The Subjectivity of Monoculture

Nathanael Jo
MIT EECS
nathanjo@mit.edu

Nikhil Garg
Cornell Tech
ngarg@cornell.edu

Manish Raghavan
MIT Sloan & EECS
mragh@mit.edu

Abstract

Machine learning models – including large language models (LLMs) – are often said to exhibit *monoculture*, where outputs agree strikingly often. Our work begins with a pair of experiments that challenge this narrative. We find that monoculture estimates can virtually disappear based on choices made by the analyst.

Informed by these findings, we theoretically formalize what it means to measure monoculture. We show that monoculture is inherently subjective, relying on two key decisions: *1)* the baseline null model for what "independence" should look like; and *2)* the population of models and items under consideration. We conclude with concrete guidelines for future evaluators to ensure the robustness of their results.

1 Introduction

A growing body of empirical work suggests that AI models, including both predictive and generative models, produce homogeneous outputs. Across domains—including text/image generation and decision making—separately trained models often produce strikingly similar responses [12, 15, 19, 26, 30, 37–39]. Throughout the literature, a common perspective has emerged: Models don't just agree, they agree *too much*, resulting in "algorithmic monoculture" [20].

Why does monoculture matter? In resource allocation settings such as hiring or lending, similar predictive systems may lead institutions to systematically privilege or disadvantage the same groups [17]. In market settings, firms that independently deploy similar pricing algorithms may inadvertently sustain higher prices and harm consumers [18]. More broadly, diversity is often a desirable property of sociotechnical systems because it can improve robustness to error, reflect a plurality of norms or preferences, and promote creative or novel outputs. A large literature has emerged analyzing the downstream consequences of algorithmic monoculture [2, 6, 9, 28, 29, 32].

However, claims about monoculture are by nature relative, because it suggests that outputs are more correlated than one would expect under some ideal world. As we will show, this comparison cannot be made in a vacuum. Instead, it requires two subjective choices made by the analyst, and these choices drive the conclusions they reach about monoculture:

1. What is the baseline (null model) against which agreement is measured?
2. On what population of models and instances do we measure excess agreement?

Baselines for excess agreement. The literature contains two distinct approaches to constructing baselines against which excess agreement is measured. One is to use human subjects [1, 12, 26, 38, 39]. While human baselines are useful for normative comparisons about diversity, they require committing to a reference population and a notion of human diversity.

A second approach, and one we take here, is to measure excess agreement relative to a "null model," under which responses are conditionally independent [15, 16, 19, 37]. A null model is a particular data-generating process that creates an expectation of how much agreement is reasonable. Any observed agreement in excess of the null model is potentially evidence of monoculture. A popular null model

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

is capability-based [15, 18]. In its simplest form, two models with accuracies (p_1, p_2) are viewed as independent if their expected agreement is the chance overlap $p_1p_2 + (1-p_1)(1-p_2)$ on each item.

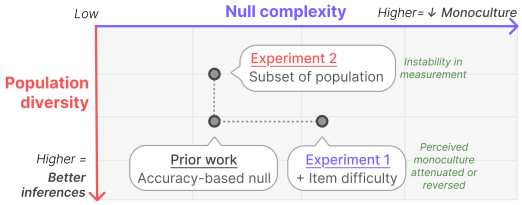


Figure 1: Summary of our main contributions.

Is this the "right" null model? Perhaps—but this baseline might ignore other relevant features such as **question difficulty**: if two models both succeed on easy items and fail on hard ones, their agreement may exceed the $p_1p_2 + (1-p_1)(1-p_2)$ baseline even in the absence of any meaningful dependence. Or perhaps some models perform well on math questions, whereas others excel at writing code. Again, this would lead to excess agreement beyond what the simple baseline would predict (Null complexity in Fig 1). Generally, a null model establishes the structure we *expect* to see in data, giving us a baseline against which to measure *unexpected* correlation.

Complicating things further is that we generally lack "ground truth" information about latent structural properties like item difficulty. Instead, we must infer them from data: e.g., the difficult questions are precisely the ones that many models get wrong. With a small set of models, one might incorrectly attribute all correlation to heterogeneous item difficulty, thereby failing to detect any excess correlation. In other words, latent structural properties like item difficulty or model capability cannot be defined in a context-free way; they are defined only relative to a population of models and items. As a result, even after fixing a null model, conclusions about excess agreement depend critically on the set of models and questions over which it is fit and evaluated (Population diversity in Fig 1).

Our contributions. We begin in Section 2 by demonstrating empirically that inferences about monoculture can change drastically given different choices on null complexity and population diversity. Motivated by these challenges, we present theory to understand the subjectivity of monoculture in Sections 3 and 4. Finally, we propose guidelines and best practices for future evaluations on monoculture in Section 5.[1]

2 Two Empirical Challenges

We first start by demonstrating two empirical challenges, as summarized in Figure 1. Experiment 1 shows that a minor addition to the null model can **substantially attenuate, and in some cases reverse, prior findings**. Experiment 2 shows that inferences of monoculture also depend heavily on the population of models or questions used in the dataset. Paradoxically, **measuring monoculture is hard precisely when there truly is monoculture.** Both suggest that empirical measurements of monoculture are perhaps more fragile than prior work suggests.

2.1 Preliminaries

Our experiments use two notable works on monoculture as baselines: Goel et al. [15] and Kim et al. [19]. As we discuss in the prior section, these works primarily assume a **capability-based null model**. Kim et al. [19] measure agreement rates conditional on both models being incorrect, compared to a null model where all the incorrect choices have an equal probability of being chosen. Meanwhile, Goel et al. [15] measure how much two models agree on their predicted probabilities beyond what would be expected from their marginal accuracies alone. See Appendix C.2 for details.

In contrast to prior works, our experiments include one additional complexity in our null model: **question difficulty**. Intuitively, if two models both succeed on easy items and fail on hard ones, their agreement may exceed the capability-only baseline even if they are truly independent. More generally, our framework can handle alternative null models—e.g., models that account for topic specialization. To operationalize this addition of question difficulty, we use a popular model of capability from psychometrics: *item response theory* (IRT) [13, 22].

[1] Code can be found in `https://github.com/nathanaj99/subjectivity_monoculture/`. Lean verifications for proofs are given in this repository: `https://github.com/nikhgarg/SubjectivityOfMonocultureLean/` [14].

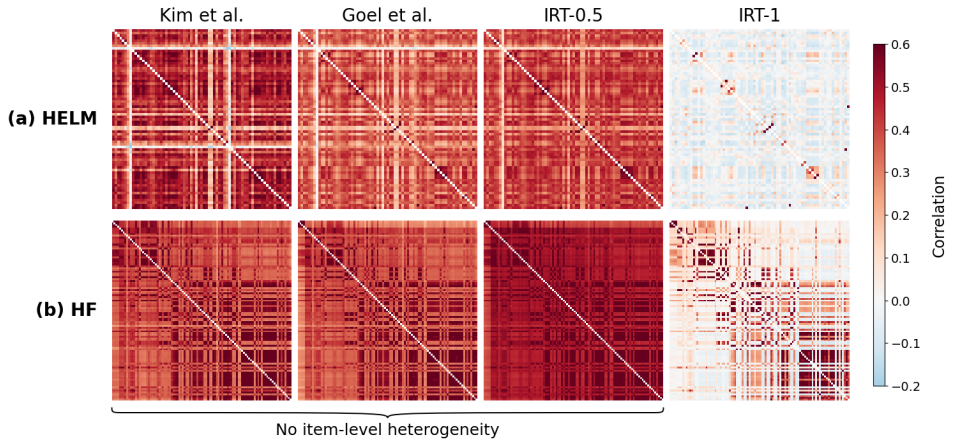


Figure 2: Residual correlation matrices for **HELM** (a) and **HF** (b) (first 100 models), using different nulls: from left to right, a baseline from Kim et al. [19], from Goel et al. [15], from a 1D IRT with no item difficulty, and from a 1D IRT with item difficulties. Each item is a question-answer choice pair. The first three null models **do not** include item heterogeneity, and so the full IRT model is attenuated compared to the rest because item difficulties absorb much of the apparent positive correlation.

The IRT Model. IRT posits that each respondent j possesses a latent ability vector θ_j and that each item i has parameters (a_i, d_i) describing the item's discrimination and difficulty, respectively. The probability of a correct response ($Y_{ij} = 1$) is modeled as an increasing function σ (e.g., logit or probit). We note that several recent works apply IRT to evaluating LLMs, see Appendix A. Multidimensional IRT [31] extends this framework by letting θ_j, a_i lie in $\mathbb{R}^K$. Formally, we have:

Definition 1 (K-dimensional IRT)**.** Fix $m \geq 1$. For $K \in \mathbb{N}$, let $\mathcal{N}_K$ be the set of laws on $\{0,1\}^m$ generated by a K-dimensional IRT model. Conditional on item parameters (a_i, d_i) and model abilities $(\theta_j)_{j=1}^m$, responses are independent across models and satisfy

$$\Pr(Y_{ij} = 1 \mid a_i, d_i, \theta_j) = \sigma(a_i^\top \theta_j + d_i).$$

We set σ to be a probit function. In the context of our experiments (and in general), this model can also capture multiple choice data by encoding correctness at the answer-choice level (i.e., correctly choosing the right answer, or correctly *not* choosing the wrong answer). Beyond multiple choice, we emphasize that **our insights generally hold for more complex outputs**, see Appendix D.1.

We estimate the (a_i, d_i, θ_j) using standard methods, see Appendix C.1. Then, define the following:

$$\hat{p}_{ij} = \Phi\left(\hat{a}_i^\top \hat{\theta}_j + \hat{d}_i\right), \quad R_{ij} := Y_{ij} - \hat{p}_{ij}, \quad \tilde{R}_{ij} := R_{ij} - \frac{1}{n}\sum_{i=1}^{n} R_{ij},$$

where $\hat{p}$ is the estimated probability of correctness, Φ is the cdf of a standard normal, R represents the (itemwise) residuals, and $\tilde{R}$ is the debiased R to account for average calibration error in $\hat{p}$ (see Prop. 7). Finally, we compute the corresponding residual correlation

$$\widehat{\Sigma} := E^{-\frac{1}{2}} \widehat{C} E^{-\frac{1}{2}}, \quad \text{where } C := \frac{1}{n}(\tilde{R})^\top \tilde{R}, \quad E := \text{diag}(\widehat{C}). \tag{1}$$

By construction, $\widehat{\Sigma}$ has unit diagonal and is bounded in $[-1, 1]$. $\widehat{\Sigma}$ **is our measure of excess correlation or monoculture throughout the paper**. If $\widehat{\Sigma}_{k\ell} \gg 0$, then models k and ℓ agree more than we would expect under the null model. We emphasize that our inference procedure is not intended for out-of-sample prediction, but to infer parameters within a fixed population. This is standard in IRT models. *Conditional on a fixed population*, there is virtually no variability ($\pm$1e-4) in estimating (a_i, d_i, θ_j), and accordingly for $\widehat{\Sigma}$.

