CFI: CAPABILITY FISHER INFORMATION FOR BENCHMARK DIAGNOSIS IN LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In current practice, LLM benchmarks are often evaluated by the reliability of the model rankings they produce, yet stable rankings do not necessarily imply sufficient and balanced measurement of the capabilities they are intended to assess. We propose Capability Fisher Information (CFI), a measurement-theoretic framework for diagnosing benchmark structure through the distribution of information in a benchmark-specific capability space. CFI represents each item as an information contribution characterized by a capability direction and strength, aggregates these contributions into a benchmark-level information operator, and supports three complementary diagnostics: Coverage, Uniformity, and Redundancy. Together, they expose capability blind spots, information imbalance, and repeated probing that overall accuracy or item counts cannot reveal. Across text-only and multimodal benchmarks, CFI reveals structural deficiencies overlooked by conventional statistics, while controlled structural perturbations and held-out model experiments validate the effectiveness of its diagnostics.

1 INTRODUCTION

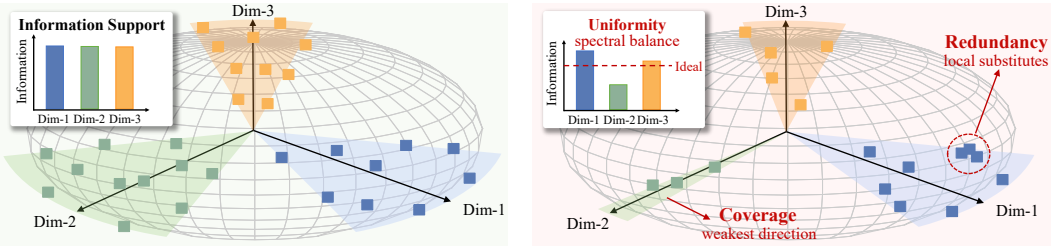
As large language models continue to expand their capabilities across knowledge, reasoning, and multimodal understanding, the number and scope of evaluation benchmarks have grown rapidly (Chang et al., 2024). Yet most existing evaluation paradigms focus on ranking models, while paying far less attention to whether the benchmarks themselves provide reliable measurements of intended capabilities. From a measurement perspective, model capabilities can be viewed as forming a multidimensional capability space, with each benchmark item serving as a probe along one or more directions in that space. The quality and distribution of these probes are therefore central to the validity of model scores, rankings, and the capability conclusions drawn from them.

However, current benchmark construction largely follows a scale-driven strategy: *adding more items, domains, and task formats to place an increasingly dense set of probes across this space*. Yet any finite benchmark can observe only a limited portion of the underlying capability space, and a denser set of probes does not necessarily yield better measurement. For instance, the probes may leave important capabilities underrepresented, or repeatedly examine similar regions. Prior work has shown that carefully constructed compact evaluations can sometimes preserve the conclusions of much larger test sets (Liang et al., 2022; Perlitz et al., 2024; Polo et al., 2024). Benchmark quality therefore depends not only on scale, but more fundamentally on how effectively a finite set of probes distributes measurement information across the target capability space. However, existing work lacks a systematic framework for characterizing and diagnosing benchmark structure from this perspective.

To address this gap, we draw on classical measurement theory and multidimensional item response theory (Lord and Novick, 2008; Embretson and Reise, 2004; Reckase, 2009) to propose Capability Fisher Information (CFI), a Fisher-information-inspired framework for benchmark diagnosis. CFI turns benchmark diagnosis into a computable analysis of how measurement information is distributed across a benchmark-specific capability space. Fig. 1 illustrates this idea by contrasting an ideal benchmark structure with one exhibiting structural deficiencies: insufficient support along some capability directions, imbalanced information allocation, and redundant probing.

Specifically, we use the benchmark's predefined evaluation dimensions to construct an approximately orthonormal, benchmark-specific capability space. Each item is projected into this space to determine its capability direction and represented as a weighted rank-one information matrix. It characterizes the

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



(a) Ideal Benchmark Structure (b) Structurally Deficient Benchmark

Figure 1: Ideal and structurally deficient benchmark structures under the CFI framework.

capability directions along which individual items contribute measurement information, quantifies the strength of these contributions. Then, aggregating these item-level matrices yields a benchmark-level information operator. Based on the spectral properties of this operator, we derive Coverage and Uniformity, and complement them with an item-level Redundancy measure. Coverage captures the weakest-supported capability direction, Uniformity measures the balance of information across directions, and Redundancy quantifies repeated probing among semantically and capability-wise similar items. In summary, our work makes the following contributions:

- **A measurement-theoretic view of benchmark quality.** We introduce Capability Fisher Information (CFI), which quantifies how benchmark items support directions in a benchmark-specific capability space, enabling structural analysis beyond aggregate scores.
- **Diagnostic metrics for benchmark structure.** Building on CFI, we develop three complementary metrics that target distinct structural defects in benchmark item placement: Coverage, Uniformity, and Redundancy.
- **Empirical validation of diagnostic behavior.** We validate that CFI metrics reflect specific structural defects in benchmarks, relate to subset evaluation quality on held-out models, and remain consistent across variations in the capability-space construction.

