^2 - \hat{\rho}_O^2$, both fitted by the decoder bank on the grid's own 50/20/30 split of the genes. At full depth this gives 0.87, against 0.89 in Table 13, which uses a different split, and the exact 0.880 (Table 3).

G.2 Real data.

The four real 10x cohorts (10x Genomics, 2021, 2024, 2022) are cast in the same form. We keep the 2000 most frequently detected genes among those detected in at least 50 cells. The sequencing depth of gene g in cell i is $d_{gi} = \ell_i f_g$, where ℓ_i is the library size of cell i and f_g the fraction of all counts that belong to gene g . The cells are split in half at random: biological overdispersion parameter b_g is replaced by its method-of-moments estimate $\hat{b}_{\text{MoM}}$ (Appendix I) from the first half, and the target, O_g, I_g , the features and the scGPT input use the second half, so the ceiling cannot recover its own target estimate. K is the number of cells in the second half. The features are those of the synthetic model, with $b_{\max}$ set to 1.1 times the largest $\hat{b}_{\text{MoM}}$ among training genes.

H EXACT REFERENCE POINTS AND VALIDATION OF THE ESTIMATION PROCEDURE

We test the floor–ceiling construction in two independently specified generative models: the random-effects meta-analysis model and a synthetic single-cell RNA-sequencing model. In both cases the latent distribution and observation model are known, so the reference points can be computed exactly, independently of the estimation procedure.

Population reference points. Let T be the probe target, let O denote the restricted information, and let $I = (O, X)$ denote the full information after observing the additional evidence X . Under squared-error loss, the optimal predictors measurable with respect to these information sets are

$$m_O(O) = \mathbb{E}[T \mid O], \quad m_I(I) = \mathbb{E}[T \mid I].$$

The corresponding population reference points are

$$\rho_O^2 = 1 - \frac{\mathbb{E}[\text{Var}(T \mid O)]}{\text{Var}(T)} = R^2(T, m_O(O)),$$

and

$$\rho_I^2 = 1 - \frac{\mathbb{E}[\text{Var}(T \mid I)]}{\text{Var}(T)} = R^2(T, m_I(I)).$$

Their difference,

$$\Delta = \rho_I^2 - \rho_O^2,$$

is the headroom supplied by X beyond O .

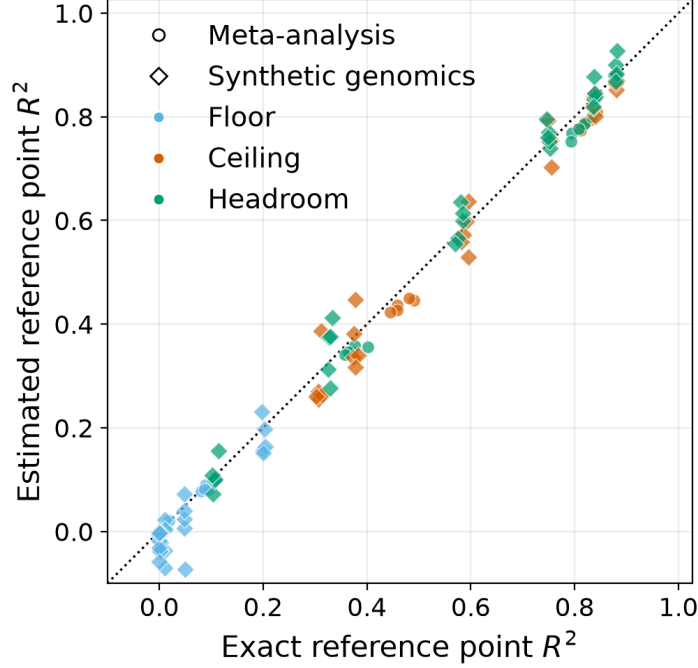


Figure 7: **The estimation procedure recovers exact reference points.** Each marker compares an estimated reference point (decoder bank; vertical axis) with its exact value (horizontal axis), for the meta-analysis (circles) and the synthetic single-cell model (diamonds). Points lie close to the identity line (dotted) across regimes and reference-point types. Slightly negative estimated floors are held-out R^2 values where the exact floor is near zero, not negative population floors.

These definitions follow directly from the L^2 projection property of conditional expectation. Because $O \subseteq I$,

$$\mathbb{E}[\text{Var}(T \mid I)] \leq \mathbb{E}[\text{Var}(T \mid O)],$$

and hence

$$0 \leq \rho_O^2 \leq \rho_I^2 \leq 1, \quad \Delta \geq 0.$$

Moreover, for any representation $H = \Phi(I)$ and any probe h ,

$$\text{MSE}(T, h(H)) \geq \text{MMSE}(T \mid I),$$

which implies

$$R^2(T, h(H)) \leq \rho_I^2.$$

Thus the population ceiling ρ_I^2 bounds every representation constructed from I , while ρ_O^2 is the amount of predictability already attainable without using the additional evidence X .

Exact meta-analysis reference points. For the random-effects model,

$$x_i \mid \mu, \tau^2, \sigma_i^2 \sim \mathcal{N}(\mu, \sigma_i^2 + \tau^2), \quad \mu \sim \mathcal{N}(0, 1), \quad \tau^2 \sim \text{U}(0, \tau_{\max}),$$

with $\tau_{\max} = 0.5$ in distribution and $\tau_{\max} = 1.5$ out of distribution (Appendix A), we define

$$O = \sigma_{1:K}^2, \quad I = (\sigma_{1:K}^2, x_{1:K}),$$

and

$$T = \log S_1 = \log \sum_{i=1}^K \frac{1}{\sigma_i^2 + \tau^2}.$$

Because τ^2 is independent of O , the exact restricted-information predictor is

$$m_O(O) = \int \log \left(\sum_{i=1}^K \frac{1}{\sigma_i^2 + \tau^2} \right) p(\tau^2) d\tau^2.$$

To obtain the exact full-information predictor, we marginalize the nuisance mean μ . For a candidate τ^2 , define

$$\nu_i = \sigma_i^2 + \tau^2, \quad A_1 = \sum_i \nu_i^{-1}, \quad A_x = \sum_i x_i \nu_i^{-1}, \quad A_{xx} = \sum_i x_i^2 \nu_i^{-1}.$$

The marginal log likelihood is

$$\log p(x_{1:K} \mid \tau^2, O) = -\frac{1}{2} \left[\sum_i \log \nu_i + \log(1 + A_1) + A_{xx} - \frac{A_x^2}{1 + A_1} \right] + \text{const.}$$

Therefore,

$$p(\tau^2 \mid I) \propto p(x_{1:K} \mid \tau^2, O) p(\tau^2),$$

and

$$m_I(I) = \int \log \left(\sum_{i=1}^K \frac{1}{\sigma_i^2 + \tau^2} \right) p(\tau^2 \mid I) d\tau^2.$$

The one-dimensional prior and posterior integrals are evaluated by numerical quadrature over τ^2 . The exact floor and ceiling are then obtained by evaluating m_O and m_I on independent held-out meta-analyses in both the in-distribution and out-of-distribution regimes.

Exact reference points of the synthetic single-cell model. The model, target T_g and information sets $O_g = d_{g,1:K}$ and $I_g = (y_{g,1:K}, d_{g,1:K})$ are defined in Appendix G. Write $T(d_g, b) = \log \sum_{i=1}^K (d_{gi}^{-1} + b)^{-1}$. Because b_g is independent of the depth vector d_g ,

$$p(b_g \mid O_g) = p(b_g),$$

and the exact floor predictor is

$$m_O(d_g) = \int T(d_g, b) p(b) db.$$

Its conditional residual variance is

$$v_O(d_g) = \int T(d_g, b)^2 p(b) db - m_O(d_g)^2.$$

With the negative-binomial likelihood of the counts, $y_{gi} \mid \lambda_g, b_g, d_{gi} \sim \text{NB}(\text{mean} = d_{gi}\lambda_g, \text{size} = \lambda_g/b_g)$ (Appendix G), independent across cells,

$$p(b_g \mid y_g, d_g) \propto p(b_g) \int p(y_g \mid \lambda, b_g, d_g) p(\lambda) d\lambda.$$

The exact full-information predictor is

$$m_I(y_g, d_g) = \int T(d_g, b) p(b \mid y_g, d_g) db,$$

with conditional residual variance

$$v_I(y_g, d_g) = \text{Var}(T_g \mid y_g, d_g).$$

The exact reference points of the synthetic single-cell model are therefore

$$\rho_O^2 = 1 - \frac{\mathbb{E}[v_O(O)]}{\text{Var}(T)}, \quad \rho_I^2 = 1 - \frac{\mathbb{E}[v_I(I)]}{\text{Var}(T)}, \quad \Delta = \rho_I^2 - \rho_O^2.$$

The integrals over (λ, b) are evaluated on a fixed numerical grid: 150 cells in b on $(0, b_{\max})$, with edges $b_{\max} u^2$ for u equally spaced on $[0, 1]$ so that the grid is dense near $b = 0$, and 50 points in $\log \lambda$, equally spaced on $\mu_\lambda \pm 3.5 \sigma_\lambda$. Under binomial thinning (Appendix G) the same computation is applied to the thinned (y', d') . The numerator and denominator are computed using the same quadrature rule, ensuring the total-variance identity

$$\text{Var}(T) = \mathbb{E}[\text{Var}(T \mid O)] + \text{Var}(\mathbb{E}[T \mid O])$$

up to numerical precision.