2.2 ● Experiment 1: Incorporating item difficulty reduces excess correlation

Datasets. We use two datasets following Kim et al. [19]: **HELM** [21] (on MMLU data) and an Open LLM Leaderboard on Hugging Face[2] (hereafter, **HF**). **HELM** comprises $n = 14{,}042$ multiple-choice

[2] `https://huggingface.co/collections/open-llm-leaderboard/open-llm-leaderboard-2`

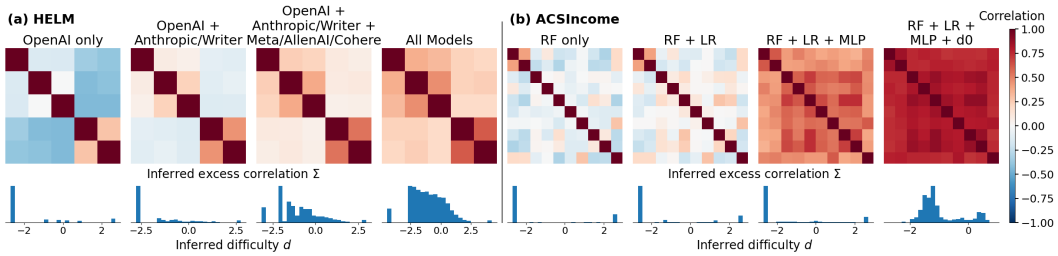


Figure 3: [Top] Inferred excess correlation matrix $\hat{\Sigma}$ from the two-stage procedure, on **HELM** (a) and **ACSIncome** (b) data. For **HELM**, the correlation matrix is shown for OpenAI models only. From left to right, more models are injected into the population for inference, starting from only OpenAI models to all models. For **ACSIncome**, the correlation matrix is shown for random forest models (RF) only, following a similar pattern as **HELM**. [Bottom] Histogram of inferred question difficulties d for all questions in the dataset.

questions spanning 57 academic and professional subjects, evaluated across $m = 72$ language models. These include prominent proprietary models (e.g., GPT and Claude families) alongside open-weight models. **HF** contains $n = 11{,}994$ multiple-choice questions from MMLU-Pro, a more challenging successor to MMLU, with greater emphasis on reasoning and typically 10 answer choices rather than 4. These questions are evaluated across $m = 451$ predominantly open-weight LLMs.

Note on multiple choice treatment. The IRT model in Definition 1 parametrizes the probability of correctness for each question i. In our datasets, however, we have multiple choices that encode more specific information about *when* and *which types* of mistakes occur. We model this richness by expanding each unit i in the IRT model as a **question-answer choice**. For example, in our experiments for **HELM**, with $n = 14{,}042$ questions and four answer choices each, we fit an IRT model with $n = 56{,}168$.

Setup. In addition to the models of Kim et al. [19] and Goel et al. [15], we also fit two 1-dimensional IRT models. First, we consider a restricted model in which item difficulty is fixed across all items ($d_i \equiv 0$), so that variation in performance is explained solely by model ability (call this IRT-0.5). This regime serves as our closest analogue to the baselines, which similarly do not adjust for item-level variation. Second, we fit a full 1-dimensional IRT model with item difficulty parameters, thereby introducing item heterogeneity (IRT-1). All experiments can be executed in minutes using a CPU.

Results. Figure 2 reports the empirical residual correlation matrices $\hat{\Sigma}$ under constant item difficulty (IRT-0.5) and item heterogeneity (IRT-1), for both **HELM** and **HF** (restricted to the first 100 models for visualization). Relative to Goel et al. [15] and Kim et al. [19], the correlation inferences for IRT-0.5 are slightly higher because it is relatively less expressive: IRT-0.5 only accounts for raw accuracies as a baseline, while both previous works incorporate more specific information about disagreement rates. However, crucially, the correlation inferences under IRT-1 are substantially attenuated compared to the other three, which do not model item heterogeneity—some even flipping from strongly positive to slightly negative. This result highlights the role of increasing expressivity: models may appear correlated simply because they succeed or fail on the same easy or hard questions.

In sum, by adding a single change to the null model (question difficulty), our estimates of monoculture are significantly attenuated or reversed. The natural interpretation from Figure 2 is that models dramatically less monoculture than prior works find, because their behavior on questions can be explained by questions being easy or difficult. But as we will show, there is an alternative explanation: because our null model parameters are learned from data, the *set* of LLMs used impacts inferences about those parameters, and therefore, our conclusions about monoculture. We explore this in our next experiment.

2.3 ● Experiment 2: Inferences depend on the population of questions and models

We now fix a null model (IRT-1) and vary the models used in our analysis. As we will show, greater model diversity improves inferences of monoculture, which may indicate that **HELM** suffers from poor diversity. In other words, **measuring monoculture is hard precisely when monoculture exists.**

Datasets. We use two datasets: **HELM** as described above, and **ACSIncome** [11], which is a prediction task to determine whether an individual earns over $50,000. We use this dataset because it allows finer-grained control over model classes and their inductive biases. Specifically, we train 57 models across three classes: 40 random forests (RF), 2 logistic regressions (LR), and 15 multilayer perceptrons (MLP). To generate variation within each class, we vary the number of trees, maximum depth, and splitting criterion for RF; regularization strength for LR; and hidden-layer architecture for MLP. We only train two logistic regressions because variation in parameters does not change the learned predictor. For RF and MLP, we additionally train five independently initialized instances per hyperparameter configuration, introducing variation from stochastic training. We train these models on a 50% training split and evaluate them on the held-out 50%. For each model, we convert held-out predictions into a binary correctness variable, which forms the dataset used in our experiment.

Setup. We begin by restricting attention to a subset of models that are expected to be highly correlated due to shared inductive biases. For **ACSIncome**, we first infer $\widehat{\Sigma}$ using only the random forest models; for **HELM**, we start with only OpenAI models, which likely share underlying architectures and training. We then progressively enlarge the population by adding other models and re-estimating $\widehat{\Sigma}$ at each step. In **ACSIncome**, for example, the experiment involves adding logistic regression (LR), then multilayer perceptrons (MLP), then an artificial set of 25 models where each model has an independent probability of success $p \sim \mathcal{N}(0.7, 0.2)$ (d0). We include this final step to illustrate the case where "ground truth" item difficulties are all identical and independent; in this regime, we can only explain correlation via monoculture. All experiments can be executed in minutes using a CPU.

Results. Figure 3 shows that when we restrict the population to a highly correlated subset of models (left subfigures), inferences about correlation reduce to noise. In **ACSIncome** (b), random forests (RF) do not seem correlated despite the inductive biases shared within the model class, since there are no additional disagreements in the dataset to contextualize the agreements within the RF class. Intuitively, the item difficulties we learn correspond to difficulty *for RF models specifically*. Similarly in **HELM** (a), OpenAI models do not seem correlated despite sharing similar base architectures and training data. These inferences change as we progressively add more diverse models.

In particular, we learn a more general notion of difficulty that holds across a broader class of models (see Figure 3, bottom, which plots the distribution of $\hat{d}$). When the model population is homogeneous, many items are inferred to be very easy or very hard. Without enough disagreement across models, the null model explains agreement by reassigning it to item difficulty rather than to shared model behavior. As more diverse models are introduced, disagreements across models constrain the null fit, spreading $\hat{d}$ distribution more evenly and increasing inferred monoculture $\widehat{\Sigma}$ for the original set.

2.4 Discussion

The experiments above provide two possible explanations for the true structure of monoculture in **HELM**. Experiment 1 suggests that there may truly be no monoculture for **HELM** models if one believes that the IRT is an appropriate null model. Meanwhile, Experiment 2 suggests that there may still be monoculture, but that it may be hidden by the very fact that the dataset is a monoculture. The latter is a plausible explanation especially because **HELM** comprises only closed-source models, while **HF** includes open-source models with greater diversity in contributors and motivations behind developing the models. At this point, we cannot disambiguate which explanation is correct. In Sections 3 and 4, we propose theory that provides a conceptual lens with which to understand these competing explanations. We will also conduct interpretable diagnostics in Section 5.1 to further uncover the true structure of monoculture.

3 • Subjectivity of Null Model

In Experiment 1, we demonstrated that small changes in the null model can drastically change our inferences on monoculture. We now formalize what it means to assess monoculture relative to a null model. To do so, we introduce a *null model of independence*: a family of joint distributions in which all dependence arises through latent parameters.

Setup. We observe, for each item $i \in \{1, \ldots, n\}$ and model $j \in \{1, \ldots, m\}$, an output $Y_{ij} \in \mathcal{Y}$, taken from some unknown probability distribution P. Throughout the paper, we instantiate our theory in the simplest setting, $\mathcal{Y} = \{0, 1\}$, where Y_{ij} encodes correctness. This choice allows us to highlight

the core identification issues without additional complexity, as well as reflecting the paradigm others have studied. **We emphasize that our insights generally hold for more complex outputs**; we show how to generalize our framework to other output spaces in Appendix D.1.

3.1 The Null Model of Independence

We argue that claims of monoculture are only meaningful relative to a chosen *null model* of independence. In the binary setting $\mathcal{Y} = \{0, 1\}$, each model output Y_{ij} is a Bernoulli random variable. Let Θ be an arbitrary parameter space encoding such latent factors (e.g., item properties, model abilities, or both). For each $\theta \in \Theta$, define a *product-Bernoulli model* on $\{0, 1\}^m$ by

$$P_\theta(y_1, \ldots, y_m) \;:=\; \prod_{j=1}^{m} p_{\theta j}^{y_j} (1 - p_{\theta j})^{1-y_j},$$

where each $p_{\theta j} \in [0, 1]$ is the conditional success probability of model j under parameter θ. A null model is the family of distributions $\mathcal{P}_{\text{null}} := \{P_\theta : \theta \in \Theta\}$, and a joint law P is consistent with the null model if it lies in this set. In other words, under the null model, all inter-model agreement can be explained by the shared latent parameter θ together with **conditional independence across models**.

There are numerous "reasonable" choices of Θ. As in Experiment 1, we demonstrated that a simpler null model that only accounts for accuracy might imply a higher observed correlation than if we additionally included item difficulties. In other words, expanding Θ enlarges $\mathcal{P}_{\text{null}}$. In the extreme, if Θ is allowed to be arbitrarily rich, then $\mathcal{P}_{\text{null}}$ can express any distribution on $\{0, 1\}^m$. The next theorem formalizes this observation.[3] All proofs can be found in Appendix B.

Theorem 1. *For any probability distribution P on $\{0, 1\}^m$, there exists a probability measure H on $[0, 1]^m$ and a latent vector $P_i = (P_{i1}, \ldots, P_{im}) \sim H$ such that, conditional on P_i, the coordinates are independent Bernoulli:*

$$Y_{ij} \mid P_i \;\perp\!\!\!\perp\; (j = 1, \ldots, m), \qquad \mathbb{P}(Y_{ij} = 1 \mid P_i) = P_{ij},$$

and the unconditional distribution of Y_i equals P:

$$\mathbb{P}(Y_i = y) \;=\; \int \prod_{j=1}^{m} p_j^{y_j} (1 - p_j)^{1-y_j} \, dH(p_1, \ldots, p_m).$$

Moreover, there exists a discrete H supported on at most 2^m points with this property.