2 Related Work

LLM Evaluation Benchmarks. Modern LLM benchmarks are often constructed as large question banks emphasizing breadth across domains and task formats (Cobbe et al., 2021; Hendrycks et al., 2021; 2020; Liang et al., 2022). While this improves coverage in a practical sense, most benchmark studies do not provide formal guarantees about measurement properties such as coverage balance across intended capabilities, discriminative signal, or evaluation reliability. Further challenges include benchmark overuse and data contamination (Bowman and Dahl, 2021; Xu et al., 2024; Sainz et al., 2023), motivating dynamic or evolving benchmarks (Kiela et al., 2021). Our work complements these efforts by treating benchmark design as a measurement problem and introducing structural diagnostics for benchmark quality assessment.

Measurement Theory and Information-Theoretic Test Design. In psychometrics, classical test theory and item response theory (IRT) model test items as probes of latent abilities, linking measurement quality to notions such as reliability, discriminability, and identifiability (Rasch, 1993; Embretson and Reise, 2004; Lord, 2012). Information geometry and Fisher information provide a principled way to characterize how measurement sensitivity depends on how informative samples populate a capability space (Fisher, 1925; Amari, 2016). These ideas have also inspired work on example selection and informativeness in machine learning (Settles, 2009). We adapt this perspective to LLM benchmark auditing by using Fisher-information-inspired geometry to derive benchmark structure diagnostics that are interpretable signals for each capability direction.

Benchmark Quality Analysis. Prior work studies practical issues in existing benchmarks, including limitations of relying on single-number metrics (Ethayarajh and Jurafsky, 2020) and methods for improving robustness, calibration, and evaluation protocols (Dror et al., 2018; Guo et al., 2017; Liang et al., 2022). Relatedly, (Qian et al., 2026) proposes systematic metrics for benchmark reliability and uses them to select more stable evaluation subsets. In contrast, our contribution targets the benchmark

structure itself: we diagnose whether benchmark items provide appropriate and non-degenerate measurement signal across predefined capability directions, and we validate diagnostic validity via controlled perturbations and calibration sensitivity.

3 METHOD

3.1 FRAMEWORK OVERVIEW

Let $\mathcal{B} = \{x_i\}_{i=1}^n$ denote a benchmark containing n evaluation items. Model capabilities can be conceptually represented by a global capability space $\mathcal{F}$. However, $\mathcal{F}$ is inherently open-ended and continually evolves with models, tasks, and evaluation objectives, making it impossible to define or exhaustively enumerate. Consequently, a finite benchmark $\mathcal{B}$ can only target a selected set of capabilities, inducing a benchmark-specific local space

$$\mathcal{F}_{\mathcal{B}} \subseteq \mathcal{F}, \qquad \dim(\mathcal{F}_{\mathcal{B}}) = d, \tag{1}$$

where $\mathcal{F}_{\mathcal{B}}$ represents the portion of the global capability space that $\mathcal{B}$ is designed to measure. Accordingly, our goal is not to assess whether $\mathcal{B}$ covers the global capability space $\mathcal{F}$, but to diagnose whether its item probes are appropriately distributed within $\mathcal{F}_{\mathcal{B}}$.

To make this diagnosis computable, we summarize each benchmark item by an operational *information contribution* in the benchmark-specific capability space. Concretely, we represent item x_i by a unit probing direction v_i and a nonnegative information strength α_i, and aggregate them into a benchmark-level information operator $\mathrm{CFI}(\mathcal{B}, \alpha) = \sum_{i=1}^n \alpha_i v_i v_i^\top$. Based on this operator, we construct three diagnostics of benchmark structure: Coverage, Uniformity, and Redundancy. We next justify why this rank-one direction–strength decomposition is Fisher-information motivated, and then specify how v_i and α_i are estimated from the benchmark.