Domain	Target	Condition	Floor	Ceiling	Headroom
Meta-analysis	$\log S_1$	ID	0.012 ± 0.004	0.822 ± 0.014	0.809 ± 0.017
		OOD	0.092 ± 0.011	0.466 ± 0.019	0.374 ± 0.017
	τ^2	ID	0.000 ± 0.000	0.758 ± 0.013	0.758 ± 0.013
		OOD	0.000 ± 0.000	0.404 ± 0.017	0.404 ± 0.017
Synthetic genomics	T_g	Full depth	0.000 ± 0.000	0.880 ± 0.001	0.880 ± 0.001

Table 3: Exact reference points for the meta-analysis and synthetic-genomics validation experiments (mean $\pm$ sample standard deviation across five seeds).

Validation against exact values. We independently estimate the same reference points from observed samples without giving the estimator access to the true latent variable. Each seed is divided into disjoint training, model-selection, and evaluation partitions. Candidate regressors from the decoder bank are fitted on the training partition, the regressor class is fixed using the selection partition, and the selected class is refitted on the combined training and selection data before evaluation.

This produces

$$\hat{\rho}_O^2 = R^2(T, \hat{m}_O(O)), \quad \hat{\rho}_I^2 = R^2(T, \hat{m}_I(I)), \quad \hat{\Delta} = \hat{\rho}_I^2 - \hat{\rho}_O^2.$$

The validation compares

$$(\rho_O^2, \hat{\rho}_O^2), \quad (\rho_I^2, \hat{\rho}_I^2), \quad (\Delta, \hat{\Delta})$$

for every seed, for both meta-analysis regimes, and for six thinning fractions $p \in \{1, 0.40, 0.16, 0.063, 0.025, 0.01\}$ of the synthetic single-cell model (Appendix G). Figure 7 shows the comparison: floors, ceilings and headroom estimated by the estimation procedure lie close to their exact values in both models, and Table 3 lists the exact reference points.

I FEATURE ENGINEERING FOR ESTIMATED REFERENCE POINTS

The exact reference points (Appendix H) condition on the raw information sets O and I . The estimated reference points are instead fitted by the decoder bank (Appendix B) on fixed-length feature vectors $\phi_O(O)$ and $\phi_I(I)$:

$$\hat{\rho}_O^2 = R^2(T, \hat{m}_O(\phi_O(O))), \quad \hat{\rho}_I^2 = R^2(T, \hat{m}_I(\phi_I(I))),$$

where T is the probe target and $\hat{m}$ the fitted regressor on feature vectors. Each feature vector is a deterministic function of its own information set only. It never uses the latent variable (τ^2 or b) or the target, and ϕ_I always contains ϕ_O . The features are fixed in advance and are identical across seeds, conditions and sequencing depths.

Notation. For a vector $a_{1:K}$, $\bar{a} = K^{-1} \sum_i a_i$; $\text{sd}(a) = (K^{-1} \sum_i (a_i - \bar{a})^2)^{1/2}$; $s^2(a) = (K - 1)^{-1} \sum_i (a_i - \bar{a})^2$; $a_{(1)} \leq \dots \leq a_{(K)}$ are the order statistics; $q_\alpha(a)$ is the empirical α -quantile with linear interpolation; $\text{med}(a) = q_{0.5}(a)$. We write $\varepsilon = 10^{-12}$ and $\text{slog}(a) = \log \max(a, \varepsilon)$.

Meta-analysis. Features are computed from all $K = 32$ studies of a meta-analysis (the final position). Let $\ell_i = \log \sigma_i^2$ and $w_i = 1/\sigma_i^2$. Restricted features ϕ_O (43 columns), from $O = \sigma_{1:K}^2$:

1. sorted log sampling variances $\ell_{(1)}, \dots, \ell_{(K)}$ (32);
2. $\text{mean}(\ell)$, $\text{sd}(\ell)$, $\min_i \ell$, $\max_i \ell$ (4);
3. quantiles $q_\alpha(\ell)$ for $\alpha \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ (5);
4. the log total and log squared total fixed-effect precision, $\log \sum_i w_i$ and $\log \sum_i w_i^2$ (2).

Full features ϕ_I (114 columns), from $I = (\sigma_{1:K}^2, x_{1:K})$:

1. ϕ_O (43);
2. the effects $x_1, \dots, x_K$ in study order (32);
3. $\ell_1, \dots, \ell_K$ in study order, so that each x_i is paired with its own σ_i^2 (32);
4. mean x , $\text{sd}(x)$, $\min_i x$, $\max_i x$ (4);
5. the fixed-effect mean $\hat{\mu}_{\text{FE}} = \sum_i w_i x_i / \sum_i w_i$, Cochran's $Q = \sum_i w_i (x_i - \hat{\mu}_{\text{FE}})^2$, and the DerSimonian–Laird estimate

$$\hat{\tau}_{\text{DL}}^2 = \max\left\{0, \frac{Q - (K - 1)}{\sum_i w_i - \sum_i w_i^2 / \sum_i w_i}\right\} \quad (3).$$

No information is lost. $T = \log S_1$ is symmetric in the studies and τ^2 is independent of O , so $\mathbb{E}[T \mid O]$ depends on $\sigma_{1:K}^2$ only through the sorted vector in item 1, and $\mathbb{E}[T \mid O] = \mathbb{E}[T \mid \phi_O(O)]$. Likewise, I is recovered exactly from items 2–3 of ϕ_I . The remaining columns are deterministic summaries that ease fitting. The estimated and exact reference points therefore target the same population quantities.

Synthetic single-cell model. Features are computed separately for each gene using its $K = 40$ cells (Appendix G). They are constructed from the observed counts y and sequencing depths d after binomial thinning (Appendix G). Define the normalized count $r_i = \frac{y_i}{d_i}$, the moment estimate of mean expression $\hat{\lambda} = \text{mean}(r)$, and the sample variance $s^2 = s^2(r)$.

The restricted feature vector ϕ_O has 8 columns and uses only $O = d_{1:K}$. It contains:

1. the number of cells K (1);
2. $\text{mean}(\log d)$, $\text{sd}(\log d)$, $\min_i \log d$, $\max_i \log d$, and $\text{med}(\log d)$ (5);
3. the log total and log squared total technical variance, $\text{slog} \sum_i d_i^{-1}$ and $\text{slog} \sum_i d_i^{-2}$ (2).

The full feature vector ϕ_I has 27 columns and uses $I = (y_{1:K}, d_{1:K})$. We first define three quantities that summarize biological variation beyond Poisson counting noise: (1) the expected Poisson variance, $v = \hat{\lambda} \frac{1}{K} \sum_i d_i^{-1}$, where $\hat{\lambda}$ is the estimated mean expression, K is the number of cells, and d_i is the sequencing depth of cell i ; (2) the excess variance, $e = \max(0, s^2 - v)$, where s^2 is the sample variance of the normalized counts $r_i = y_i/d_i$; and (3) the method-of-moments estimate of biological dispersion, $\hat{b}_{\text{MoM}} = e / \max(\hat{\lambda}, \varepsilon)$, where ε is a small constant used to avoid division by zero.

These definitions follow from $\text{Var}(r_i) = \lambda/d_i + \lambda b$, where λ/d_i is the variance expected from Poisson counting noise and λb is the additional variance due to biological dispersion b . We also define the depth-weighted mean $\bar{r}_d = \sum_i y_i / \sum_i d_i$, and the weighted dispersion statistic $Q = \sum_i d_i (r_i - \bar{r}_d)^2$.