In other words, there exists a sufficiently expressive null that makes the data look independent conditional on the model parameters Θ. From hereon in this section, we suppress Θ and work directly with abstract families of distributions. All results apply uniformly over any choice of Θ.

3.2 Absorption of Excess Correlation

Theorem 1 hints at the central tension of measuring monoculture: **the more structure we allow the null model to capture, the less apparent correlation remains to be explained.** If the null is too simple, it under-explains the data and makes monoculture appear everywhere; if the null is too rich, it explains away all dependence and makes monoculture impossible to detect. To formalize this idea, we introduce a "null ladder" in Definition 2: a nested sequence of increasingly expressive null models. As we climb the ladder, the null becomes more permissive and the gap between the observed law P and the best-fitting null model must shrink.

Definition 2 (Null ladder)**.** Let $\mathcal{P}$ be the set of all laws on $\{0, 1\}^m$. A *null ladder* is a nested sequence $(\mathcal{N}_K)_{K \geq 1}$ with

$$\mathcal{N}_1 \subseteq \mathcal{N}_2 \subseteq \cdots \subseteq \mathcal{N}_K \subseteq \mathcal{N}_{K+1} \subseteq \cdots \subseteq \mathcal{P}.$$

Definition 3 (Excess at level K)**.** Let $D : \mathcal{P} \times \mathcal{P} \to [0, \infty)$ be a discrepancy with $D(P, Q) \geq 0, D(P, P) = 0$, such as TV distance or f-divergence. For $P \in \mathcal{P}$, define the *minimal excess* at level K and the *best-fitting null*:

$$Q_K^\star \in \arg\inf_{Q \in \mathcal{N}_K} D(P, Q), \quad E_{\min}^P(K) \;:=\; D(P, Q_K^\star).$$

[3] Note that this finite mixture representation is a simple instance of a more general phenomenon: exchangeable or symmetric laws can often be written as mixtures of independent structures, as in de Finetti-type theorems.

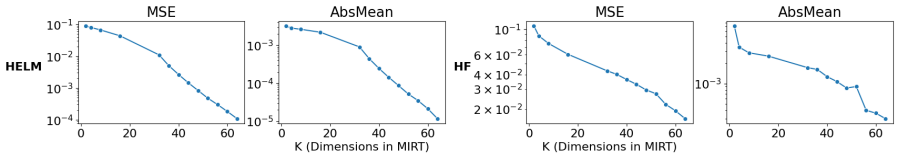


Figure 4: Mean square error and absolute mean of the residual correlations $\widehat{\Sigma}$, as a function of K: dimensions in the multidimensional IRT model, i.e., increading complexity of null model. Left (right) shows results on the **HELM** (**HF**) dataset. As K increases, residual correlations that are unexplained by the null model tend toward zero, meaning that increasingly expressive null models can absorb model correlations.

When the infimum is not attained, let $(Q_K^\star)_{K\geq 1}$ be any sequence in $\mathcal{N}_K$ s.t. $D(P, Q_K^\star) \to E_{\min}^P(K)$.

The quantity $E_{\min}^P(K)$ measures how much of P remains unexplained at the Kth level of the ladder. With this quantity, we can now state our next result, which simply formalizes the intuition that a more expressive model can only explain *more* of the data, never less.

Proposition 2. *For any null ladder $(\mathcal{N}_K)$ and $P \in \mathcal{P}$, $\cdots \geq E_{\min}^P(K) \geq E_{\min}^P(K+1) \geq \cdots$.*

Proposition 2 focuses on the global discrepancy between P and the null, rather than *cross-model correlations* unexplained by the null, which we formalize next. Theorem 3 shows that, provided P can be approximated arbitrarily well along the ladder, these residual covariances vanish as K increases. In other words, a sufficiently rich null model can always reinterpret what appears to be cross-model correlation as arising from latent structure rather than monoculture.

Theorem 3. *Fix a parametrized null ladder $(\mathcal{N}_K)_{K\geq 1}$ and let $Q_K^\star$ be a best-fitting null law at level K as in Definition 3. Suppose that $D(P, Q_K^\star) \xrightarrow[K\to\infty]{} 0$, where D is any discrepancy that controls expectations of bounded functions, such as total variation.*

For each K, let $p_{ij}^{(K)} := \Pr_{Q_K^\star}(Y_{ij} = 1)$ denote the null-implied conditional success probability for item i and model j under $Q_K^\star$. Then, for every fixed pair of distinct models $j \neq \ell$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_P\Big[(Y_{ij} - p_{ij}^{(K)})(Y_{i\ell} - p_{i\ell}^{(K)}) \Big] \right| \xrightarrow[K\to\infty]{} 0.$$

Connection to Empirical Results. One type of null ladder is IRT-0.5 and IRT-1 from Experiment 1. IRT-1 (varied item difficulty) subsumes IRT-0.5 (constant item difficulty) in complexity and is therefore able to absorb more of the pairwise correlations as per Thm 3. Multi-dimensional IRT is another null ladder, and in Figure 4 we show that $\widehat{\Sigma}$ tend to zero as dimension K increases. See Appendix C.3 for details.

4 • Relativity of Population

In Experiment 2 (Section 2.3), we demonstrated that inferences of monoculture are less reliable precisely when there truly is monoculture. We now theoretically formalize these observations. Intuitively, greater diversity results in more informative disagreements between models that better reveal the underlying correlation structure.

4.1 Preliminaries

For notational clarity, we introduce a more specific parametrization of the null model $\Theta = (U, V)$:

$$Y_{ij} \mid (U_i, V_j) \perp\!\!\!\perp \forall j, \quad \Pr(Y_{ij} = 1 \mid U_i, V_j) = f(U_i, V_j), \tag{2}$$

where $U_i \in \mathbb{R}^r$ are item factors, $V_j \in \mathbb{R}^r$ are model factors, $f : \mathbb{R}^r \times \mathbb{R}^r \to (0, 1)$ is a fixed link satisfying mild regularity, and $r \in \mathbb{N}$ is fixed. Note that this includes the IRT models as special cases.

For any choice of item set I and model set J, let $P_{I,J}$ denote the joint distribution of model responses on items I and models J. Let $\mathcal{N}_r(I, J)$ denote the set of distributions that can be explained by an r-dimensional null model (2) using only items I and models J.

Definition 4 (Population-specific target)**.** Given any discrepancy measure $D(P, Q)$ (e.g. TV or an f-divergence), define the *best-fitting null* as

$$Q^\star_{I,J} \in \arg\min_{Q \in \mathcal{N}_r(I,J)} D(P_{I,J}, Q). \tag{3}$$

Under Definition 4, any notion of "excess correlation" built from residuals under $Q^\star_{I,J}$ is a function of (I, J). As before, we use Eq. (1) as our notion of excess correlation.

4.2 Inferences depend on the population

It is almost immediate that changing the population changes the null fit, which also changes the residual covariances, as formalized in the following proposition. In this sense, notions such as difficulty, ability, and excess correlation are not intrinsic properties of isolated models or items, but are defined relative to the population under consideration.

Proposition 4 (Population relativity of the null fit)**.** *Fix r and D. The optimizer $Q^\star_{I,J}$ in* (3) *may vary as I, J change: there may exist item/model sets (I, J) and (I', J') with $I' \subseteq I$ and $J' \subseteq J$ such that the induced marginals do not match: $\big(Q^\star_{I,J}\big)\big|_{(I',J')} \neq Q^\star_{I',J'}$.*

Here, $\big(Q^\star_{I,J}\big)\big|_{(I',J')}$ denotes the marginal predictions for items I' and models J' induced by a null model fit on the larger population (I, J).

What is the "right" population to choose? For example, if all questions are very similar in difficulty, then we cannot expect to learn much about model ability. Similarly, if all models behave nearly identically, we cannot distinguish whether they are truly homogeneous or simply being evaluated on trivially easy or hard questions. This intuition suggests that diversity is useful, as we formalize next.

Heterogeneity measure. Let $(U^\star, V^\star)$ be any fitted null representation achieving $Q^\star_{I,J}$ in Defn. 4. Define summary covariance matrices G of the fitted item and model factors and heterogeneity h:

$$G^\star_U := \sum_{i \in I} U^\star_i (U^\star_i)^\top, \quad G^\star_V := \sum_{j \in J} V^\star_j (V^\star_j)^\top, \quad h(I, J) := \min\big\{\lambda_{\min}(G^\star_U),\ \lambda_{\min}(G^\star_V)\big\}.$$

Here, $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue, which measures the least amount of variation along any direction in the latent space. Thus, $h(I, J)$ captures the weakest direction of variation across items or models under the fitted null. Larger $h(I, J)$ means that items and models explore more independent directions in the latent space (i.e., more heterogeneous).

We note that h is defined using latent factors, so equivalent reparametrizations can change h.[4] A principled way to remove this dependence is to evaluate h at the representation that maximizes it, and we take $(U^\star, V^\star)$ to be this representation henceforth. For example, for unrestricted MIRT-type links $f(u, v) = g(u^\top v)$, a canonical maximizer is the balanced representation ($G^\star_U = G^\star_V$), which is unique up to rotation and gives $h(I, J) = \sigma_r(\Theta^\star)$, the r-th singular value of the fitted logit matrix $\Theta^\star_{ij} = g^{-1}(p^\star_{ij})$. Thus, in the unrestricted model, h depends only on the fitted probabilities, and hence only on the population (I, J). As we show in the next result, heterogeneity improves the local stability (i.e., conditioning) of the null fit.

Theorem 5. *Let $p^\star_{ij} = f(U^\star_i, V^\star_j)$ be the fitted probabilities, where f is twice continuously differentiable. Assume that the null model is locally identifiable up to reparametrization, with identification strength bounded below proportionally to the heterogeneity index $h > 0$ (Assumption 1).*

For $\varepsilon > 0$, define the local ε-equivalence set *of latent representations by*

$$\mathcal{E}_\varepsilon := \Big\{(U, V) \in \mathcal{N} : \sum_{i \in I} \sum_{j \in J} \big(f(U_i, V_j) - p^\star_{ij}\big)^2 \leq \varepsilon^2\Big\},$$

where $\mathcal{N}$ is a sufficiently small neighborhood of the fitted representation. Then $\exists$ a constant $C > 0$ such that, for all sufficiently small ε,

$$\sup_{(U,V) \in \mathcal{E}_\varepsilon} \operatorname{dist}\big((U, V), \mathcal{E}_0\big) \leq \frac{C}{\sqrt{h}}\, \varepsilon,$$

[4] E.g., scaling $U^\star$ by s and $V^\star$ by $1/s$ leaves every $p^\star_{ij}$ unchanged but sends $h \to 0$ as $s \to \infty$.