3.2 THEORETICAL FORMULATION

A benchmark $\mathcal{B}$ can be viewed as a measurement instrument whose quality depends on whether its items can distinguish models with different capability profiles. Since benchmark-level measurement is aggregated from individual item responses, we begin by examining the information contributed by a single item. Let $\theta \in \mathbb{R}^d$ denote a model's latent capability state in the benchmark-specific capability space $\mathcal{F}_{\mathcal{B}}$, and let $\theta + \Delta$ denote a nearby state differing by a small capability variation Δ. For item x_i, let $q_i(y \mid \theta)$ denote the distribution of its evaluated response outcome under capability state θ. For binary-scored items, $y \in \{0, 1\}$ indicates an incorrect or correct response, respectively. If a small change from θ to $\theta + \Delta$ induces a substantial change in the response distribution, item x_i is more sensitive to the corresponding capability variation and therefore provides more measurement information. We quantify this local change using the KL divergence,

$$g_i(\Delta) \triangleq D_{\mathrm{KL}}(q_i(\cdot \mid \theta) \,\|\, q_i(\cdot \mid \theta + \Delta)) \,. \tag{2}$$

We use $g_i(\Delta)$ to quantify the local measurement information provided by item x_i for a specific capability change Δ. Because Δ can vary across different directions and magnitudes in the capability space, evaluating $g_i(\Delta)$ separately for every possible change is impractical. Under standard regularity conditions, the local behavior of $g_i(\Delta)$ can instead be summarized by the Fisher information matrix through the following second-order approximation,

$$g_i(\Delta) = \frac{1}{2}\Delta^\top \mathcal{I}_i(\theta)\Delta + o\big(\|\Delta\|_2^2\big) \,, \qquad \|\Delta\|_2 \to 0, \tag{3}$$

where $\mathcal{I}_i(\theta) \in \mathbb{R}^{d\times d}$ is the Fisher information matrix of item x_i at capability state θ. For a specific capability change Δ, the quadratic form $\Delta^\top \mathcal{I}_i(\theta)\Delta$ measures the item's local sensitivity to that change, with larger values indicating more distinguishable response distributions. The Fisher information matrix associated with the response distribution $q_i(y \mid \theta)$ is defined as

$$\mathcal{I}_i(\theta) = \sum_y q_i(y \mid \theta) \nabla_\theta \log q_i(y \mid \theta) \nabla_\theta \log q_i(y \mid \theta)^\top, \tag{4}$$

Following item response theory, we instantiate $q_i(y \mid \theta)$ using a multidimensional 2-parameter logistic (2PL) model(Reckase, 2009),

$$\pi_i(\theta) = q_i(1 \mid \theta) = \sigma\big(a_i^\top \theta - \beta_i\big) \,, \qquad q_i(0 \mid \theta) = 1 - \pi_i(\theta), \tag{5}$$

where $a_i \in \mathbb{R}^d$ is the discrimination vector of item x_i, $\beta_i \in \mathbb{R}$ is its difficulty parameter, and $\sigma(t)$ is the logistic function. Substituting Eq. 5 into Eq. 4 gives

$$\mathcal{I}_i(\theta) = \pi_i(\theta)\left(1 - \pi_i(\theta)\right) a_i a_i^\top. \tag{6}$$

Writing $a_i = \gamma_i u_i$, where $\gamma_i = \|a_i\|_2$ and $\|u_i\|_2 = 1$, further yields

$$\mathcal{I}_i(\theta) = \underbrace{\kappa_i(\theta)\gamma_i^2}_{\text{strength}} \underbrace{u_i u_i^\top}_{\text{direction}}, \qquad \kappa_i(\theta) = \pi_i(\theta)\left(1 - \pi_i(\theta)\right). \tag{7}$$

This rank-one form decomposes item information into a capability direction u_i specifying *where* the item is informative in the latent capability space, and a nonnegative strength $\kappa_i(\theta)\gamma_i^2$ determining *how much* information it contributes along that direction. Together, they characterize the item-level Fisher information matrix $I_i(\theta)$, with the latter controlling the magnitude of information contributed along the direction specified by the former. Detailed derivations are provided in Appendix B.

3.3 Capability Fisher Information

To apply the rank-one decomposition to an existing benchmark, we construct computable operational counterparts for the direction and strength terms, denoted by v_i and α_i, respectively.

3.3.1 Capability Direction

Having established the probing direction as the capability-space direction along which an item contributes information, we operationalize it by representing the item in the benchmark-specific capability space. Specifically, we first construct an approximately orthonormal basis for this space and project the item embedding onto the basis. The resulting coordinates characterize the item's relative composition across the benchmark-defined capability directions. We then normalize these coordinates to obtain v_i, which retains only the item's directional profile, while the overall magnitude of its information contribution is modeled separately by the strength term α_i.