The 27 features in ϕ_I are:

1. the 8 restricted features in ϕ_O (8);
2. two sparsity features: the detection rate $K^{-1} \sum_i \mathbb{1}[y_i > 0]$, which is the fraction of cells with a nonzero count, and $\log(1 + \max_i y_i)$, which summarizes the largest observed count (2);
3. three variance features: $\log(s^2 + \varepsilon)$ for the total observed variance of the normalized counts, $\log(v + \varepsilon)$ for the variance expected from Poisson counting noise, and $\log(e + \varepsilon)$ for the excess variance beyond that noise (3);
4. two representations of the moment estimate of biological dispersion b : $\log(\hat{b}_{\text{MoM}} + \varepsilon)$ and $\hat{b}_{\text{MoM}}$ (2);
5. four overdispersion features: Q ; $\log(1 + \max\{Q - (K - 1), 0\})$, which records the amount by which Q exceeds the reference value $K - 1$; the dispersion ratio $s^2 / \max(v, \varepsilon)$, which compares observed variance with the variance expected from Poisson noise; and the normalized-count spread $q_{0.9}(r) - q_{0.1}(r)$, the difference between the 90th and 10th percentiles of r (4);

6. seven profile-likelihood features. We evaluate the negative-binomial log-likelihood at seven candidate dispersion values $b_j = j b_{\max}/7$, for $j = 1, \dots, 7$. We fix the mean-expression parameter at $\tilde{\lambda} = \max(\hat{\lambda}, \epsilon)$, define the expected count for cell i as $m_i = d_i \tilde{\lambda}$, and define the corresponding negative-binomial shape parameter as $\kappa_j = \tilde{\lambda}/b_j$. For each candidate b_j , the log-likelihood is $L_j = \sum_i [\ln \Gamma(y_i + \kappa_j) - \ln \Gamma(\kappa_j) - \ln \Gamma(y_i + 1) + \kappa_j \log \frac{\kappa_j}{\kappa_j + m_i} + y_i \log \frac{m_i}{\kappa_j + m_i}]$. This is the likelihood from Appendix G, with λ fixed at $\tilde{\lambda}$. We use the seven relative likelihood values $L_j - \max_k L_k$ as features, so that the best grid value has value zero and the others indicate how much less likely they are (7);
7. one likelihood-based estimate of b : the grid value with the largest likelihood, b_{j^*} , where $j^* = \arg \max_j L_j$ (1).

For the synthetic model, $b_{\max} = 0.05$, which is the upper bound of the prior on biological overdispersion b . For the real cohorts, $b_{\max}$ is set to 1.1 times the largest $\hat{b}_{\text{MoM}}$ among the training genes (Appendix G).

This construction parallels the meta-analysis setting. The restricted features ϕ_O summarize the technical variances $1/d_i$, which play the same role as the sampling variances σ_i^2 . The full features ϕ_I additionally include information about biological dispersion, analogous to adding Q and $\hat{\tau}_{\text{DL}}^2$ in the meta-analysis.

The analogy for Q is only approximate. Here, $Q = \bar{r}_d \chi_P^2$, where $\chi_P^2 = \sum_i (y_i - d_i \bar{r}_d)^2 / (d_i \bar{r}_d)$ is Pearson’s statistic. Therefore Q is not standardized, and comparing Q directly with $K - 1$ in item 5 is only a heuristic rather than a calibrated DerSimonian–Laird statistic. Standardized information about excess variation is still available to the decoder through s^2/v and $\hat{b}_{\text{MoM}}$.

Both feature vectors discard some information. Unlike the meta-analysis setting, ϕ_O keeps only 8 summaries of the K sequencing depths, while ϕ_I does not include the raw vectors $y_{1:K}$ and $d_{1:K}$. Therefore, the best predictors based on ϕ_O and ϕ_I can fall below the exact population quantities ρ_O^2 and ρ_I^2 , which condition on the full raw data. Figure 7 compares these estimated and exact values.

The full feature vector ϕ_I nevertheless contains two direct estimates of the latent biological dispersion b : the moment estimator $\hat{b}_{\text{MoM}}$ and the likelihood estimator b_{j^*} . These features give the estimated reference points at all six thinning fractions in Figure 7, the full-depth ($p = 1$) floor and ceiling in Figure 4 and Table 13, and the headroom grid in Figure 6.

J SENSITIVITY TO THE CHOICE OF (O, I, T)

The quantities used for normalization in Section 4.1 are conditional on the choice of (O, I, T) . In the controlled meta-analysis and synthetic single-cell experiments, the generative model identifies a natural distinction between restricted and full information. In the observational chemistry and neuroscience applications, however, there is no unique decomposition of the measurements into O and I , because changing the choice of (O, I, T) changes the scientific estimand itself.

We therefore evaluate several prespecified, scientifically interpretable contrasts in each observational domain to determine whether the qualitative interpretation of a probe depends strongly on a particular placement of the boundary between restricted and additional information. Whenever possible, we hold the target (T) and full information I fixed and progressively enrich the restricted baseline,

$$\sigma(O_1) \subseteq \sigma(O_2) \subseteq \sigma(O_3) \subseteq \sigma(O_4) \subseteq \sigma(I).$$

Population monotonicity then implies

$$\rho_{O_1}^2 \leq \rho_{O_2}^2 \leq \rho_{O_3}^2 \leq \rho_{O_4}^2 \leq \rho_I^2, \quad \Delta(O_1, I) \geq \Delta(O_2, I) \geq \Delta(O_3, I) \geq \Delta(O_4, I),$$

although estimated reference points need not satisfy these inequalities.

J.1 Chemistry

Dataset and experimental setting. We use part 1 of the Raman-ChEMBL database, which provides molecular composition, physicochemical descriptors, three-dimensional molecular geometry, and quantum-chemical orbital energies computed with density functional theory (PBE/6-31G). We draw a fixed random sample of 50,000

molecules, keep the 48,931 that pass structural validation, and use 30,000 of these per seed. The analysis unit is one molecule. Identical SMILES are assigned to the same data split so that duplicate molecular identities do not occur across fine-tuning, validation, probe-fitting, and final-evaluation sets.

The experiment is useful for probe normalization because molecular orbital energies are partly predictable from relatively coarse molecular properties such as composition, charge, and global shape, while detailed atomic geometry provides additional information. We therefore ask whether TabPFN hidden representations contain information about orbital energies beyond what was already predictable from progressively richer non-geometric or coarse molecular summaries. TabPFN hidden-state probes use a 64-component PCA fitted on training rows only. All other probes, floors and ceilings use full features.

Prediction target. We analyze the two frontier orbital energies separately,

$$T^{\text{HOMO}} = E_{\text{HOMO}}, \quad T^{\text{LUMO}} = E_{\text{LUMO}},$$

where E_{HOMO} is the energy of the highest occupied molecular orbital and E_{LUMO} is the energy of the lowest unoccupied molecular orbital. Both are measured in Hartree.

These two quantities also determine the HOMO–LUMO gap,

$$E_g = E_{\text{LUMO}} - E_{\text{HOMO}}.$$

TabPFN is fine-tuned to predict E_g , but HOMO and LUMO are not individually used as fine-tuning targets. They are subsequently used only as targets of the representation audit.

Restricted molecular information. The restricted information O is constructed from progressively richer fixed-size molecular summaries. We consider four nested specifications for restricted information,

$$O_1 \subset O_2 \subset O_3 \subset O_4.$$

The first contains elemental composition and total formal charge,

$$O_1 = (\text{element counts}, Q),$$

where Q denotes total formal charge. This 18-dimensional representation describes which atomic species are present and their abundances, but does not specify how the atoms are arranged in space.

The second specification additionally contains three rotational constants,

$$O_2 = (O_1, R_x, R_y, R_z).$$

These quantities summarize the distribution of molecular mass around the principal axes and therefore provide coarse information about molecular size, mass distribution, and global shape without specifying individual atomic distances.

The third specification additionally includes the total dipole magnitude,

$$O_3 = (O_2, D_{\text{total}}).$$

The total dipole provides a coarse summary of molecular charge separation and therefore introduces global electrostatic information beyond composition and shape.

Finally, the richest specification is

$$O_4 = (O_3, \alpha_{\text{iso}}, R_e^2, C_v),$$

where α_{iso} is isotropic polarizability, R_e^2 denotes electronic spatial extent, and C_v is heat capacity. These quantities add information about the response and spatial extent of the electronic distribution and about the molecule’s thermodynamic state. Thus the restricted baseline becomes progressively more informative while remaining a compact tabular description of the molecule.

Full information I . The full information is held fixed across all four specifications:

$$I = (O_4, G_{\text{MBTR}}),$$

where G_{MBTR} is the unchanged monolithic MBTR geometry block.

The MBTR representation encodes species-resolved pairwise inverse-distance information. It describes which types of atoms occur at which relative distances. It is invariant to translation, rotation, and atom ordering. Unlike the tabular descriptors in the restricted information, it provides a detailed representation of molecular geometry.

As O_j becomes richer, the variables that I adds become more specific. Above O_1 , it includes the full physico-chemical and geometric information absent from composition alone, whereas above O_4 the remaining increment is specifically the detailed MBTR geometry block.

The associated headroom,

$$\hat{\Delta}_j = \hat{\rho}_I^2 - \hat{\rho}_{O_j}^2,$$

measures how much additional predictability of the orbital energy remains after conditioning on the corresponding molecular baseline.

Model representation. TabPFN is fine-tuned with LoRA using the fixed full information I to predict

$$E_g = E_{\text{LUMO}} - E_{\text{HOMO}}.$$

Neither E_{HOMO} nor E_{LUMO} is supplied as an individual fine-tuning target. LoRA adapters (rank 8, $\alpha = 16$, dropout 0.05) are added to the attention and MLP projections of every TabPFN v3 block and to its output projection, and are fitted on molecules disjoint from those used for probing.

After fine-tuning, hidden representations $H^{(\ell)}$ are extracted from the audited TabPFN layers. Probes from the decoder bank are then trained separately to predict T^{HOMO} and T^{LUMO} from these representations, producing $R_{H^{(\ell)}}^2$.