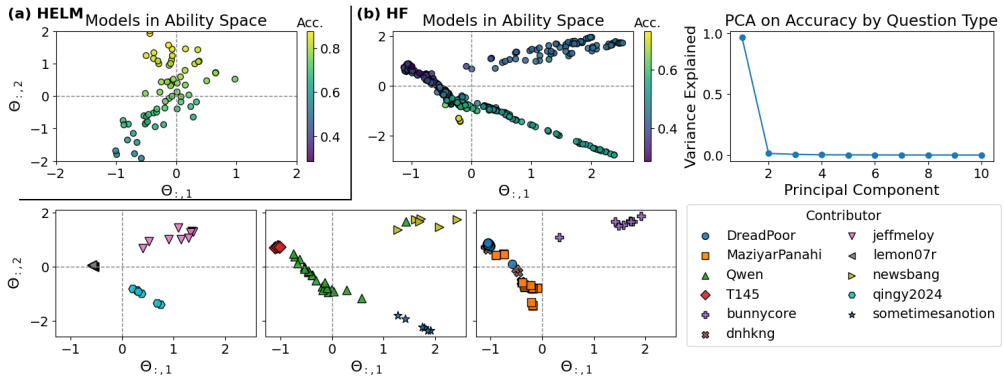


Figure 5: (a) Scatter plot of inferred model ability Θ in a 2-D IRT model, colored by accuracy over entire benchmark dataset (**HELM**). (b; top middle) Same as (a), but for **HF** data. (b; top right) Percentage of variance explained by principal component, when applying PCA on model accuracy stratified by question type. (b; bottom) Scatter plot of inferred ability Θ, disaggregated by contributor.

where dist *is Euclidean distance in* (U, V) *and* $\mathcal{E}_0 := \{(U, V) : f(U_i, V_j) = p^\star_{ij}\ \forall i, j\}$ *is the set of representations that reproduce the fitted probabilities exactly.*

We measure distance to $\mathcal{E}_0$ rather than to $(U^\star, V^\star)$ because null models have exact invariances: e.g., for an IRT link $f(u, v) = g(u^\top v)$, replacing (U_i, V_j) by $(AU_i, A^{-\top}V_j)$ for any invertible A leaves every probability unchanged. In that case, $\mathcal{E}_0$ consists exactly of these rescalings and rotations of $(U^\star, V^\star)$, and so our theorem bounds the fit *after optimally re-aligning its latent coordinates*. Note that Assumption 1 holds automatically for IRT/MIRT, see Lemma 8.

Intuitively, if many items (or models) are similar, the null model has substantial freedom: it can shift latent factors across similar directions without meaningfully changing its predicted probabilities. Since our inferences about monoculture depend on these latent factors, the fact that there exists many factor representations that yield similar observed distributions means that we cannot make strong claims about monoculture. However, when items and models are diverse, this flexibility disappears: the null parameters become locally well-identified, and the fitted representation is more stable. Thus, having datasets and models with heterogeneity is itself helpful in measuring monoculture.

Connection to Empirical Results. In the regime where models are homogeneous from Experiment 2, Thm 5 suggests that many possible (U, V) values can plausibly explain the data. Our particular inference procedure happened to explain the observed agreement by pushing question difficulties to be extremely hard or easy, rather than recovering the true correlation structure. However, greater diversity improves the conditioning of our inferences as per Thm 5.

5 What should we do in practice?

We now provide guidelines for evaluators to better analyze monoculture in practice.

Guidelines

1. Select and justify the following choices:

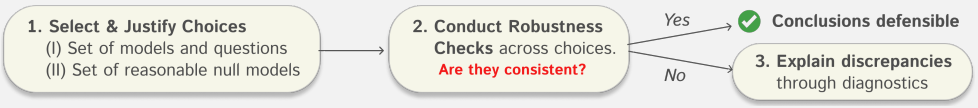
   (I) **The set of models and questions to ensure a diverse and representative sample.** For models, use additional information about contributors and architectures. For questions, ensure that they encompass the various skills within the scope of the benchmark.

   (II) **The set of reasonable null models.** Deciding a null baseline should be a *theory-driven process* rather than a data-driven one, informed by one's beliefs about the data generating process and how models behave. Evaluators should decide **what variation should the null be allowed to explain away.** Should agreement induced by overall model ability, item difficulty, or shared architectures and training paradigms be treated as expected or as part of the phenomenon of interest? This decision involves various factors, such as:

   (a) *What level of granularity is appropriate?*

(b) *What assumptions of exchangeability or conditional independence are defensible?*
(c) *How much flexibility should the null have?*

2. Conduct robustness checks: **Compare monoculture results across the set of reasonable null models.** If the results do not change across null models, then the conclusions are defensible and valid. If results differ (as in Section 2.2), go to Step 3. This principle is similar to best practices in empirical work, where robustness checks test whether or not results are valid against different assumptions about the world/data generating process.
3. Explain discrepancies: **Run appropriate diagnostics for further insights on correlation structure.** See the next section for an example of finding interpretable explanations.



5.1 Our empirical findings as an exercise

Experiment 1 and Experiment 2 are instances of the first two steps of this guideline. We first selected two reasonable null models and several sets of models (Step 1). Then, we inferred monoculture based on these decisions as robustness checks (Step 2). However, we found substantial inconsistencies, especially between our two null models. What remains to be done is Step 3 (explaining discrepancies). One idea to explain discrepancies is to find interpretable structures among models. To do so, we fit a $K = 2$ IRT model to both **HELM** and **HF** and inspect the geometry of the resulting latent spaces.

Our results are shown in Figure 5. For **HELM**, we find that model abilities collapse onto a single dimension that is highly correlated with overall accuracy (subfig (a)), indicating that the 1D IRT may be a reasonable descriptive model of the dataset. However, this finding is not evidence that **HELM** models truly behave independently, as one would conclude from Figure 2(a, right). Rather, it suggests that **HELM** models have limited coverage for detecting dependence beyond a single ability dimension. If **HELM** models are all produced by similar training pipelines or architectures, then we would need additional diversity (e.g., open-source models) in order to reveal what kinds of questions only this family of models systematically gets right or wrong. In other words, while the 1D IRT is sufficient in characterizing the models in **HELM**, we still cannot answer if there is truly monoculture.

HF (subfig (b)) presents a more definitive picture. Model abilities are more dispersed in the two-dimensional latent space and form two distinct clusters that are not strongly correlated with overall accuracy (top middle plot). A natural hypothesis is that these clusters correspond to specialization (e.g., mathematical ability vs reading comprehension). However, a PCA analysis of model accuracies stratified by question type does not support this explanation (top right plot). Instead, the clustering appears to align more closely with model provenance, including common contributors, model families, or development pipelines (bottom plots). This suggests that a 2D (or higher) IRT model is **not appropriate** because the additional structure may be reflected by other sources of heterogeneity like model provenance. We can therefore conclude that there exists some monoculture structure (as in Fig 2, right), because it cannot be otherwise explained by an appropriate richer null model.

5.2 Conclusion

We have shown that inferences of monoculture are sensitive to two key choices: *(1)* the baseline null model (Fig 2, Section 3); and *(2)* the population of models and questions (Fig 3, Section 4). This subjectivity means that robustness checks should be a best practice to ensure that claims of monoculture are consistent across various modeling assumptions, as we establish in Section 5.

Limitations & Future Work. While our theoretical framework extends to richer outputs like free-form text, it remains unclear what a suitable null model would look like in these settings. Building such models and empirically testing for monoculture are promising directions for future work.

Acknowledgments

NG is supported by NSF CAREER IIS-2339427, NASA, the William T. Grant Foundation, and research awards from the Cornell Tech Urban Tech Hub, Google, Mastercard, and Amazon. NJ and MR are supported by the MIT Generative AI Impact Consortium (MGAIC). All opinions are those of the authors and not of any funding agencies, supporters, or other institutions.

References

[1] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425, 2024.

[2] Jackie Baek, Hamsa Bastani, and Shihan Chen. Hiring under congestion and algorithmic monoculture: Value of strategic behavior. *arXiv preprint arXiv:2502.20063*, 2025.

[3] Emily Black, Klas Leino, and Matt Fredrikson. Selective ensembles for consistent predictions. In *International Conference on Learning Representations*, 2022.

[4] Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 850–863, 2022.

[5] Emily Black, Zifan Wang, and Matt Fredrikson. Consistent counterfactuals for deep models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=St6eyiTEHnG`.

[6] Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems*, 35:3663–3678, 2022.

[7] Jane Castleman, Nimra Nadeem, Tanvi Namjoshi, and Lydia Liu. Rethinking math benchmarks: Implications for AI in education. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 66–82. PMLR, 2025. URL `https://proceedings.mlr.press/v273/castleman25a.html`.

[8] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2144–2155. PMLR, 2021. URL `https://proceedings.mlr.press/v139/coston21a.html`.

[9] Kathleen Creel and Deborah Hellman. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1): 26–43, 2022. doi: 10.1017/can.2022.3.

[10] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL `https://jmlr.org/papers/v23/20-1335.html`.

[11] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

[12] Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290.

[13] Susan E. Embretson and Kim A. Diehl. Item response theory. In *Encyclopedia of Psychology*, volume 4, pages 378–380. American Psychological Association and Oxford University Press, 2000. doi: 10.1037/10519-153. URL `https://psycnet.apa.org/doi/10.1037/10519-153`.

[14] Nikhil Garg. Econcslib: Ai-assisted lean formalization for economics & computation research. *arXiv preprint arXiv:2606.13306*, 2026. URL https://arxiv.org/abs/2606.13306.

[15] Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 19621–19678. PMLR, 2025. URL https://proceedings.mlr.press/v267/goel25b.html.

[16] Mila Gorecki and Moritz Hardt. Monoculture or multiplicity: Which is it? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5456. URL https://papers.neurips.cc/paper_files/paper/2025/hash/ef4f4a6beb8b14b2d70a7ef5b386375d-Abstract-Conference.html.

[17] Shomik Jain, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson. Algorithmic pluralism: A structural approach to equal opportunity. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 197–206, 2024.

[18] Nathanael Jo, Ashia Wilson, Kathleen A Creel, and Manish Raghavan. Homogeneous algorithms can reduce competition in personalized pricing. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3365. URL https://papers.nips.cc/paper_files/paper/2025/hash/91625e1de0ab1c8f2c0bb2c7742d5321-Abstract-Conference.html.

[19] Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 30038–30066. PMLR, 2025. URL https://proceedings.mlr.press/v267/kim25e.html.

[20] Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021. doi: 10.1073/pnas.2018340118.

[21] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW.

[22] Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980. URL https://www.ets.org/research/policy_research_reports/publications/book/1980/jexj.html.

[23] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 34303–34326. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

[24] Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42, 2019. doi: 10.1016/j.artint.2018.09.004.