Capability grouping. Let $\ell_i \in 1, \ldots, d$ denote the capability label of item x_i, where d is the number of target capabilities. These labels partition the benchmark items into d capability groups,

$$\mathcal{G}_\mathcal{B} = \{G_j\}_{j=1}^d, \qquad G_j = \{x_i \in \mathcal{B} : \ell_i = j\}, \tag{8}$$

where G_j contains the items associated with the j-th capability. When the benchmark provides predefined capability labels, such as the subject categories in MMLU-Pro, we use them directly to determine ℓ_i. Otherwise, we identify latent capability groups by clustering.

Capability representation. Let $E(\cdot)$ denote a frozen encoder with output dimension m. We represent each item by an ℓ_2-normalized embedding $z_i = E(x_i)/\|E(x_i)\|_2 \in \mathbb{R}^m$ and derive a semantic prototype $b_j \in \mathbb{R}^m$ for each capability group,

$$\bar{z}_j = \frac{1}{|G_j|} \sum_{i \in G_j} z_i, \qquad b_j = \frac{\bar{z}_j}{\|\bar{z}_j\|_2}, \qquad B = [b_1, \ldots, b_d] \in \mathbb{R}^{m \times d}. \tag{9}$$

Here, $\bar{z}_j$ is the mean embedding of the items in G_j, and B is the capability prototype matrix whose j-th column b_j is the normalized semantic prototype of the j-th target capability.

Item projection. The capability prototypes define the benchmark-specific subspace but are not necessarily orthogonal. We therefore calibrate the prototype basis via *regularized Gram whitening*, which reduces correlations among the prototype directions and provides an approximately orthonormal coordinate system for the subspace spanned by the original prototypes. Let $B \in \mathbb{R}^{m \times d}$ denote the prototype matrix. We define

$$\widetilde{B} = B\left(B^\top B + \epsilon I_d\right)^{-1/2}, \qquad \widetilde{B}^\top \widetilde{B} \approx I_d, \tag{10}$$

where I_d denotes the $d \times d$ identity matrix and ϵ is a small constant used for numerical stability. The columns of $\widetilde{B} = [\widetilde{b}_1, \ldots, \widetilde{b} * d] \in \mathbb{R}^{m \times d}$ form an approximately orthonormal basis of $\mathcal{F} * \mathcal{B}$ and are used to represent benchmark items in this space. Given $\widetilde{B}$, we define the capability direction of item x_i as $v_i = \widetilde{B}^\top z_i / \|\widetilde{B}^\top z_i\|_2 \in \mathbb{R}^d$. Implementation details for capability grouping, prototype construction, and the regularized Gram whitening procedure are provided in Appendix E.

3.3.2 INFORMATION STRENGTH

In the 2PL formulation, the information strength along an item's probing direction is given by the nonnegative term $\kappa_i(\theta)\gamma_i^2$, combining response uncertainty and discrimination magnitude. Directly estimating this quantity for an existing LLM benchmark would require fitting a multidimensional response model, whose latent parameters are typically unavailable. We therefore introduce an operational strength α_i and consider two regimes: uniform strength, which focuses on directional structure, and reference-weighted strength, which incorporates empirical response variation from a pool of reference models.

Uniform strength. We set $\alpha_i = 1$ for all items, so every item contributes one unit of information along its probing direction. This provides a model-independent view of the benchmark's directional structure without requiring additional model responses.

Reference-weighted strength. To reflect response variation across a model population, we introduce a reference model pool spanning a range of capability levels. For each item x_i, let $\widehat{p}_i$ denote the fraction of reference models that answer it correctly, and define $\alpha_i = \widehat{p}_i(1 - \widehat{p}_i)$. This quantity is small when nearly all models either succeed or fail on the item, and is maximized when their responses are most evenly split. It therefore captures the item's empirical response variation and, to some extent, its discriminative utility within the reference population. Since this form corresponds to the response-uncertainty term $\kappa_i(\theta) = \pi_i(\theta)(1 - \pi_i(\theta))$ in the 2PL Fisher information, we use it as an operational proxy for the observable component of theoretical information strength. Its dependence on the reference pool is further analyzed in Appendix C.

3.3.3 DEFINITION AND PROPERTIES

Having obtained the capability direction and information strength of each item, we combine them to characterize the measurement information of $\mathcal{B}$ in the benchmark-specific capability space.