Thus, the estimated reference points $\hat{\rho}_{O_j}^2$ and $\hat{\rho}_I^2$ correspond to the chosen (O_j, I, T) , and $R_{H^{(\ell)}}^2$ measures the orbital-energy information accessible from the model’s internal representation. TabPFN receives O_4 and a 99%-variance PCA of the MBTR block; the ceiling probe uses the raw MBTR, so this PCA does not limit the ceiling.

What is the scientific question? The raw question

“How well can a TabPFN hidden state predict HOMO or LUMO?”

is insufficient because substantial orbital-energy information may already be predictable from molecular composition or other coarse physicochemical properties. The question posed by normalization is instead

“How should orbital-energy decodability be interpreted relative to what was already predictable from a declared molecular summary and what additional predictability is available from the full molecular descriptor set I ?”

For each orbital-energy target and each choice of (O_j, I, T) , we therefore compare $\hat{\rho}_{O_j}^2$, $\hat{\rho}_I^2$ and $R_{H^{(\ell)}}^2$. The first quantity measures what is predictable from the restricted molecular description, the second measures what is predictable from the fixed geometry-aware representation, and their difference defines the estimated headroom $\hat{\Delta}_j$. The model probe is interpreted relative to these quantities rather than on an absolute R^2 scale.

The sequence $O_1 \subset O_2 \subset O_3 \subset O_4$ further provides a sensitivity analysis for the choice of restricted information. O_1 asks what the complete molecular representation adds beyond composition and charge; O_2 additionally conditions on coarse molecular size and shape; O_3 further conditions on global charge separation; and O_4 conditions on a rich physicochemical state before asking what detailed geometry adds. These baseline specifications define distinct estimands. If the interpretation of the TabPFN probe changes across them, that dependence identifies which classes of molecular information account for the apparent decodability.

Result. Most of each raw probe score is already predictable from the physicochemical baseline O_4 . Above it, the fine-tuned TabPFN recovers 69% (HOMO) and 82% (LUMO) of the additional predictability that detailed geometry provides (Table 13). Neither orbital energy was a fine-tuning target, so substantial information about both is present in the representation even though the model was trained only on their difference. The interpretation is stable across the nested baselines O_1 – O_4 .

J.2 Neuroscience

Related work and motivation. Neuroscience has long used decoding to ask what neural populations represent (Quiñero Quiroga and Panzeri, 2009; Naselaris et al., 2011; Glaser et al., 2020), and has documented its limits (Kriegeskorte and Douglas, 2019; Ritchie et al., 2019; Hebart and Baker, 2018). Two common practices resemble our reference points. First, decoding analyses control for confounds by removing the variance explained by variables that are not of interest (Snoek et al., 2019), and encoding studies split variance between feature sets (de Heer et al., 2017). Our floor plays this role, with the baseline stated explicitly and chosen by the scientific question. Second, encoding models are compared against a noise ceiling: the part of the neural response that repeats across identical trials, which no model can exceed (Hsu et al., 2004; Schoppe et al., 2016; Lage-Castellanos et al., 2019). Unlike a noise ceiling, which is limited by trial-to-trial variability in the response, the ceiling we define is limited by how much the available information reveals about the target.

Dataset and experimental setting. We use the Area 2 dataset of Chowdhury et al. (2022), consisting of neural population recordings from proprioceptive Area 2 of somatosensory cortex during an arm-reaching task with externally imposed mechanical perturbations. On perturbation trials, the animal holds or moves a manipulandum while a brief planar bump is applied through the interface. The dataset contains simultaneously recorded neural activity together with hand position, hand velocity, interface force, and the commanded perturbation. The analysis unit is one perturbation trial.

The experiment is useful for probe normalization because the subsequent mechanical response is partly predictable from the state of the limb before the perturbation, while knowledge of the externally imposed bump provides additional information. We therefore ask whether post-perturbation neural population activity contains information about the ensuing force response beyond what was already predictable from the measured pre-bump state.

Prediction target. We analyze the two planar force components separately. For component $c \in \{x, y\}$, the target is the baseline-corrected interface-force response during the first 100 ms after bump onset,

$$T_c = \overline{F_c(t)}_{0 \leq t < 100 \text{ ms}} - \overline{F_c(t)}_{-100 \leq t < 0 \text{ ms}}.$$

Thus T_c measures the change in force following the perturbation relative to the force already present immediately before it. Baseline correction removes pre-existing load, tonic force, and constant sensor offsets, so the target reflects the perturbation-evoked mechanical response rather than the absolute force reading.

Restricted pre-bump information. The restricted information O is constructed entirely from measurements before bump onset. The primary pre-bump window is $[-100, 0)$ ms, which summarizes the limb state immediately before the perturbation while excluding post-bump information. Averaging within this interval reduces measurement noise relative to conditioning on a single time point and provides a local estimate of the state from which the perturbation begins.

We consider progressively richer restricted-information specifications:

$$\begin{aligned} O_1 &= (\bar{x}_{\text{pre}}, \bar{y}_{\text{pre}}), \\ O_2 &= (O_1, \bar{v}_{x_{\text{pre}}}, \bar{v}_{y_{\text{pre}}}), \\ O_3 &= (O_2, \bar{F}_{x_{\text{pre}}}, \bar{F}_{y_{\text{pre}}}). \end{aligned}$$

Here each bar denotes the mean over $[-100, 0)$ ms. These contrasts progressively condition on hand location, motion, and the force already being exerted before the bump.

Baseline	Target	Floor mean ± SD	Ceiling mean ± SD	Headroom mean ± SD	Probe mean ± SD	U_H mean ± SD
Neuroscience						
O_1	Force x	-0.013 ± 0.019	0.987 ± 0.009	1.000 ± 0.025	0.745 ± 0.014	0.758 ± 0.019
O_1	Force y	-0.002 ± 0.006	0.997 ± 0.000	0.999 ± 0.006	0.573 ± 0.019	0.575 ± 0.017
O_2	Force x	0.004 ± 0.009	0.987 ± 0.009	0.983 ± 0.016	0.745 ± 0.014	0.754 ± 0.018
O_2	Force y	-0.013 ± 0.028	0.997 ± 0.000	1.010 ± 0.028	0.573 ± 0.019	0.580 ± 0.011
O_3	Force x	-0.002 ± 0.019	0.987 ± 0.009	0.989 ± 0.025	0.745 ± 0.014	0.755 ± 0.019
O_3	Force y	-0.003 ± 0.005	0.997 ± 0.000	1.000 ± 0.005	0.573 ± 0.019	0.575 ± 0.019
O_4	Force x	0.111 ± 0.023	0.987 ± 0.009	0.876 ± 0.028	0.745 ± 0.014	0.724 ± 0.018
O_4	Force y	0.088 ± 0.024	0.997 ± 0.000	0.909 ± 0.024	0.573 ± 0.019	0.533 ± 0.025
Chemistry						
O_1	HOMO	0.378 ± 0.004	0.628 ± 0.005	0.250 ± 0.007	0.555 ± 0.014	0.710 ± 0.065
O_1	LUMO	0.458 ± 0.020	0.693 ± 0.007	0.236 ± 0.014	0.660 ± 0.012	0.855 ± 0.071
O_2	HOMO	0.381 ± 0.003	0.628 ± 0.005	0.247 ± 0.003	0.555 ± 0.014	0.705 ± 0.070
O_2	LUMO	0.460 ± 0.007	0.693 ± 0.007	0.233 ± 0.011	0.660 ± 0.012	0.857 ± 0.066
O_3	HOMO	0.383 ± 0.003	0.628 ± 0.005	0.245 ± 0.005	0.555 ± 0.014	0.703 ± 0.069
O_3	LUMO	0.457 ± 0.016	0.693 ± 0.007	0.236 ± 0.012	0.660 ± 0.012	0.856 ± 0.070
O_4	HOMO	0.391 ± 0.004	0.628 ± 0.005	0.237 ± 0.005	0.555 ± 0.014	0.693 ± 0.073
O_4	LUMO	0.504 ± 0.016	0.693 ± 0.007	0.190 ± 0.012	0.660 ± 0.012	0.820 ± 0.089

Table 4: Estimated reference points and probe scores for the nested baselines $O_1 \subset O_2 \subset O_3 \subset O_4$, reported as mean $\pm$ sample standard deviation (SD) across five seeds.

The richest baseline, O_4 , additionally captures whether these quantities were changing immediately before perturbation. For any pre-bump variable $q(t)$, we define

$$\Delta q_{\text{pre}} = \overline{q(t)}_{-30 \leq t < 0 \text{ ms}} - \overline{q(t)}_{-100 \leq t < -70 \text{ ms}}.$$

We compute this late-minus-early change separately for planar position, velocity, and force, yielding six additional features:

$$\Delta x, \Delta y, \Delta v_x, \Delta v_y, \Delta F_x, \Delta F_y.$$

Thus,

$$O_4 = (O_3, \Delta x, \Delta y, \Delta v_x, \Delta v_y, \Delta F_x, \Delta F_y).$$

These features distinguish a stable pre-bump state from one in which the limb or interface force was already drifting immediately before perturbation, while remaining entirely pre-perturbation.