[25] Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International conference on machine learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6765–6774. PMLR, 2020. URL https://proceedings.mlr.press/v119/marx20a.html.

[26] Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Feiz5HtCD0.

[27] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pages 809–818. PMLR, 2020. URL https://proceedings.mlr.press/v124/pawelczyk20a.html.

[28] Kenny Peng and Nikhil Garg. Monoculture in matching markets. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2605. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/95249d42f559a0cfaf282fdf26fe2e69-Abstract-Conference.html.

[29] Kenny Peng and Nikhil Garg. Wisdom and foolishness of noisy matching markets. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 675–675, 2024.

[30] Manish Raghavan. Competition and diversity in generative AI. *Management Science*, 2026. doi: 10.1287/mnsc.2025.02761. URL https://pubsonline.informs.org/doi/10.1287/mnsc.2025.02761. Published online October 16, 2026.

[31] Mark D. Reckase. Multidimensional item response theory. In *Handbook of Statistics*, volume 26, pages 607–642. Elsevier, 2006. doi: 10.1016/S0169-7161(06)26018-8. URL https://www.sciencedirect.com/science/article/abs/pii/S0169716106260188.

[32] Kevin Ren, Manish Raghavan, and Nikhil Garg. Hitting a moving target: Test-time adaptation for AI text detection under continual distribution shift. In *Advances in Neural Information Processing Systems*, 2026. URL https://arxiv.org/abs/2606.25152. To appear.

[33] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3(10): 896–904, 2021. doi: 10.1038/s42256-021-00396-x.

[34] Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, 2021. doi: 10.18653/v1/2021.acl-long.346. URL https://aclanthology.org/2021.acl-long.346/.

[35] Mara Schilling-Wilhelmi, Nawaf Alampara, and Kevin Maik Jablonka. Lifting the benchmark iceberg with item-response theory. In *AI for Accelerated Materials Design-ICLR 2025*, 2025. URL https://openreview.net/forum?id=ZyVQqK7mcP.

[36] Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, 2022.

[37] Connor Toups, Rishi Bommasani, Kathleen Creel, Sarah Bana, Dan Jurafsky, and Percy S Liang. Ecosystem-level analysis of deployed machine learning reveals homogeneous outcomes. *Advances in Neural Information Processing Systems*, 36:51178–51201, 2023. doi: 10.52202/075280-2228. URL https://papers.nips.cc/paper_files/paper/2023/hash/a0b1082fc7823c4c68abcab4fa850e9c-Abstract-Conference.html.

[38] Johan Ugander and Ziv Epstein. The art of randomness: Sampling and chance in the age of algorithmic reproduction. *Harvard Data Science Review*, 6(4), 2024. doi: 10.1162/99608f92.f5dcab1a. URL https://hdsr.mitpress.mit.edu/pub/x5yq8vmk.

[39] Fan Wu, Emily Black, and Varun Chandrasekaran. Generative monoculture in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yZ7sn9pyqb.

[40] Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in benchmarks? rethinking large language model benchmarking with item response theory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):35085–35093, 2026. doi: 10.1609/aaai.v40i41.40814. URL https://ojs.aaai.org/index.php/AAAI/article/view/40814.

A Other Related Work

Model Evaluations using IRT. While our results apply broadly to any choice of null models, in this work we often use IRT models as a null because they incorporate item-level heterogeneity. Many recent works have used IRT in their evaluations [7, 23, 24, 34, 35, 40], acknowledging that questions vary in characteristics and some may be more informative than others. However, all of these works use IRT in the context of evaluating model capabilities rather than, as we do, inter-model correlations.

Model Multiplicity. Model multiplicity suggests that highly expressive models admit many different solutions that are consistent with the same training data, often reflecting different inductive biases or arbitrary choices among equivalent hypotheses [4, 25]. There have been many works that seek to characterize or mitigate multiplicity across a variety of contexts [3, 5, 8, 10, 27, 33, 36]. In these works, multiplicity is typically framed as a property of the *model class*: more flexible hypotheses allow for a greater number of solutions. However, our results suggest another perspective in which multiplicity is also shaped by *the population of models used for evaluation*. We discuss this further in Section 5.

B Additional Theorems and Proofs

B.1 Additional Results

Lemma 6 (Nestedness of ability nulls)**.** *For all K, we have $\mathcal{N}_K \subseteq \mathcal{N}_{K+1}$.*

Proof. Embed the K-dimensional latent θ_i into $\mathbb{R}^{K+1}$ by $\theta'_i = (\theta_i, 0)$ and extend $a'_j = (a_j, 0)$. Then $\sigma(a_i'^{\top}\theta'_j + b_i) = \sigma(a_i^{\top}\theta_j + b_i)$, so every law representable at dimension K is also representable at $K{+}1$. □

Proposition 7 (Debiasing residuals removes calibration error)**.** *Let $Y_{ij} \in \{0, 1\}$ denote the observed response for item $i \in \{1, \ldots, n\}$ and model/respondent $j \in \{1, \ldots, m\}$. Assume the item rows $Y_i = (Y_{i1}, \ldots, Y_{im})$ are i.i.d. from some law P. For each (i, j), let $p^\star_{ij} := \Pr(Y_{ij} = 1)$ denote the* true *success probability and let $\hat{p}_{ij} \in (0, 1)$ be any (possibly misspecified) model-based estimate of $p^\star_{ij}$.*

Define the raw residual

$$e_{ij} \;:=\; Y_{ij} - \hat{p}_{ij},$$

and, for a fixed model j, define the sample mean residual across items

$$\bar{e}_{\cdot j} \;:=\; \frac{1}{n}\sum_{i=1}^{n} e_{ij} \;=\; \frac{1}{n}\sum_{i=1}^{n}(Y_{ij} - \hat{p}_{ij}).$$

Then the population mean of the raw residual equals the population calibration error:

$$\mathbb{E}[e_{ij}] \;=\; \mathbb{E}[Y_{ij} - \hat{p}_{ij}] \;=\; \mathbb{E}[p^\star_{ij} - \hat{p}_{ij}].$$

Since the response rows are i.i.d., writing $p_j = \Pr(Y_{ij} = 1)$,

$$\bar{e}_{\cdot j} - \frac{1}{n}\sum_{i}(p_j - \hat{p}_{ij}) = \frac{1}{n}\sum_{i} Y_{ij} - p_j \xrightarrow{p} 0 \qquad \text{as } n \to \infty.$$

Thus the mean residual tracks the empirical average calibration error, even when the fitted probabilities depend on the sample.

Consequently, the debiased *residual*

$$\tilde{e}_{ij} \;:=\; e_{ij} - \bar{e}_{\cdot j} \;=\; (Y_{ij} - \hat{p}_{ij}) - \frac{1}{n}\sum_{i'=1}^{n}(Y_{i'j} - \hat{p}_{i'j})$$

satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{e}_{ij} = 0$$

exactly. Moreover,

$$\tilde{e}_{ij} = \xi_{ij} - \bar{\xi}_{\cdot j} + \delta_{ij} - \bar{\delta}_{\cdot j},$$

i.e. it removes the average calibration error but retains centered itemwise misspecification.

In particular, sample covariances (or correlations) computed from the debiased residuals $\tilde{e}_{ij}$ estimate the residual dependence *across models, relative to the fitted probabilities; demeaning removes average calibration error but may leave dependence induced by item-specific misspecification*

B.2 Proofs

Proof of Theorem 1.

Proof. Define a discrete probability measure $H^\star$ on the vertices $\{0,1\}^m \subset [0,1]^m$ by placing mass $P(y)$ at the point $p = y$: $H^\star(\{y\}) = P(y)$ for each $y \in \{0,1\}^m$. Given $P_i \sim H^\star$, define the conditional law of Y_i by independent Bernoulli coordinates with success probabilities P_{ij}:

$$\mathbb{P}(Y_{ij} = 1 \mid P_i = p) = p_j, \qquad j = 1, \dots, m.$$

If $P_i = y \in \{0,1\}^m$, then each Y_{ij} is almost surely equal to y_j, so

$$\mathbb{P}(Y_i = y \mid P_i = y) = \prod_{j=1}^{m} y_j^{y_j} (1 - y_j)^{1 - y_j} = 1.$$

Integrating over $H^\star$ gives, for any $y \in \{0,1\}^m$,

$$\mathbb{P}(Y_i = y) = \int \mathbb{P}(Y_i = y \mid P_i = p)\, dH^\star(p) = \sum_{y' \in \{0,1\}^m} \mathbb{I}\{y' = y\} P(y') = P(y).$$

Thus P admits the desired mixture representation with $H = H^\star$. By construction, $H^\star$ is supported on at most 2^m points. □

Proof of Proposition 2.

Proof. Since the null ladder is nested, we have $\mathcal{N}_K \subseteq \mathcal{N}_{K+1}$ for every K. Therefore, minimizing over $\mathcal{N}_{K+1}$ gives at least as much flexibility as minimizing over $\mathcal{N}_K$. Hence,

$$E^P_{\min}(K+1) = \inf_{Q \in \mathcal{N}_{K+1}} D(P, Q) \le \inf_{Q \in \mathcal{N}_K} D(P, Q) = E^P_{\min}(K).$$

Thus $E^P_{\min}(K)$ is nonincreasing in K. □

Proof of Theorem 3.

Proof. Fix K and distinct $j \neq \ell$. Define the function $g_K : \{0,1\}^{n \times m} \to \mathbb{R}$ by

$$g_K(y) \;:=\; \frac{1}{n} \sum_{i=1}^{n} (y_{ij} - p_{ij}^{(K)})(y_{i\ell} - p_{i\ell}^{(K)}).$$

Since each $y_{ij} \in \{0,1\}$ and $p_{ij}^{(K)} \in [0,1]$, we have $|y_{ij} - p_{ij}^{(K)}| \le 1$, hence

$$|g_K(y)| \le \frac{1}{n} \sum_{i=1}^{n} 1 = 1 \quad \text{for all } y,$$

so $\|g_K\|_\infty \le 1$.

By definition of the null model in Section 3, for each item i,

$$\mathbb{E}_{Q^\star_K}\Big[(Y_{ij} - p_{ij}^{(K)})(Y_{i\ell} - p_{i\ell}^{(K)})\Big] = \mathbb{E}_{Q^\star_K}[Y_{ij} - p_{ij}^{(K)}]\, \mathbb{E}_{Q^\star_K}[Y_{i\ell} - p_{i\ell}^{(K)}] = 0,$$

because $\mathbb{E}_{Q^\star_K}[Y_{ij} - p_{ij}^{(K)}] = \Pr_{Q^\star_K}(Y_{ij} = 1) - p_{ij}^{(K)} = 0$. Averaging over i gives $\mathbb{E}_{Q^\star_K}[g_K(Y)] = 0$.