Definition 3.1 (Capability Fisher Information)**.** Given a benchmark $\mathcal{B} = \{x_i\}_{i=1}^n$, let $v_i \in \mathbb{R}^d$ denote the unit probing direction of item x_i, and let $\alpha_i \geq 0$ denote its information strength. We define the operational information matrix of x_i as

$$\widehat{\mathcal{I}}_i = \alpha_i v_i v_i^\top. \tag{11}$$

The CFI of $\mathcal{B}$ is defined by aggregating the information matrices of its items,

$$\mathrm{CFI}(\mathcal{B}, \alpha) = \sum_{i=1}^{n} \widehat{\mathcal{I}}_i = \sum_{i=1}^{n} \alpha_i v_i v_i^\top. \tag{12}$$

It summarizes how the measurement information provided by $\mathcal{B}$ is distributed across the benchmark-specific capability space. CFI is monotone under item addition: if $\mathcal{B}' \subseteq \mathcal{B}$, then $\mathrm{CFI}(\mathcal{B}, \alpha) - \mathrm{CFI}(\mathcal{B}', \alpha) = \sum_{i \in \mathcal{B} \setminus \mathcal{B}'} \widehat{\mathcal{I}}_i \succeq 0$. Thus, adding items never decreases the information operator.

3.4 BENCHMARK STRUCTURE DIAGNOSTICS

A well-constructed benchmark should provide sufficient support for its intended capability space, distribute measurement information evenly across capability directions, and avoid excessive duplication among items. Accordingly, based on the CFI and the relationships among item probes, we define three complementary metrics to assess the structural quality of a benchmark.

Coverage. A benchmark may contain many items yet still leave some intended capability directions weakly supported. To capture this worst-case deficiency, we define Coverage as the smallest eigenvalue of the CFI operator,

$$\mathrm{Cov}(\mathcal{B}, \alpha) = \lambda_{\min}(\mathrm{CFI}(\mathcal{B}, \alpha)). \tag{13}$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue, corresponding to the minimum information provided by the benchmark over all unit directions in its capability space. A small value indicates that at least one direction receives insufficient support, while a value of zero indicates that some direction is completely uncovered.

Uniformity. Coverage captures the least-supported capability direction, but does not characterize the overall balance of information across the capability space. We therefore define Uniformity as the ratio of the geometric mean to the arithmetic mean of the CFI eigenvalues,

$$\mathrm{Unif}(\mathcal{B}, \alpha) = \frac{\det(\mathrm{CFI}(\mathcal{B}, \alpha))^{1/d}}{\mathrm{tr}(\mathrm{CFI}(\mathcal{B}, \alpha))\,/d}. \tag{14}$$

where $\det(\cdot)$ and $\mathrm{tr}(\cdot)$ denote the determinant and trace of a matrix, respectively. Uniformity lies in $[0, 1]$, attaining 1 when information is evenly distributed across all capability directions and approaching 0 as it becomes concentrated in fewer directions.

Redundancy. A benchmark may contain many items yet still include substantial local duplication, with multiple items probing similar capability directions and semantic content. We define benchmark Redundancy as the average, across all items, of the maximum pairwise overlap between each item and any other item,

$$\mathrm{Red}(\mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i} r_{ij}, \qquad r_{ij} = (v_i^\top v_j)^2\, \mathrm{sim}(i, j). \tag{15}$$

Here, $(v_i^\top v_j)^2$ measures their overlap in capability direction, while $\mathrm{sim}(i, j) \in [0, 1]$ measures semantic similarity between the two items. In our implementation, we use clipped cosine similarity between normalized item embeddings, $\mathrm{sim}(i, j) = \max\{0, z_i^\top z_j\}$, following common embedding-based semantic similarity practice (Reimers and Gurevych, 2019; Gao et al., 2021). The definition itself is agnostic to this particular choice of sim and can accommodate alternative semantic similarity estimators without changing Eq. 15. A larger value indicates greater local redundancy.

4 Experiments

In this section, we evaluate whether CFI provides useful diagnostics of benchmark structure. First, we use CFI to characterize the structure of existing benchmarks. Second, we test whether its diagnostics respond as expected to controlled changes in benchmark structure. Third, we examine whether the structural differences measured by CFI are associated with the fidelity and stability of subset evaluations. Specifically, we consider how well a subset preserves full-benchmark model rankings and scores, and how stable the resulting rankings remain under item resampling. We further examine the robustness of these findings to key modeling and implementation choices.

4.1 Experimental Setup

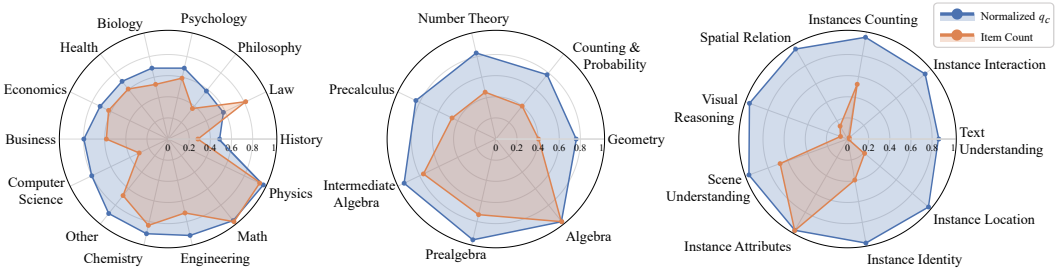
Benchmarks and capability dimensions. We conduct experiments on a collection of text-only and multimodal benchmarks.[1] For each benchmark, we fix the capability dimensions used to construct its benchmark-specific capability space. When native category annotations are available, these dimensions are derived from the annotated categories; otherwise, we identify capability groups by clustering item representations. The same capability definition is used throughout all analyses of a given benchmark, so comparisons remain benchmark-internal.