Full information I . The full information is held fixed across all four specifications. It augments the richest pre-bump state with the experimentally commanded planar perturbation,

$$I = (O_4, b_x, b_y),$$

where (b_x, b_y) is the commanded bump vector. Because $O_1 \subset O_2 \subset O_3 \subset O_4$, every specification satisfies $\sigma(O_j) \subseteq \sigma(I)$. The ceiling $\hat{\rho}_I^2$ is therefore fitted once per target and seed and shared by all four contrasts; only the floor changes between them.

The associated headroom,

$$\hat{\Delta}_j = \hat{\rho}_I^2 - \hat{\rho}_{O_j}^2,$$

measures how much additional predictability of the subsequent force response is available from knowing the perturbation after conditioning on a specified description of the pre-bump body state.

Neural representation. The representation H is the post-perturbation Area 2 population activity extracted from the neural response following bump onset. A decoder-bank probe is trained to predict T_x or T_y from this population representation (spike counts), separately within each recording session with predictions pooled across sessions, producing R_H^2 ; like the ceiling, this probe score is shared by all four contrasts.

This setting differs from the artificial-network experiments because H is a recorded biological population state rather than a deterministic transformation of the measured I . Area 2 activity may contain proprioceptive, biomechanical, or internally generated information that is absent from our behavioral measurements.

What is the scientific question? The raw question

“How well can post-bump Area 2 activity predict the force response?”

is insufficient because part of that response may already be predictable from the state of the hand before the perturbation. The question posed by normalization is instead

“How should neural decodability be interpreted relative to what was already predictable before the bump and what additional predictability is provided by knowing the bump?”

For each target and each choice of (O_j, I, T) , we therefore compare

$$\hat{\rho}_{O_j}^2, \quad \hat{\rho}_I^2, \quad R_H^2.$$

The first quantity is the floor from the pre-bump state alone, the second is the estimated full-information reference, and their difference is the estimated headroom $\hat{\Delta}_j$. The neural probe score is interpreted relative to these quantities rather than on an absolute R^2 scale.

The sequence $O_1 \subset O_2 \subset O_3 \subset O_4$ further provides a sensitivity analysis for the choice of restricted information. O_1 asks whether bump information matters beyond initial hand position alone; O_2 additionally conditions on ongoing motion; O_3 conditions on pre-existing interface force; and O_4 further accounts for recent pre-bump trends. These baseline specifications define distinct estimands. If the interpretation of the neural probe changes across them, that dependence identifies which aspects of the pre-perturbation state are responsible for the apparent decodability.

Result. Under O_4 , floors are near 0.1 and the full-information reference is near 1, so normalized and raw scores nearly coincide: Area 2 activity captures about 72% of the perturbation-driven force response along x and 53% along y under the richest pre-bump baseline O_4 (Table 13). Because neural activity is not a function of the measured inputs, the reference is not an upper bound on the probe score. The interpretation is stable across O_1 – O_4 ; only O_4 , which adds pre-bump trends, raises the floor appreciably.

K NORMALIZING PUBLISHED LLM PROBES

K.1 Common setup

Floors without the model. In every analysis, O is a simple description of the input that uses no language model: character n -grams of a name, word and character n -grams of a statement or conversation, or summaries of an Othello move sequence. The floor therefore needs no access to the model and can be computed for any published result whose inputs, labels, and splits are released.

What the floor measures. Here O is computed from the input string itself, so in principle it holds the same information as the input. What differs is what can be learned from it: the floor is the best readout that a decoder from the bank (Appendix B) can learn from O using only the probe’s training data. The headroom above it is what the model adds, through knowledge from its training or through computation.

Metric. We use R^2 for continuous targets and accuracy for categorical ones. The decoder is chosen on a validation split taken from the training rows and evaluated once on the test rows. A floor can never truly be below chance; an estimated floor well below chance means the readout’s cues invert outside its training data. We do not compute U_H in that case and compare raw scores instead. **Ceiling.** We cannot estimate the ceiling here, because predicting a city’s location or a statement’s truth from its text needs knowledge that the probe’s training data cannot supply. We set it to 1, the highest possible value, which can only lower a positive U_H . In Othello the ceiling is exactly 1, since the board follows from the moves. When a probe scores at or near the floor, we compare the two directly, which does not depend on the ceiling. The floors themselves are lower estimates: a stronger readout could raise them and lower every U_H . The two choices push U_H in opposite directions, so we read it as approximate.

Dataset	Floor (95% CI)	Probe R^2 (7B / 70B)	U_H (7B / 70B)
World places	0.61 (0.59–0.62)	0.88 / 0.91	0.70 / 0.77
US places	0.46 (0.44–0.48)	0.80 / 0.86	0.63 / 0.75
NYC places	0.33 (0.29–0.37)	0.22 / 0.36	−0.17 / 0.04
Historical figures	0.60 (0.59–0.62)	0.79 / 0.83	0.46 / 0.58
Entertainment	0.16 (0.05–0.28)	0.79 / 0.89	0.75 / 0.86
Headlines	0.45 (0.43–0.47)	0.56 / 0.75	0.21 / 0.54

Table 5: Spelling floors for the space and time datasets of Gurnee and Tegmark (2024). The floor is the held-out R^2 of the selected decoder on character n -grams of the entity name; probe scores are the published linear probes on Llama-2-7B / 70B; U_H uses a ceiling of 1, which can only lower a positive U_H (Appendix K.1).

The space and time floors use character 2–4-gram counts (Appendix K.2), and the truth and user-attribute floors word 1–2-gram and character 2–4-gram TF-IDF features; in both cases the linear decoder uses the sparse features directly, and the nonlinear decoders use a 1,024-component truncated SVD.

K.2 Space and time

Published claim. Gurnee and Tegmark (2024) train linear ridge probes on Llama-2 activations at the last token of an entity name to predict its real-world location (world, US, and New York City places) or date (historical figures, entertainment, news headlines), and report high out-of-sample R^2 as evidence of spatial and temporal representations. For New York City, where scores are lowest, they note that the largest model performs best and suggest that larger models may form city-level spatial models.

Choice of (O, I, T) . T is a coordinate or a date, I is the entity string exactly as presented to the model, and O is its character 2–4-gram counts. The floor measures how much of T a readout can learn from spelling alone, for example from state modifiers, borough names, or numbers in the string. The original authors raise a related concern, that a probe could assemble coordinates from simpler features under supervision, and address it by holding out whole regions; the floor quantifies one specific form of it.

Matching the published evaluation. We use the published probe scores (their Table 2: linear probes at 60% layer depth, Llama-2-7B and Llama-2-70B) and re-ran one to confirm that our pipeline matches theirs: for Llama-2-7B on New York City we obtain $R^2 = 0.218$ against the published 0.219. We verified that (i) the floor uses exactly the strings the model received, including rewritten forms such as “Iowa’s Dallas County” and “<creator>’s <title>”, matching token for token on all 186,669 rows; (ii) rows and splits follow the released train–test assignment, which groups New York City locations by complex and entertainment items by creator; and (iii) targets are in their original units, with R^2 computed as in the original evaluation. Ridge regression is fitted on the n -gram counts, and gradient boosting and an MLP on a 1,024-component truncated SVD of them. Confidence intervals come from 1,000 bootstrap resamples of the test set, resampling complexes for New York City and creators for entertainment.

Result (Table 5). For world and US places, spelling accounts for half to two-thirds of the raw R^2 , yet the probes recover 63–77% of the remaining headroom, so most of what remains reflects information beyond the names. For New York City, spelling reaches $R^2 = 0.33$; the Llama-2-70B score (0.36) lies within the floor’s confidence interval and the Llama-2-7B score (0.22) lies below it. At this scale the reported scores cannot separate a city-level spatial representation from lexical cues, and the better relative performance of larger models is also consistent with better use of such cues. This agrees with the original block-holdout analysis, in which probes generalized to held-out boroughs only slightly better than chance. For temporal targets, the probes recover most of the headroom for entertainment ($U_H \geq 0.75$), more than half for historical figures with the larger model (0.58; 0.46 for the smaller), and about half for headlines with the larger model (0.54; 0.21 for the smaller). The entertainment floor is low (0.16), possibly because the released split holds out entire creators, so spelling cues learned for one creator’s titles cannot transfer to another’s. The entertainment result is robust to the floor’s uncertainty: across its interval, U_H lies between 0.71 and 0.78 for Llama-2-7B and between 0.85 and 0.88 for Llama-2-70B. New York City locations are also grouped, yet their floor stays high, plausibly because names carry borough names and numbers shared across complexes.