Therefore,

$$|\mathbb{E}_P[g_K(Y)]| = \left|\mathbb{E}_P[g_K(Y)] - \mathbb{E}_{Q_K^\star}[g_K(Y)]\right|.$$

Now use the standard total variation inequality: for any two probability laws μ, ν on a finite measurable space and any f with $\|f\|_\infty \le 1$,

$$\left|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\right| \le 2\,\mathrm{TV}(\mu, \nu).$$

Applying this with $\mu = P$, $\nu = Q_K^\star$, and $f = g_K$ yields

$$|\mathbb{E}_P[g_K(Y)]| \le 2\,\mathrm{TV}(P, Q_K^\star).$$

Finally, note that $\mathbb{E}_P[g_K(Y)]$ is exactly the left-hand side quantity:

$$\mathbb{E}_P[g_K(Y)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_P\Big[(Y_{ij} - p_{ij}^{(K)})(Y_{i\ell} - p_{i\ell}^{(K)})\Big].$$

The claimed bound follows, and since $\mathrm{TV}(P, Q_K^\star) \to 0$ by assumption, the left-hand side converges to 0. □

Remark 1. Theorem 3 is stated in terms of the model-implied probabilities $p_{ij}^{(K)}$. In practice we use fitted probabilities $\hat{p}_{ij}^{(K)}$ obtained from the K-dimensional IRT fit. If the fitted probabilities are uniformly consistent for the model-implied probabilities and the corresponding empirical residual moments obey a law of large numbers in the stated sample-size regime, then the corresponding sample residual covariances inherit the same limiting zero-residual conclusion. The conclusion also holds for correlations if the corresponding residual variances stay bounded away from zero.

Proof of Proposition 4.

Proof. The key observation is that the best-fitting null model is defined *relative to the population* through the optimization problem (3). Changing the item or model set changes both the feasible set $\mathcal{N}_r(I, J)$ and the objective being minimized.

Fix (I, J) and suppose $Q_{I,J}^\star$ is a minimizer of $D(P_{I,J}, Q)$ over $\mathcal{N}_r(I, J)$. By construction, the parameters (U, V) underlying $Q_{I,J}^\star$ must jointly explain *all* item–model interactions in (I, J) using a shared r-dimensional representation. In particular, the fitted parameters for items in I' and models in J' are influenced by the presence of additional items $I \setminus I'$ and models $J \setminus J$ through this shared low-dimensional constraint.

Now consider fitting the null model directly on the restricted population (I', J'). The optimizer $Q_{I',J'}^\star$ is free to choose parameters that best explain $P_{I',J'}$ alone, without needing to simultaneously accommodate the behavior of items or models outside this subset. Unless the restriction of $Q_{I,J}^\star$ to (I', J') already happens to be optimal for this smaller problem—which would require a special compatibility condition—there is no reason for the induced marginal $\big(Q_{I,J}^\star\big)\big|_{(I',J')}$ to coincide with $Q_{I',J'}^\star$.

Therefore, there exist populations (I, J) and strict subsets (I', J') for which

$$\big(Q_{I,J}^\star\big)\big|_{(I',J')} \neq Q_{I',J'}^\star,$$

establishing that the null fit, and hence any residual-based notion of excess correlation, depends on the chosen population. □

Local identifiability. Write $\Theta = (U, V)$, $\|\Theta\|^2 := \sum_{i\in I}\|U_i\|^2 + \sum_{j\in J}\|V_j\|^2$, and $p(\Theta) := (f(U_i, V_j))_{i\in I, j\in J}$ for the prediction map, with Jacobian $\mathrm{D}p$. For a perturbation $(\Delta U, \Delta V)$, the linearized change in predictions at $(U^\star, V^\star)$ is

$$\Delta p_{ij}^{\mathrm{lin}} := \langle \nabla_u f(U_i^\star, V_j^\star), \Delta U_i\rangle + \langle \nabla_v f(U_i^\star, V_j^\star), \Delta V_j\rangle.$$

Assumption 1 (Local identifiability up to reparametrization)**.** $\mathrm{D}p$ has constant rank near $(U^\star, V^\star)$, and there is $c > 0$ such that

$$\sum_{i,j}\big(\Delta p_{ij}^{\mathrm{lin}}\big)^2 \;\ge\; c\,h\left(\sum_i \|\Delta U_i\|^2 + \sum_j \|\Delta V_j\|^2\right)$$

for every perturbation $(\Delta U, \Delta V)$ orthogonal to $\ker \mathrm{D}p(U^\star, V^\star)$.

The kernel of $\mathrm{D}p$ consists of perturbations that leave every prediction unchanged to first order; under constant rank, these are exactly the directions along $\mathcal{E}_0$, i.e., reparametrizations of the null. Assumption 1 therefore says that item and model perturbations cannot cancel in the predictions unless they amount to a reparametrization. This is not implied by smoothness alone, but it holds automatically for IRT/MIRT links.

Lemma 8 (IRT/MIRT links)**.** *Let $f(u, v) = g(u^\top v)$ with g twice continuously differentiable and $g' > 0$ (e.g., logistic or probit), and suppose $h > 0$. Then Assumption 1 holds with $c = m^2$, where $m := \min_{i,j} g'\big((U_i^\star)^\top V_j^\star\big) > 0$. The same holds when some coordinates of every V_j are fixed constants, e.g., $V_j = (\theta_j, 1)$ to encode item intercepts.*

Proof. Write $\ell_{ij} := \langle V_j^\star, \Delta U_i\rangle + \langle U_i^\star, \Delta V_j\rangle$, so that $\Delta p_{ij}^{\mathrm{lin}} = g'_{ij}\,\ell_{ij}$ with $g'_{ij} := g'\big((U_i^\star)^\top V_j^\star\big) \geq m$. Hence $\sum_{i,j}(\Delta p_{ij}^{\mathrm{lin}})^2 \geq m^2 \sum_{i,j} \ell_{ij}^2$, and $\ker \mathrm{D}p = \ker \ell$.

Kernel and constant rank. Stack $U_i^\star$ and $V_j^\star$ as the rows of $U^\star \in \mathbb{R}^{|I|\times r}$ and $V^\star \in \mathbb{R}^{|J|\times r}$, so that $G_U^\star = (U^\star)^\top U^\star$ and $G_V^\star = (V^\star)^\top V^\star$; since $h > 0$, both matrices have full column rank. If $\ell \equiv 0$, i.e., $\Delta U (V^\star)^\top + U^\star \Delta V^\top = 0$, multiplying on the right by $V^\star (G_V^\star)^{-1}$ gives $\Delta U = U^\star B^\top$ for some $B \in \mathbb{R}^{r\times r}$, and then $U^\star\big(B^\top (V^\star)^\top + \Delta V^\top\big) = 0$ forces $\Delta V = -V^\star B$. Thus

$$\ker \mathrm{D}p = \big\{(\Delta U_i, \Delta V_j) = (BU_i^\star,\, -B^\top V_j^\star) :\ B \in \mathbb{R}^{r\times r}\big\},$$

the tangent directions of the reparametrizations $(AU_i, A^{-\top} V_j)$ at $A = I$. This kernel has dimension r^2 at every full-rank (U, V), an open set, so $\mathrm{D}p$ has constant rank near $(U^\star, V^\star)$.

Lower bound. A perturbation is orthogonal to the kernel iff $\sum_i \langle \Delta U_i, BU_i^\star\rangle = \sum_j \langle \Delta V_j, B^\top V_j^\star\rangle$ for all B, i.e., iff

$$S := \sum_i \Delta U_i (U_i^\star)^\top = \sum_j V_j^\star \Delta V_j^\top.$$

Expanding the square,

$$\sum_{i,j} \ell_{ij}^2 = \sum_{i,j} \langle V_j^\star, \Delta U_i\rangle^2 + \sum_{i,j} \langle U_i^\star, \Delta V_j\rangle^2 + 2\sum_{i,j} \langle V_j^\star, \Delta U_i\rangle \langle U_i^\star, \Delta V_j\rangle.$$

The first term equals $\sum_i \Delta U_i^\top G_V^\star \Delta U_i \geq \lambda_{\min}(G_V^\star) \sum_i \|\Delta U_i\|^2$, and similarly the second is at least $\lambda_{\min}(G_U^\star) \sum_j \|\Delta V_j\|^2$. The cross term is nonnegative:

$$\sum_{i,j} (V_j^\star)^\top \Delta U_i (U_i^\star)^\top \Delta V_j = \sum_j (V_j^\star)^\top S\, \Delta V_j = \mathrm{tr}\Big(S \sum_j \Delta V_j (V_j^\star)^\top\Big) = \mathrm{tr}(SS^\top) = \|S\|_F^2 \ \geq\ 0.$$

Therefore $\sum_{i,j} \ell_{ij}^2 \geq h\big(\sum_i \|\Delta U_i\|^2 + \sum_j \|\Delta V_j\|^2\big)$, and Assumption 1 holds with $c = m^2$.

Fixed coordinates. If coordinate k of every V_j is fixed, perturbations have $(\Delta V_j)_k = 0$, and kernel directions within this subspace are those with B's k-th column equal to zero. Orthogonality to these gives $S_U := \sum_i \Delta U_i (U_i^\star)^\top$ and $S_V := \sum_j V_j^\star \Delta V_j^\top$ equal outside column k, while S_V vanishes on column k. The cross term is then $\langle S_U, S_V\rangle_F = \|S_V\|_F^2 \geq 0$, and the rest of the argument is unchanged. □

Proof of Theorem 5.

Proof. Note that $\Theta \in \mathcal{E}_\varepsilon$ iff $\Theta \in \mathcal{N}$ and $\|p(\Theta) - p^\star\| \leq \varepsilon$, and $\mathcal{E}_0 = p^{-1}(p^\star)$. Let k be the constant rank of $\mathrm{D}p$ near $\Theta^\star := (U^\star, V^\star)$ and $\sigma_k(\cdot)$ its k-th singular value, i.e., the smallest nonzero one. Assumption 1 states that $\sigma_k\big(\mathrm{D}p(\Theta^\star)\big)^2 \geq c\,h$.

Step 0: Choose the neighborhood. Fix $\rho_0 > 0$ such that, on the ball $B(\Theta^\star, \rho_0)$: (a) by the constant rank theorem, $\mathcal{E}_0 \cap B(\Theta^\star, \rho_0)$ is a C^1 submanifold whose tangent space at each point $\bar\Theta$ is $\ker \mathrm{D}p(\bar\Theta)$; (b) by continuity of σ_k (the rank being constant), $\sigma_k\big(\mathrm{D}p(\bar\Theta)\big)^2 \geq c\,h/2$; (c) since f is C^2, there is $L < \infty$ with $\|p(\Theta) - p(\bar\Theta) - \mathrm{D}p(\bar\Theta)(\Theta - \bar\Theta)\| \leq L\|\Theta - \bar\Theta\|^2$. Take $\mathcal{N} := B(\Theta^\star, \rho)$ with $\rho := \min\big\{\rho_0/2,\ \sqrt{ch/2}\,/(2L)\big\}$.