Evaluation measures. We use the three CFI diagnostics introduced in Section 3 to characterize benchmark structure: Coverage (Cov), Uniformity (Unif), and Redundancy (Red). For each subset, we additionally use three measures to evaluate the results it produces, which we collectively refer to as the *subset evaluation measures*. Ranking preservation (τ_{cat}) compares the model rankings obtained on the subset with those on the full benchmark within each capability group, and then averages the agreement across groups. Score preservation is measured by the mean absolute error (MAE) between subset and full-benchmark model scores. Ranking stability (τ_{self}) measures how stable the model rankings remain under item resampling, again computed within capability groups and averaged across groups. Formal definitions are provided in Appendix D.

[1] **Text-only:** MMLU-Pro (Wang et al., 2024), MATH$_{TEST}$ (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021). **Multimodal:** AI2D$_{TEST}$ (Kembhavi et al., 2016), MMMU$_{DEV\ VAL}$ (Yue et al., 2024), MMT-Bench$_{VAL}$ (Ying et al., 2024), SEEDBench$_{IMG}$ (Li et al., 2024).

Table 1: **Structural diagnostics of the full evaluated item set for each benchmark.** Coverage and Uniformity are reported under the uniform-strength and reference-weighted regimes, while Redundancy is weight-independent and reported once.

Benchmark	#Items	d	Uniform		Reference-weighted		Red ↓
			Cov ↑	Unif ↑	Cov ↑	Unif ↑	
Text-only Benchmarks							
MMLU-Pro	12,032	14	181.33	0.61	27.95	0.60	0.57
MATH_{TEST}	5,000	7	98.55	0.44	14.27	0.44	0.64
GSM8K	8,792	3	604.58	0.53	62.29	0.54	0.59
Multimodal Benchmarks							
AI2D_{TEST}	3,088	15	4.62	0.12	0.61	0.12	0.89
$\text{MMMU}_{DEV\ VAL}$	1,050	6	12.13	0.28	2.12	0.28	0.79
MMT-Bench_{VAL}	3,127	32	4.96	0.19	0.74	0.20	0.88
SEEDBench_{IMG}	14,232	9	49.81	0.13	4.98	0.13	0.87



(a) MMLU-Pro **(b) MATH_{TEST}** **(c) SEEDBench_{IMG}**

Figure 2: Radar profiles of Item Count vs. directional information mass q_c across categories for MMLU-Pro, MATH_{TEST}, and SEEDBench_{IMG}. For visual comparison, we normalize both quantities by their maximum values within each benchmark.

Reference and held-out models. We use two disjoint model pools for different parts of the experiment. Reference models are used only to estimate item-level response variation for the reference-weighted CFI. Held-out models are used only to compute the subset evaluation measures described above. In this way, the models used to determine CFI weights are different from those used to evaluate ranking preservation, score preservation, and ranking stability.

4.2 Benchmark-Level Structural Profiles

Benchmark-level structural diagnostics. We first characterize the structural profile of the full evaluated item set for each benchmark. Table 1 reports Coverage and Uniformity under the uniform-strength and reference-weighted regimes, together with the weight-independent Redundancy diagnostic. These quantities are intended primarily for benchmark-internal analysis: in particular, raw Coverage is directly comparable only when both the capability space and weighting regime are fixed.

As shown in Table 1, the benchmarks differ in how evenly information is distributed across capability directions and in how much local redundancy they contain. For example, MMLU-Pro has a Uniformity of 0.61 and Redundancy of 0.57, indicating relatively even support across directions with less local overlap. In contrast, SEEDBench_{IMG} has a much lower Uniformity of 0.13 and a higher Redundancy of 0.87, indicating a more uneven distribution of information and stronger local overlap. Reference weighting changes the scale of Coverage, but Uniformity remains similar within each benchmark, suggesting that the overall balance across directions is largely preserved under reweighting.