Floor features	Decoder	Floor acc. (95% CI)	Linear floor acc.	Chance acc.	Occupied acc.	Probe acc. (95% CI)	Linear probe acc.	U_H	U_H linear probe
Black/White/Empty									
O_1 : position	GBM	0.62 (0.61–0.62)	0.56	0.46	0.39	0.89 (0.89–0.90)	0.74	0.72	0.33
O_2 : recency	GBM	0.66 (0.65–0.67)	0.60	0.46	0.47	0.89 (0.89–0.90)	0.74	0.68	0.24
O_3 : placement	GBM	0.80 (0.80–0.81)	0.76	0.46	0.64	0.89 (0.89–0.90)	0.74	0.45	−0.30
Mine/Theirs/Empty									
O_1 : position	MLP	0.62 (0.61–0.63)	0.56	0.46	0.40	0.98 (0.98–0.99)	0.99	0.96	0.96
O_2 : recency	GBM	0.66 (0.66–0.67)	0.57	0.46	0.47	0.98 (0.98–0.99)	0.99	0.95	0.96
O_3 : placement	GBM	0.81 (0.81–0.82)	0.80	0.46	0.65	0.98 (0.98–0.99)	0.99	0.92	0.92

Table 6: Floors, probe accuracy and headroom-normalized decodability for the Black/White/Empty and Mine/Theirs/Empty targets under nested floor feature sets $O_1 \subset O_2 \subset O_3$. The ceiling is exactly 1, since the board is a deterministic function of the move sequence. Occupied acc. is the floor accuracy on occupied squares only. U_H is the fraction of the headroom recovered by the probe.

K.3 Othello world model

Published claim. Li et al. (2022) train probes on Othello-GPT activations to predict the state of each board square, and report 0.88–0.98 accuracy for nonlinear probes and 0.76–0.79 for linear probes on a black/white/empty encoding for synthetic tasks, as evidence of an emergent world model. Nanda et al. (2023) show that a mine/theirs/empty encoding, relative to the player to move, is linearly decodable at higher accuracy.

Choice of (O, I, T) . T is the state of one square, I is the move sequence, and O is a nested set of move summaries that never simulate flips: O_1 records only the move number and the square’s position; O_2 adds whether the square was among the last four moves and by which player, and its distance to the last move; O_3 adds whether the square has ever been played, by which player (relative to the player to move), and how many moves ago. O_3 is thus the board as it would be if no disc were ever flipped. The floor measures how much of the board state is predictable without computing flips; for O_3 , the headroom is exactly what the flip dynamics add.

Model and probes. We use the released Othello-GPT checkpoint of Li et al. (2022) and evaluate it on 1,000 games of uniformly random legal play, the same distribution as its training data. The model predicts a legal next move for 99.94% of positions (58,958/58,992; 98.67% after a pass, which the input does not mark). Floor and probes share one set of rows: eight random positions per game and every square at each (512,000 rows), split by game into 60% fit, 20% selection and 20% test. Rows at the final move (1.6%) are excluded from the probe because the model reads at most 59 moves. For each square we fit the MLP and linear decoders of the decoder bank on the 512-d residual stream after the embedding or after each of the 8 blocks. The layer and each square’s decoder are chosen by McFadden R^2 on the selection games, refitted on fit and selection, and scored once on the test games; 95% CIs resample test games. The selected layer is 5 for black/white/empty and 6 for mine/theirs/empty. Probe accuracy is 0.892 (0.887–0.897) for black/white/empty, with a linear probe at 0.745, and 0.985 (0.980–0.989) for mine/theirs/empty, with a linear probe at 0.986. Our probes are trained on about 4,700 positions per square, fewer than in the original work, so their accuracies, and hence U_H , are lower bounds.

Result (Table 6). The floor rises from 0.62 (O_1) to ≈ 0.80 (O_3). For black/white/empty, our nonlinear probe (0.89) recovers 72% of the headroom above O_1 but only 45% above O_3 : much of its accuracy is available from disc placement without flips. The linear black/white probe (0.74) recovers 33% above O_1 and falls below the O_3 floor ($U_H = -0.30$), so its accuracy is explained by placement alone. For mine/theirs/empty the representation is linear: the linear probe reaches 0.99 and recovers 92% of the headroom even above O_3 (96% above O_1), matching the nonlinear probe. The world-model claim therefore holds most strongly in the relative encoding. The floor depends on which move summaries are allowed; O_3 is the richest summary we tested that still excludes flips. A

Dataset	Decoder	Floor acc. (95% CI)	Linear acc.
cities	Linear	0.46 (0.40–0.52)	0.46
neg_cities	Linear	0.45 (0.39–0.50)	0.45
sp_en_trans	Linear	0.65 (0.52–0.76)	0.65
neg_sp_en_trans	GBM	0.49 (0.38–0.62)	0.56
larger_than	Linear	0.99 (0.97–1.00)	0.99
smaller_than	Linear	0.97 (0.96–0.99)	0.97
cities_cities_conj	Shallow GBM	0.56 (0.50–0.62)	0.54
cities_cities_disj	Linear	0.53 (0.48–0.59)	0.53
companies_true_false	Shallow GBM	0.53 (0.46–0.59)	0.43
common_claim_true_false	Linear	0.62 (0.59–0.65)	0.62
counterfact_true_false	GBM	0.53 (0.52–0.54)	0.50

Table 7: Wording-only floors for the truth datasets. The floor is the test accuracy of the selected decoder on word and character n -grams of the statement; Linear acc. is the linear decoder’s. All datasets are balanced, so chance accuracy is 0.50 (0.51 for `sp_en_trans` and `neg_sp_en_trans`). The ceiling is taken as 1, since the truth value is fixed by the statement and world knowledge.

floor that approximates flips heuristically would be higher and would lower U_H .

K.4 Truth directions

Published claim. Marks and Tegmark (2023) report linear directions in LLM activations that separate true from false statements and transfer across datasets, including to negated statements.

Choice of (O, I, T) . The target T is the binary truth label, I is the full statement, and O is its wording-only representation using n -gram features. The floor therefore measures how well truth can be predicted from surface form alone, without access to model activations or pretrained world knowledge. We evaluate on a random split. For transfer experiments, the floor decoder is trained on the same source dataset as the activation probe and evaluated unchanged on the target dataset; we call this the transfer floor.

Result. Table 7 shows that wording alone is close to chance (0.5) on most truth datasets. For cross-dataset transfer (Table 8), we compare the published probes with this transfer floor.

When trained on `larger_than` and `smaller_than` and tested on `sp_en_trans`, the floor is 0.49, while the published 13B and 70B probes both reach 0.97. Similarly, training on `cities` and `neg_cities` and testing on `neg_sp_en_trans` gives a floor of 0.50, compared with probe accuracies of 0.96 and 0.99. In both cases, the probe transfers substantially better than surface wording alone.

Train → Test	Decoder	Floor acc. (95% CI)	Linear acc.	Published probe acc. (13B / 70B)
<code>larger_than + smaller_than</code> → <code>sp_en_trans</code>	GBM	0.49 (0.44–0.55)	0.50	0.97 / 0.97
<code>cities + neg_cities</code> → <code>neg_sp_en_trans</code>	Shallow GBM	0.50 (0.45–0.55)	0.50	0.96 / 0.99
<code>cities</code> → <code>neg_cities</code> ¹	Shallow GBM	0.35 (0.32–0.37)	0.47	0.75 / 0.37
<code>larger_than</code> → <code>smaller_than</code>	Linear	0.00 (0.00–0.00)	0.00	0.07 / 0.55

Table 8: Wording-only floors for cross-dataset transfer, against the published mass-mean probe accuracy on Llama-2-13B and Llama-2-70B. Floors are trained on the source datasets and scored on the target; chance accuracy is 0.50.

The other two transfers give a different pattern. For `larger_than`→`smaller_than`, the floor is 0.00, while the 13B and 70B probes obtain 0.07 and 0.55. For `cities`→`neg_cities`, the floor is 0.35, while the probes obtain

¹The decoder is selected on held-out source rows, which contain unseen city–country pairs, so no decoder is informative there (shallow GBM pseudo- $R^2 = 0.035$, linear pseudo- $R^2 = 0.001$); `neg_cities` repeats every `cities` pair with “not” inserted and the label flipped, so pairs that shallow GBM memorised during refitting are wrong on the target (0.35, below the 0.50 chance level), while the near-prior linear decoder stays at chance (0.47).