Step 1: Project onto the nearest exact representation. Fix $\Theta \in \mathcal{E}_\varepsilon$. Since $\mathcal{E}_0$ is closed and contains $\Theta^\star$, it has a point $\bar{\Theta}$ nearest to Θ, and $\Delta := \Theta - \bar{\Theta}$ satisfies $\|\Delta\| = \operatorname{dist}(\Theta, \mathcal{E}_0) \le \|\Theta - \Theta^\star\| < \rho$. Hence $\|\bar{\Theta} - \Theta^\star\| < 2\rho \le \rho_0$, so (a)–(c) apply at $\bar{\Theta}$ (and on the segment from $\bar{\Theta}$ to Θ). As $\bar{\Theta}$ minimizes the distance to Θ over the manifold in (a), Δ is orthogonal to its tangent space: $\Delta \perp \ker \mathrm{D}p(\bar{\Theta})$.

Step 2: Lower bound the change in predictions. Since $p(\bar{\Theta}) = p^\star$, (c) and (b) give

$$\varepsilon \;\ge\; \|p(\Theta) - p^\star\| \;\ge\; \|\mathrm{D}p(\bar{\Theta})\Delta\| - L\|\Delta\|^2 \;\ge\; \Big(\sqrt{ch/2} - L\|\Delta\|\Big)\|\Delta\| \;\ge\; \tfrac{1}{2}\sqrt{ch/2}\,\|\Delta\|,$$

where the last step uses $\|\Delta\| < \rho \le \sqrt{ch/2}\,/(2L)$.

Step 3: Conclude. Rearranging, $\operatorname{dist}(\Theta, \mathcal{E}_0) = \|\Delta\| \le \sqrt{8/c}\,\varepsilon/\sqrt{h}$. As $\Theta \in \mathcal{E}_\varepsilon$ was arbitrary, the theorem holds with $C = \sqrt{8/c}$ (so $C = \sqrt{8}/m$ for IRT/MIRT links, by Theorem 8). □

C Experiment Details

C.1 Inferring IRT parameters

We estimate the IRT parameters $\{a_i, d_i\}_{i=1}^n$ and $\{\theta_j\}_{j=1}^m$ via gradient-ascent on the joint log-likelihood

$$\ell(a, d, \theta) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{m}\Big[Y_{ij}\log p_{ij} + (1 - Y_{ij})\log(1 - p_{ij})\Big],$$

with ℓ_2 penalties on a_i, b_i, and θ_j to ensure stability and identifiability. We also apply periodic whitening transformations to θ_j so that the covariance of θ_j remains close to the identity; this removes rotational and scale indeterminacies inherent to IRT models, see [31] for details. We then recover the estimated probability of correctness (where Φ is the cdf of a standard normal):

$$\hat{p}_{ij} = \Phi\Big(\hat{a}_i^\top \hat{\theta}_j + \hat{d}_i\Big).$$

C.2 Experiment 1 (Details)

Baselines from prior work. Both Kim et al. [19] and Goel et al. [15] operate on multiple-choice data. Let $A_{ij} \in \{1, \dots, K_i\}$ denote the option selected by model j on item i, with correct option $a_i^\star$.

Kim et al. focus on agreement conditional on both models being incorrect:

$$q_{j\ell} := \Pr(A_{ij} = A_{i\ell} \mid A_{ij} \neq a_i^\star,\; A_{i\ell} \neq a_i^\star),$$

and compare this quantity to the chance baseline

$$q_{j\ell}^{\text{ch}} := \mathbb{E}\left[\frac{1}{K_i - 1}\right],$$

which corresponds to uniform random selection among the incorrect options. Excess correlation is defined as the difference between these two quantities:

$$\text{Excess} := q_{j\ell} - q_{j\ell}^{\text{ch}}$$

$$\kappa_{j\ell}^{\text{err}} := \frac{q_{j\ell} - q_{j\ell}^{\text{ch}}}{1 - q_{j\ell}^{\text{ch}}}$$

Goel et al. define a probabilistic agreement measure based on the inner product of predictive distributions. In the discrete (argmax-only) setting, their observed agreement reduces to

$$c_{\text{obs}} \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{A_{ij} = A_{i\ell}\},$$

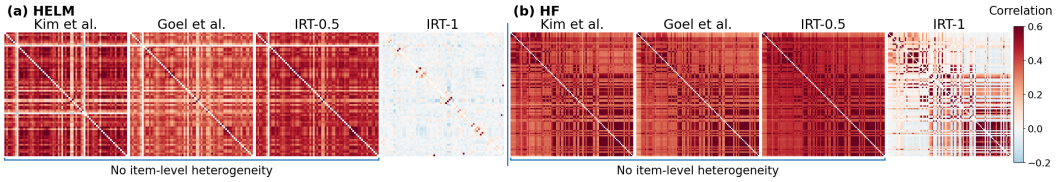


Figure 6: Residual correlation matrices for models in **HELM** (a) and **HF** (b), using different nulls: from left to right, a baseline from Kim et al. [19], from Goel et al. [15], from a 1D IRT with no item difficulty, and from a 1D IRT with item difficulties. Each item is a question. The first three null models **do not** include item heterogeneity. As such, the corresponding excess correlation for the full IRT model is attenuated compared to the others because item difficulties absorb much of the apparent positive correlation.

while their chance agreement baseline is

$$c_{\text{exp}} \;=\; p_j p_\ell \;+\; (1-p_j)(1-p_\ell)\,\mathbb{E}\left[\frac{1}{K_i-1}\right],$$

where p_j denotes model j's overall accuracy. Their CAPA statistic is a normalized excess-over-chance measure,

$$\text{CAPA}_{j\ell} \;=\; \frac{c_{\text{obs}} - c_{\text{exp}}}{1 - c_{\text{exp}}}.$$

Both baselines adjust for differences in model capability, but treat all items as exchangeable.

Our null: 1-dimensional IRT on multiple-choice correctness. We now evaluate excess correlation using a strictly more expressive null model. Specifically, we fit a *1-dimensional IRT model* ($K = 1$) to the same multiple-choice data, where

$$Y_{ij} := \mathbf{1}\{A_{ij} = a_i^\star\}$$

indicates whether model j selected the correct option on item i. The null model is

$$\Pr(Y_{ij} = 1) \;=\; \Phi(a_i\theta_j + b_i),$$

with item-specific difficulty and discrimination parameters (a_i, b_i) and model ability θ_j.

This model strictly subsumes the baselines above: if all items share identical parameters, the IRT model reduces to a capability-only model, whereas allowing (a_i, b_i) to vary introduces item heterogeneity. Thus, in the language of Section 3.2, this represents a higher rung in the null ladder.

Residual dependence. After fitting the model, we compute residuals

$$R_{ij} = Y_{ij} - \hat{p}_{ij},$$

de-bias them as in Experiment 1, and estimate the residual covariance matrix

$$\widehat{C} \;=\; \frac{1}{n}\tilde{R}^\top \tilde{R}.$$

This quantity captures excess model-to-model dependence not explained by either model ability or item difficulty.

C.2.1 Additional Results

In Figure 2, we used a null model where each item is a (question, answer-choice) pair in order to provide a fair comparison to previous methods. In Figure 6, we use a null model where each item is at the question level. The results do not change significantly, indicating that our findings remain robust to the choice of null model.

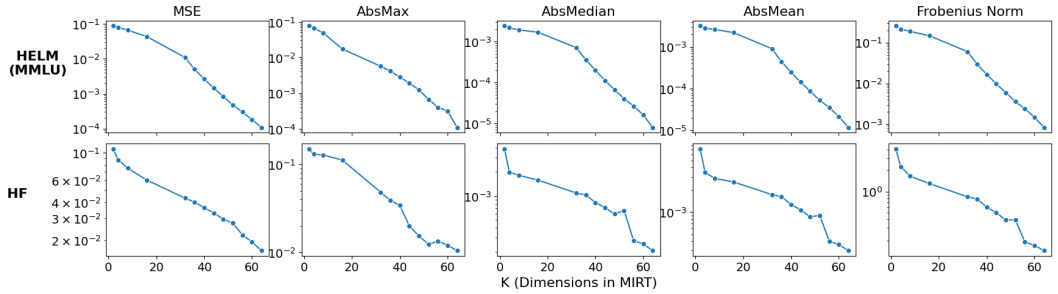


Figure 7: Mean square error (MSE) and summary statistics of the distribution of pairwise residual correlations, as a function of K: the dimensions in the multidimensional IRT model. Top (bottom) row shows results on the **HELM** (**HF**) dataset. As K increases, the residual correlations that are unexplained by the independent null model tend toward zero, meaning that a sufficiently expressive null model can arbitrarily absorb model correlations.

C.3 Figure 4: Increasing dimensions reduces excess correlation

Setup. Consider a sequence of IRT models with increasing dimensions K. For each $K \in \{1, 2, 4, 8, 16, 32, 36, \ldots, 60, 64\}$, we fit a K-dimensional ability model via the procedure from Section 2.2. For each K, we report fit via mean squared error: $\mathrm{MSE}^{(K)} := \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(Y_{ij} - \hat{p}_{ij}^{(K)}\right)^2$. We also calculate the excess correlation matrix $\widehat{\Sigma}$ for each K.

In order to summarize the pairwise covariance distribution, we compute the following summary statistics from C for each K:

$$\text{(i) AbsMax}^{(K)} := \max_{j\neq\ell}\left|\widehat{\Sigma}_{j\ell}^{(K)}\right|, \qquad \text{(ii) AbsMean}^{(K)} := \frac{1}{m(m-1)}\sum_{j\neq\ell}\left|\widehat{\Sigma}_{j\ell}^{(K)}\right|,$$

$$\text{(iii) Median}^{(K)} := \mathrm{median}_{j\neq\ell}\left|\widehat{\Sigma}_{j\ell}^{(K)}\right|, \qquad \text{(iv) } \|\widehat{\Sigma}^{(K)}\|_{\mathrm{off},F} := \sqrt{\sum_{j\neq\ell}\left(\widehat{\Sigma}_{j\ell}^{(K)}\right)^2}.$$

Results. Figure 7 shows the quantities of interest (y-axis) across K dimensions of the IRT model (x-axis) on both **HELM** and **HF**. By Proposition 2, enlarging the feasible class from $\mathcal{N}_K$ to $\mathcal{N}_{K+1}$ cannot increase the optimal discrepancy, and we indeed observe a monotone decrease of $\mathrm{MSE}^{(K)}$ as K increases. Similarly, by Theorem 3, we see that the pairwise residual covariances go to 0 as K increases. This result suggests that unexplained correlation vanishes for sufficiently complex null models. Note that the figures are on a log scale, meaning that increasing dimensions exponentially decreases excess correlation.

D Extensions

D.1 Generalizing to other output spaces

The framework accommodates a wide range of choices for $\mathcal{Y}$, including:

(1) **Binary correctness:** $\mathcal{Y} = \{0, 1\}$, where $Y_{ij} = 1$ denotes a correct answer.

(2) **Probabilistic correctness:** $\mathcal{Y} = [0, 1]$, where Y_{ij} is a calibrated probability of correctness.