Category-level information allocation. We further compare the number of items in each category with its corresponding effective information mass to examine how measurement information is distributed across categories. For category c, we compute the reference-weighted information mass

$$q_c(\mathcal{B}, \alpha) = v_c^\top \mathrm{CFI}(\mathcal{B}, \alpha) v_c, \tag{16}$$

where v_c denotes the calibrated diagnostic direction associated with category c. Further details are provided in Appendix I. Whereas item count measures nominal representation, q_c measures the amount of reference-weighted information allocated along that direction.

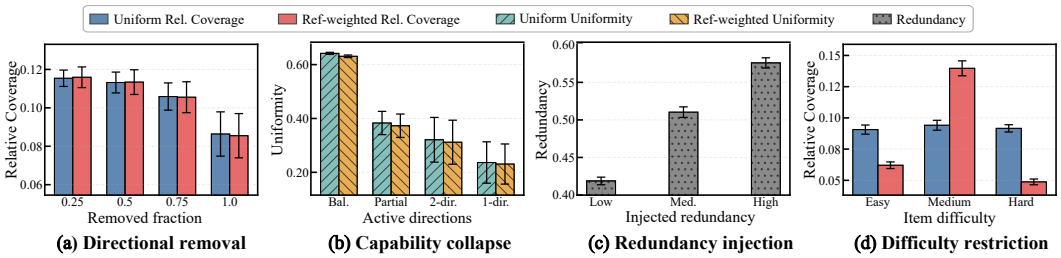


Figure 3: **Sensitivity of CFI diagnostics to controlled structural perturbations on MMLU-Pro.** Bars show means over 50 seeded subsets; error bars indicate one standard deviation. Relative Coverage is normalized by the full-benchmark value under the same weighting regime.

Fig. 2 shows clear differences in how information is distributed across categories. MMLU-Pro is relatively uneven, with higher information mass in Physics and Math and lower values elsewhere. MATH_{TEST} is more balanced, although Geometry remains less supported. In SEEDBench_{IMG}, item counts vary sharply across categories, while normalized information mass is comparatively even.

These results show that category size is not a reliable proxy for information support. More items do not necessarily provide more information, since similar items may repeatedly measure the same directions. Thus, q_c better reflects information allocation across categories than item count alone.

4.3 Sensitivity to Controlled Structural Perturbations

In this section, we test whether the CFI diagnostics capture the structure properties they are designed to measure by examining their responses to controlled changes in benchmark structure. On MMLU-Pro, we fix the subset budget at $b = 0.10$ and introduce four perturbations: (1) directional removal, which weakens information along a selected capability direction; (2) capability collapse, which concentrates coverage within a reduced set of capability directions; (3) redundancy injection, which increases local overlap among items; and (4) difficulty restriction, which narrows the range of item difficulty. Full construction details are provided in Appendix F.

The results confirm that the CFI diagnostics detect the intended structural changes. For example, Fig. 3(d) shows that reference-weighted Coverage peaks at intermediate difficulty and decreases for easy and hard items, while uniform-strength Coverage remains comparatively stable.

4.4 Association with Subset Evaluation Measures on Disjoint Models

We evaluate whether CFI provides a diagnostic of subset quality by testing whether its structural diagnostics are associated with how well a subset preserves full-benchmark model rankings and scores and how stable its rankings remain under item resampling. We fix the budget at $b = 0.10$ and construct 200 subsets from MMLU-Pro using four sampling strategies: random sampling, category-balanced sampling, capability-collapse sampling, and redundancy-injected sampling, with 50 subsets generated under each strategy. Neither the CFI diagnostics nor the evaluation measures are used to guide subset construction. We then compute the CFI diagnostics and subset evaluation measures defined in Section 4.1, and report their pairwise Spearman correlations across subsets. Full construction details are provided in Appendix G.

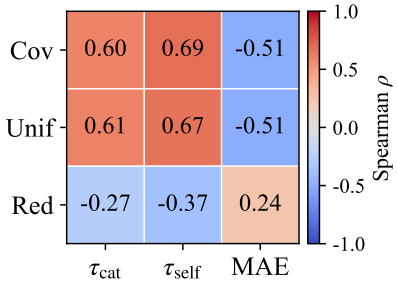
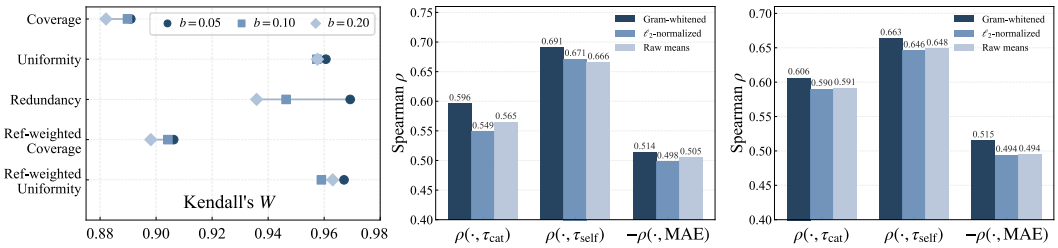


Figure 4: Spearman correlations between CFI diagnostics and evaluation measures across MMLU-Pro subsets.