Attribute	Evaluation	Floor acc. (95% CI)	Chance acc.	Published probe acc. (reading / control)	U_H (reading / control)
Age					
	Pooled	0.98 (0.97–0.99)	0.25	0.98 / 0.96	0.11 / −0.78
	User turns	0.98 (0.97–0.99)	0.25	–	–
	GPT-3.5	0.98 (0.97–0.99)	0.25	–	–
	Llama-2-Chat	0.93 (0.89–0.96)	0.25	–	–
	GPT-3.5→Llama-2-Chat	0.88 (0.86–0.90)	0.25	–	–
Gender					
	Pooled	0.92 (0.89–0.94)	0.50	0.94 / 0.91	0.27 / −0.09
	User turns	0.91 (0.88–0.93)	0.50	–	–
	GPT-3.5	0.95 (0.93–0.97)	0.50	–	–
	Llama-2-Chat	0.77 (0.71–0.82)	0.50	–	–
	GPT-3.5→Llama-2-Chat	0.76 (0.73–0.78)	0.50	–	–
Education					
	Pooled	0.97 (0.96–0.98)	0.33	0.96 / 0.93	−0.24 / −1.17
	User turns	0.98 (0.97–0.99)	0.33	–	–
	GPT-3.5	0.97 (0.96–0.98)	0.33	–	–
Socioeconomic status					
	Pooled	0.96 (0.95–0.98)	0.33	0.97 / 0.95	0.14 / −0.43
	User turns	0.95 (0.93–0.97)	0.33	–	–
	GPT-3.5	0.97 (0.95–0.98)	0.33	–	–
	Llama-2-Chat	0.97 (0.95–0.99)	0.33	–	–
	GPT-3.5→Llama-2-Chat	0.90 (0.87–0.92)	0.33	–	–

Table 9: Wording-only floors for user attributes. *Pooled*: both role-play sources, as the published probes were trained. *User turns*: the user’s turns only. *GPT-3.5*, *Llama-2-Chat*: one source. *GPT-3.5 → Llama-2-Chat*: fitted on GPT-3.5 conversations and tested on Llama-2-Chat ones. The released data contain no Llama-2-Chat conversations for education, so those rows are absent. The floor is the test accuracy of the selected decoder (linear in every row). U_H uses a ceiling of 1 and is unstable here because the headroom is small.

0.75 and 0.37. Both floors fall well below chance (0.50) because cues learned on the source dataset invert on the target, so headroom normalization is not applicable (Section 4.1); we compare the probe accuracies directly with the floor and with chance.

K.5 User attributes in conversations

Published claim. Chen et al. (2024) train linear probes on Llama-2-Chat-13B to read a user’s age, gender, education, and socioeconomic status from a conversation, as evidence that the model builds an internal model of its user. They report two kinds of probe. *Reading* probes are applied after a prompt that asks the model about the user, and reach 0.94–0.98. *Control* probes are applied at the end of the user’s message and are used for steering, and reach 0.91–0.96.

Choice of (O, I, T) . T is the user’s attribute, I is the conversation, and O is its surface wording: word 1–2-gram and character 2–4-gram TF-IDF features, with no pretrained model. The linear decoder uses the sparse features directly, and the other decoders use a 1,024-component truncated SVD of them. The linear decoder wins every split.

Matching the published evaluation. We use the released synthetic conversations, in which GPT-3.5 and Llama-2-Chat role-play users with a given attribute (2,900–4,500 conversations per attribute), pooled across both sources as the published probes were trained. We use a stratified 80/20 split, with 20% of the training rows held out for decoder selection. Chance accuracy is the share of the most common class.

Additional evaluations. To see where the cue comes from, we also fit the floor on the user’s turns only, on each source separately, and on GPT-3.5 conversations tested on Llama-2-Chat ones. **Result** (Table 9). Wording alone predicts every attribute almost as well as the published probes. Each published reading probe

lies within the floor’s 95% interval: 0.98 against a floor of 0.98 for age, 0.94 against 0.92 for gender, 0.96 against 0.97 for education, and 0.97 against 0.96 for socioeconomic status. Every control probe falls below the floor. The published accuracies are therefore largely explained by cues in the wording.

The floor from the user’s turns alone (0.91–0.98) matches the floor from the full conversation, so the cues are in what the role-played user writes. Trained on GPT-3.5 conversations and tested on Llama-2-Chat ones, the floor falls to 0.76–0.90, so part of the cue is the generating model’s style rather than the attribute itself.

Reading U_H near a ceiling. With floors of 0.92–0.98, the headroom is only 0.02–0.08, so a difference of a few points between probe and floor becomes a large U_H . The reading probes give U_H of 0.11, 0.27, and 0.14 for age, gender, and socioeconomic status, and -0.24 for education; the control probes give -0.09 to -1.17 . These values mainly reflect how small the headroom is, so we read this analysis by the raw comparison above rather than by U_H .

Caveats. The conversations were written to reflect each attribute, so the label is visible in the text by design. This shows that the published accuracies cannot establish a user model, not that the model lacks one; testing that would need data where the attribute is not cued in the wording.

L CONDITIONAL PROBING AND PREDICTIVE $\mathcal{V}$ -INFORMATION

Hewitt et al. (2021), based on predictive $\mathcal{V}$ -information (Xu et al., 2020), ask how much information about a linguistic property, such as part of speech, a contextual layer of a language model (ELMo, RoBERTa) contains beyond what the non-contextual word embeddings already provide. The baseline B is therefore the embedding layer, and the quantity of interest is the information that a deeper layer H adds to it, $I_{\mathcal{V}}(H \rightarrow Y \mid B) = H_{\mathcal{V}}(Y \mid B) - H_{\mathcal{V}}(Y \mid B, H)$, estimated by training one probe on B and one on the concatenation $[B; H]$. They contrast this with the earlier practice of comparing a probe on H with a probe on B separately. That standalone comparison penalizes H for any information about word identity it has discarded, even when it adds new contextual information. Under conditional probing, deeper layers turn out to carry more information beyond the embeddings than standalone comparisons had suggested.

Our floor plays the role of the baseline defined by them, with two differences: O is a declared information set chosen by the scientific question rather than another layer of the model, and we pair it with a ceiling ρ_I^2 , the predictability supported by the full input.

L.1 Comparing probe normalization with conditional probes

Here, T is the target, O the observed information that defines the floor ρ_O^2 , I the full input, H the representation and B a baseline (here $B = O$). R_X^2 is the held-out R^2 of a probe on X , and $[B; H]$ is the concatenation of B and H . The normalized score, U_H has the form of a skill score. In our method, two reference points, floor and ceiling are what O and I themselves determine about T , not a fixed baseline and a perfect score.

For continuous targets, conditional probing reduces to a difference in R^2 . $\mathcal{V}$ -information measures how much better a probe predicts T when it is given X than when it is given nothing, with prediction quality scored by log-loss. Suppose the probe outputs a Gaussian with mean $\varphi(X)$ and fixed variance $\frac{1}{2}$. Its log-loss is then $-\log p(T \mid X) = (T - \varphi(X))^2 + \text{const}$, so minimizing log-loss is the same as minimizing squared error, and the best achievable log-loss is the probe’s MSE plus a constant. With no input, the best prediction is the mean of T , whose MSE is $\text{Var}(T)$. Subtracting,

$$I_{\mathcal{V}}(X \rightarrow T) = \text{Var}(T) - \text{MSE}_X = \text{Var}(T) R_X^2.$$

The same subtraction with baseline B gives

$$\frac{I_{\mathcal{V}}(H \rightarrow T \mid B)}{\text{Var}(T)} = \frac{\text{MSE}_B - \text{MSE}_{B,H}}{\text{Var}(T)} = R_{B,H}^2 - R_B^2.$$

For continuous targets, conditional probing is therefore just the gain in R^2 from adding H to B , so we report it in R^2 units for the meta-analysis. For Othello the targets are categorical, so we report the corresponding gain in accuracy.

$U_{H|O}$ is **conditional probing with $B = O$ divided by the headroom**. The normalized score U_H compares the probe on H with the floor directly. We define a conditional variant, $U_{H|O}$ that credits H only for what it adds to O :

$$U_{H|O} := \frac{R_{O,H}^2 - \rho_O^2}{\rho_I^2 - \rho_O^2},$$

where $R_{O,H}^2$ is the held-out R^2 of a probe given both O and H . Its numerator is conditional probing with $B = O$, so $U_{H|O}$ is conditional probing normalized by the headroom.

In the population, $U_H \leq U_{H|O} \leq 1$. The two scores share the denominator $\rho_I^2 - \rho_O^2$ and differ only in whether the probe sees H alone or H together with O . With Bayes-optimal predictors, more information never lowers R^2 . Since (O, H) contains H , $R_{O,H}^2 \geq R_H^2$, which gives $U_{H|O} \geq U_H$. Since (O, H) is a function of I when $H = \Phi(I)$, $R_{O,H}^2 \leq \rho_I^2$, which gives $U_{H|O} \leq 1$. The gap

$$U_{H|O} - U_H = \frac{R_{O,H}^2 - R_H^2}{\rho_I^2 - \rho_O^2}$$

is the information about T that O carries and H has discarded. It is zero when H retains everything in O that is relevant to T . In the population, U_H is therefore conservative: it can understate, but not overstate, what H adds beyond O . With finite samples, estimated reference points and a fixed decoder family, either inequality can fail; out of distribution in Table 10, for example, U_H slightly exceeds $U_{H|O}$.

L.2 Meta-analysis.

H is the last layer of the trained xl transformer at the final position. O is the summary statistics of σ^2 (Appendix I). Every probe (for O , H , $[O; H]$) is the linear decoder from the decoder bank of Table 2.