(3) **Logits or score vectors:** $\mathcal{Y} = \mathbb{R}^V$, such as the model's pre-softmax logits over a vocabulary of size V.

(4) **Natural language:** $\mathcal{Y} = \mathcal{T}^*$, the space of token sequences.

Generalizing to probabilistic correctness is immediate: one simply replaces each Bernoulli variable with a real value in $[0, 1]$ and defines independence conditional on latent parameters in the same manner. Richer output spaces (logits and natural language) require more care. For logits, one natural approach is to view each Y_{ij} as a vector of real-valued features and study conditional independence of

these vectors given latent parameters. For natural language, a possible strategy is to embed outputs via a mapping $f : \mathcal{T}^* \to \mathbb{R}^d$ (e.g. semantic or style embeddings) and treat the embedded representations $X_{ij} = f(Y_{ij})$ as the objects of analysis. In this case, an "independent" null model would posit that, conditional on latent structure, these embeddings are independent across models. However, the appropriate parameterization of such latent structure (e.g. global factors capturing topic, stylistic preferences, or discourse-level strategies) is not yet clear and presents an interesting direction for future work.

D.2 Inferring Σ

In this work, we define and calculate monoculture Σ as the pairwise correlations of the residuals between the outcomes Y and the null model fitted probabilities $\hat{p}$. This gives us a metric that is agnostic to the particular inference algorithm used to infer Σ. However, one may want to directly infer Σ from the data, and in this section we provide one principled way of doing so.

Kernel parameterization and estimation of Σ. We first introduce a latent Gaussian variable Z_{ij}, where $Y_{ij} = \mathbb{I}[Z_{ij} \geq 0]$. We can therefore equivalently re-write the generative model as:

$$Z_{ij} \;=\; \theta_j - b_i + \varepsilon_{ij}, \qquad \varepsilon_{i,\cdot} \sim \mathcal{N}(0, \Sigma), \tag{4}$$

where for each item i the vector $\varepsilon_{i\cdot} \in \mathbb{R}^M$ (across M models) has covariance matrix $\Sigma \in \mathbb{R}^{M \times M}$, shared across items. The null model corresponds to $\Sigma = I_M$, i.e., conditionally independent models given (θ, b). Any deviation $\Sigma \neq I_M$ is interpreted as *excess correlation* between models beyond what is induced by sharing the same item difficulties and model abilities.

In the alternative model (4), assume the excess-correlation matrix $\Sigma \in \mathbb{R}^{M \times M}$ is parameterized via low-dimensional embeddings for each model. Let $v_j \in \mathbb{R}^K$ denote the embedding for model j, and let $V \in \mathbb{R}^{M \times K}$ collect these row-wise. We define pairwise correlations by

$$\rho_{jk} \;=\; v_j^\top v_k, \qquad j \neq k,$$

and enforce unit variance by setting

$$\Sigma \;=\; VV^\top \;+\; \mathrm{diag}\big(1 - \|v_1\|_2^2, \ldots, 1 - \|v_M\|_2^2\big). \tag{5}$$

If we maintain $\|v_j\|_2 \leq 1$ for all j, then Σ is positive semidefinite and has unit diagonal:

$$\Sigma_{jj} \;=\; \|v_j\|_2^2 + (1 - \|v_j\|_2^2) \;=\; 1.$$

In the implementation, we explicitly use $\rho_{jk} = v_j^\top v_k$ for $j \neq k$ in the bivariate probit calculations and fix $\Sigma_{jj} = 1$ via (5).

Model-implied and empirical covariances. Let $Y \in \{0,1\}^{N \times M}$ be the correctness matrix (items $\times$ models), and let $(\hat{\theta}, \hat{b})$ be the 1D IRT estimates from Stage 1. Define

$$\hat{\mu}_{ij} = \hat{\theta}_j - \hat{b}_i, \qquad \hat{p}_{ij} = \Phi(\hat{\mu}_{ij}), \qquad \hat{m}_j = \frac{1}{N}\sum_{i=1}^{N} \hat{p}_{ij}, \qquad \bar{Y}_j = \frac{1}{N}\sum_{i=1}^{N} Y_{ij}.$$

For a given pair (j,k) and correlation ρ_{jk}, the latent pair (Z_{ij}, Z_{ik}) for item i is modeled as

$$\begin{pmatrix} Z_{ij} \\ Z_{ik} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \hat{\mu}_{ij} \\ \hat{\mu}_{ik} \end{pmatrix}, \begin{pmatrix} 1 & \rho_{jk} \\ \rho_{jk} & 1 \end{pmatrix} \right).$$

The model-implied joint success probability for item i is

$$p_{11}^{(i)}(j,k;\rho_{jk}) \;:=\; \mathbb{P}\big(Y_{ij} = 1, Y_{ik} = 1\big) \;=\; \Phi_2\big(\hat{\mu}_{ij}, \hat{\mu}_{ik}; \rho_{jk}\big),$$

where $\Phi_2(\cdot,\cdot;\rho)$ is the bivariate normal CDF with correlation ρ. The model-implied covariance between $Y_{\cdot j}$ and $Y_{\cdot k}$ is

$$C_{jk}^{\text{model}}(V) \;=\; \frac{1}{N}\sum_{i=1}^{N} p_{11}^{(i)}(j,k;\rho_{jk}) \;-\; \hat{m}_j \hat{m}_k, \qquad \rho_{jk} = v_j^\top v_k. \tag{6}$$

The empirical covariance is

$$C_{jk}^{\text{emp}} \;=\; \frac{1}{N}\sum_{i=1}^{N} Y_{ij}Y_{ik} \;-\; \bar{Y}_j\bar{Y}_k. \tag{7}$$

For each iteration, the implementation computes these quantities in a streaming/tiled fashion over items and model blocks, but conceptually they are just (6)–(7).

Loss and gradients in correlation space. We fit V by matching model-implied and empirical covariances off the diagonal. The loss is

$$\mathcal{L}(V) \;=\; \sum_{1\le j<k\le M} \big(C_{jk}^{\text{model}}(V) - C_{jk}^{\text{emp}}\big)^2. \tag{8}$$

To differentiate with respect to ρ_{jk}, note that

$$\frac{\partial}{\partial \rho_{jk}} p_{11}^{(i)}(j,k;\rho_{jk}) \;=\; \varphi_2\big(\hat{\mu}_{ij},\hat{\mu}_{ik};\rho_{jk}\big),$$

where φ_2 is the bivariate normal PDF with correlation ρ_{jk}. Hence

$$\frac{\partial C_{jk}^{\text{model}}}{\partial \rho_{jk}} \;=\; \frac{1}{N}\sum_{i=1}^{N} \varphi_2\big(\hat{\mu}_{ij},\hat{\mu}_{ik};\rho_{jk}\big) \;=:\; \bar{\varphi}_{jk},$$

and by the chain rule,

$$\frac{\partial \mathcal{L}}{\partial \rho_{jk}} \;=\; 2\big(C_{jk}^{\text{model}} - C_{jk}^{\text{emp}}\big)\bar{\varphi}_{jk}. \tag{9}$$

In the code, $C_{jk}^{\text{model}} - C_{jk}^{\text{emp}}$ is stored as `diff_blk`, and $\bar{\varphi}_{jk}$ as `phibar_blk`, so that

$$G_{jk} \;:=\; \frac{\partial \mathcal{L}}{\partial \rho_{jk}} \;=\; 2\,\texttt{diff_blk}_{jk}\,\texttt{phibar_blk}_{jk}.$$

Gradients in embedding space and optimization. Because $\rho_{jk} = v_j^\top v_k$ for $j \neq k$, we have

$$\frac{\partial \rho_{jk}}{\partial v_j} = v_k, \qquad \frac{\partial \rho_{jk}}{\partial v_k} = v_j,$$

and thus

$$\frac{\partial \mathcal{L}}{\partial v_j} \;=\; \sum_{k\neq j} \frac{\partial \mathcal{L}}{\partial \rho_{jk}}\frac{\partial \rho_{jk}}{\partial v_j} \;=\; \sum_{k\neq j} G_{jk}\, v_k. \tag{10}$$

In matrix form, if G is the matrix with entries G_{jk} (symmetric and zero on the diagonal), then the gradient w.r.t. V is

$$\nabla_V \mathcal{L} \;=\; GV,$$

which is exactly what the code accumulates in the tensor `Z`. We then perform Adam updates on V:

$$V \leftarrow \mathrm{AdamStep}\big(V, \nabla_V \mathcal{L}\big),$$

with bias-corrected learning rate as in standard Adam.

Norm projection and unit variance. After each Adam update, we project each row v_j onto the unit ball:

$$v_j \;\leftarrow\; \frac{v_j}{\max\{1, \|v_j\|_2\}}. \tag{11}$$

This ensures $\|v_j\|_2 \le 1$ for all j, so that the covariance matrix Σ in (5) remains positive semidefinite with unit diagonal. Conceptually, we are learning a low-rank correlation structure $\rho_{jk} = v_j^\top v_k$ subject to the constraint that all variances are fixed at one, which preserves the scaling of the underlying 1D IRT model.

Practical details. The implementation evaluates the loss (8) and its gradients exactly but does so in a memory-efficient way by: (i) streaming over items in chunks (size `item_chunk`) and (ii) tiling

the model index set into blocks (size `model_block`) to compute C_{jk}^{model} and C_{jk}^{emp} blockwise. We use early stopping based on the best observed loss over iterations.

Justification for two-stage estimation. This two-stage procedure separates *first-order* and *second-order* structure: Stage 1 uses only the marginal success probabilities to learn (θ, b), while Stage 2 uses the residual covariances to learn Σ. Under standard regularity conditions, the Stage 1 maximum-likelihood estimates $(\hat{\theta}, \hat{b})$ are consistent and $\sqrt{N}$-accurate even if $\Sigma \neq I_M$, because the marginal Bernoulli probabilities depend only on (θ, b) and not on Σ. Plugging these into the second-stage moment-matching objective yields a consistent estimator of Σ that is asymptotically equivalent to a joint estimator but computationally simpler. Conceptually, this is analogous to generalized method-of-moments procedures that estimate parameters associated with lower-order moments before fitting higher-order structure.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We made sure our claims in the abstract and introduction properly reflect our findings.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See final paragraph of paper.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Proofs can be found in Appendix.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experiment described in high level before results in the main paper. Details can be found in Appendix.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: URL provided in footnote 1.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental setting described in the paper, with details in the Appendix.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Experiments involve one static dataset and an inference procedure that has no/little variability, see explanation in Lines 104-106.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See "Setup" paragraph in Experiments 1 and 2.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The authors have read the NeurIPS Code of Ethics and confirm that the research conducted adheres to this code.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We address the societal impacts (particularly to the science of evaluations) in Section 5.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Owners of existing assets are properly credited and their terms of use are properly respected.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Documentation is provided in the URL in footnote 1.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

Justification: This paper does not involve LLMs as any important, original, or non-standard components.