Condition	ID		OOD	
Target	$\log S_1$	τ^2	$\log S_1$	τ^2
Estimated floor $\hat{\rho}_O^2$	0.01 ± 0.01	-0.00 ± 0.00	0.09 ± 0.01	-0.00 ± 0.00
Estimated ceiling $\hat{\rho}_I^2$	0.63 ± 0.02	0.66 ± 0.01	0.41 ± 0.02	0.36 ± 0.03
Estimated headroom $\hat{\Delta} = \hat{\rho}_I^2 - \hat{\rho}_O^2$	0.62 ± 0.01	0.66 ± 0.01	0.33 ± 0.02	0.36 ± 0.03
Hidden probe R_H^2	0.60 ± 0.04	0.64 ± 0.02	0.40 ± 0.02	0.34 ± 0.02
Probe on $[O; H]$, $R_{O,H}^2$	0.59 ± 0.04	0.64 ± 0.02	0.40 ± 0.02	0.34 ± 0.03
Normalized score U_H	0.95 ± 0.04	0.97 ± 0.04	0.96 ± 0.02	0.97 ± 0.02
Conditional normalized score $U_{H O}$	0.95 ± 0.05	0.97 ± 0.04	0.95 ± 0.02	0.95 ± 0.03
Conditional probe ($B = O$), $R_{O,H}^2 - \hat{\rho}_O^2$	0.58 ± 0.04	0.64 ± 0.03	0.31 ± 0.02	0.34 ± 0.02

Table 10: Meta-analysis transformer (xl, last layer, final position): normalized headroom and conditional probing with baseline $B = O$. Every probe, floor and ceiling is the **linear decoder**; the floor and ceiling are estimated from O and I . Mean $\pm$ SD over five seeds.

We found that conditional normalized score is roughly equal to normalized score (Table 10). For both τ^2 and $\log S_1$, U_H and $U_{H|O}$ differ by at most 0.02 in every condition. Raw conditional probe gains mix up the representation and the data and normalizing by headroom separates them. From in distribution to out of distribution, Conditional Probe ($H \mid B = O$) halves from $0.58 \rightarrow 0.31$ for $\log S_1$ and $0.64 \rightarrow 0.34$ for τ^2 . Alone, this indicates that the representation got twice as bad. But the estimated headroom $\hat{\rho}_I^2 - \hat{\rho}_O^2$ also shrinks ($0.62 \rightarrow 0.33$ for $\log S_1$; $0.66 \rightarrow 0.36$ for τ^2), and the normalized score does not fall: $U_{H|O} = 0.95 \rightarrow 0.95$ for $\log S_1$ and $0.97 \rightarrow 0.95$ for τ^2 . The drop therefore reflects information missing from the shifted data, not a worse representation. Conditional probing cannot differentiate without a ceiling score.

L.3 Othello-GPT.

We fit one linear probe per square on the residual stream (block 5 for black/white/empty, block 6 for mine/theirs/empty; 1,000 games split 60/20/20), with O the move-sequence features of rung O_3 (disc placement without flips) and a ceiling of 1, since the board is a function of the moves. Because every probe here, including

Target	Black/White/Empty	Mine/Theirs/Empty
Layer	5	6
Estimated floor (rung O_3)	0.762 (0.754–0.768)	0.803 (0.797–0.808)
Ceiling (exact)	1.000	1.000
Headroom	0.238	0.197
Probe on O	0.768 (0.762–0.774)	0.808 (0.803–0.814)
Hidden probe	0.745 (0.738–0.752)	0.986 (0.981–0.990)
Probe on $[O; H]$	0.767 (0.760–0.773)	0.989 (0.984–0.992)
Normalized score U_H	-0.069	0.929
Conditional normalized score $U_{H O}$	0.021	0.942
Conditional probe ($B = O$), $\text{Acc}_{O,H} - \text{Acc}_O$	-0.001	0.180

Table 11: Othello-GPT: normalized headroom and conditional probing with baseline $B = O$ (rung O_3 : disc placement without flips). Test accuracy of the linear decoder, one probe per square; the ceiling is exactly 1 because the board is a function of the moves. The estimated floor is the linear decoder of the floor ladder (Table 6), fitted once across all squares; the probe on O fits one linear probe per square on the same features and on the probe’s rows, which exclude the final move, so the two differ slightly. The conditional probe is measured from the probe on O ; U_H and $U_{H|O}$ are measured from the estimated floor.

the floor, is linear, the floor (0.76 and 0.80) is slightly lower than the decoder-bank floor of Appendix K. The standalone and conditional scores reach the same verdict (Table 11). For black/white/empty, H adds nothing beyond disc placement: the conditional gain is -0.001 , $U_H = -0.07$ and $U_{H|O} = 0.02$. For mine/theirs/empty, H captures most of what the flips add: the conditional gain is 0.18, $U_H = 0.93$ and $U_{H|O} = 0.94$. The normalized and conditional normalized scores agree in both cases. Therefore, both methods support the conclusion of Section 4.2.3: a linear readout of the absolute board is explained by disc placement, while with mine/theirs labels a linear probe recovers the board state that requires simulating flips.

M ADDITIONAL FIGURES AND TABLES

	$\log S_1$ R^2		τ^2 R^2		$\text{MSE}_{\text{model}}/\text{MSE}_{\text{oracle}}$		MSE
	ID	OOD	ID	OOD	ID	OOD	OOD/ID
xs	0.28 ± 0.20	0.30 ± 0.04	0.22 ± 0.20	0.24 ± 0.03	1.05 ± 0.01	5.05 ± 2.25	28.3
s	0.60 ± 0.04	0.29 ± 0.03	0.54 ± 0.03	0.26 ± 0.02	1.02 ± 0.00	2.44 ± 0.37	14.2
m	0.60 ± 0.03	0.34 ± 0.03	0.60 ± 0.04	0.28 ± 0.03	1.03 ± 0.00	2.58 ± 0.73	14.9
l	0.61 ± 0.02	0.35 ± 0.05	0.58 ± 0.05	0.29 ± 0.05	1.03 ± 0.01	2.23 ± 0.59	12.9
xl	0.64 ± 0.05	0.40 ± 0.02	0.64 ± 0.02	0.34 ± 0.02	1.03 ± 0.00	2.08 ± 0.51	12.0

Table 12: In-distribution (ID) vs. out-of-distribution (OOD) probe R^2 and prediction MSE at the final checkpoint (mean $\pm$ sample standard deviation across five seeds). MSE is measured at the final position; the oracle is the conditional-Bayes estimator $S_2/(S_1 + 1)$, which uses the true τ^2 .

Target	Floor	Ceiling	Headroom	Representation probe	U_H
Neuroscience					
Force x	0.11 ± 0.02	0.99 ± 0.01	0.88 ± 0.03	0.75 ± 0.01	0.72 ± 0.02
Force y	0.09 ± 0.02	1.00 ± 0.00	0.91 ± 0.02	0.57 ± 0.02	0.53 ± 0.03
Chemistry					
HOMO	0.39 ± 0.00	0.63 ± 0.00	0.24 ± 0.00	0.56 ± 0.01	0.69 ± 0.07
LUMO	0.50 ± 0.02	0.69 ± 0.01	0.19 ± 0.01	0.66 ± 0.01	0.82 ± 0.09
Meta-analysis					
$\log S_1$ (ID)	0.01 ± 0.00	0.82 ± 0.01	0.81 ± 0.02	0.64 ± 0.05	0.78 ± 0.06
τ^2 (ID)	0.00 ± 0.00	0.76 ± 0.01	0.76 ± 0.01	0.64 ± 0.02	0.84 ± 0.04
$\log S_1$ (OOD)	0.09 ± 0.01	0.47 ± 0.02	0.37 ± 0.02	0.40 ± 0.02	0.82 ± 0.08
τ^2 (OOD)	0.00 ± 0.00	0.40 ± 0.02	0.40 ± 0.02	0.34 ± 0.02	0.85 ± 0.08
Single-cell					
Synthetic	-0.02 ± 0.02	0.87 ± 0.01	0.89 ± 0.02	0.60 ± 0.03	0.70 ± 0.03
10k human	0.60 ± 0.02	0.97 ± 0.01	0.38 ± 0.02	0.81 ± 0.02	0.57 ± 0.03
20k donors (pooled)	0.13 ± 0.04	0.97 ± 0.00	0.84 ± 0.04	0.57 ± 0.03	0.52 ± 0.03
40k breast	0.15 ± 0.06	0.87 ± 0.02	0.72 ± 0.04	0.35 ± 0.04	0.28 ± 0.05
40k lung	0.06 ± 0.06	0.86 ± 0.02	0.80 ± 0.05	0.44 ± 0.05	0.47 ± 0.05

Table 13: Headroom-normalized decodability across domains (mean $\pm$ sample standard deviation across five seeds). $U_H = (R_H^2 - \rho_O^2)/(\rho_I^2 - \rho_O^2)$ is the fraction of headroom recovered by the representation probe. Neuroscience uses the O_4 pre-bump floor and a decoder-bank probe on Area 2 response counts, fitted within each recording session; meta-analysis uses the exact reference points of Table 3 with final-checkpoint xl last-layer probes; single-cell data use reference points estimated at full sequencing depth and a probe on frozen scGPT representations (64 PCA components); chemistry uses the O_4 physicochemical floor and a probe on the LoRA-fine-tuned TabPFN.