Fig. 4 shows that Coverage and Uniformity are both strongly associated with the subset evaluation measures. Coverage has correlations of 0.60, 0.69, and -0.51 with τ_{cat}, τ_{self}, and MAE, respectively; the corresponding correlations for Uniformity are 0.61, 0.67, and -0.51. Thus, among subsets of the same size, higher Coverage and Uniformity tend to preserve full-benchmark rankings and scores better and yield more stable rankings. Redundancy shows weaker associations, with correlations of -0.27, -0.37, and 0.24, indicating only a weak relationship with the corresponding evaluation results.



(a) Encoder robustness **(b)** Coverage **(c)** Uniformity

Figure 5: **Robustness and ablations of the capability-space construction.** (a) Agreement of subset rankings on MMLU-Pro, quantified by Kendall's W across three text encoders for each subset budget. (b, c) Associations of Coverage and Uniformity with the subset evaluation measures under raw, ℓ_2-normalized, and Gram-whitened prototype constructions. For MAE, correlations are sign-reversed so that larger values consistently indicate stronger associations.

4.5 Sensitivity to the Capability-Space Construction

Our capability-space construction depends on two implementation choices: (i) the text encoder used to define the semantic geometry and (ii) the calibration applied to capability prototypes. We test whether the resulting CFI diagnostics reflect stable structural properties of the constructed subsets rather than artifacts of a particular embedding space or prototype construction.

Text-encoder robustness. We first evaluate whether CFI gives consistent structural rankings of subsets when the semantic representation changes. For each subset budget $b \in \{0.05, 0.10, 0.20\}$, we keep the subset indices fixed and recompute the capability space using three text encoders (`text-embedding-3-large`, `bge-m3`, and `e5-large-v2`). Since the absolute numerical scales across embedding spaces are not directly comparable, we measure cross-encoder robustness by comparing the rankings of the same subsets using Kendall's W. As shown in Fig. 5(a), the subset rankings are highly concordant: Uniformity and Redundancy remain stable across budgets, while Coverage exhibits greater sensitivity to the embedding geometry. Reference-weighted Coverage shows slightly higher concordance than uniform-strength Coverage.

Capability-basis calibration. Next, we isolate the effect of prototype calibration while keeping the same subsets and their evaluation results fixed from Section 4.4. We reconstruct the capability basis using three prototype constructions: raw prototype means, per-prototype ℓ_2-normalization, and Gram whitening. For each construction, we recompute uniform-strength Coverage and Uniformity and measure their Spearman correlations with the subset evaluation measures on MMLU-Pro.

Across all comparisons, Gram whitening yields the strongest associations with the subset evaluation measures: for Coverage, the correlations are $0.596/0.691/0.514$ with $\tau_{\text{cat}}/\tau_{\text{self}}/\text{MAE}$, and for Uniformity they are $0.606/0.663/0.515$. In contrast, ℓ_2-normalization alone produces only marginal changes relative to raw prototypes, suggesting that the main benefit comes from decorrelating prototype directions rather than correcting their magnitudes.

5 Conclusion

In this paper, we view benchmark quality as a question of how measurement information is allocated across capability directions, rather than how many questions are included. We introduce Capability Fisher Information (CFI), a benchmark-internal framework for characterizing benchmark structure through the distribution of measurement information in capability space. CFI provides interpretable diagnostics through Coverage, Uniformity, and Redundancy, which capture capability blind spots, information imbalance, and local overlap among probes. Experiments show that these diagnostics respond to controlled structural defects, are associated with how well benchmark subsets preserve full-benchmark rankings and scores and how stable their rankings remain, and remain robust to key modeling choices. These results suggest that benchmark refinement should focus on information allocation across capabilities rather than simply on benchmark scale.

AI USE STATEMENT

We used generative AI tools to support parts of the manuscript preparation, including (i) language polishing for clarity and grammar, and (ii) limited assistance with code writing and debugging. Importantly, the underlying research questions, methodological design, experiment planning, analysis, and interpretation of results were carried out by the authors. We retain full responsibility for the content and accuracy of the manuscript.

REPRODUCIBILITY STATEMENT

We provide detailed information required to reproduce our experiments, including dataset preprocessing, implementation details, experimental protocols, and key hyperparameter settings. The procedures for subset construction, evaluation, and sensitivity analyses are described in the main text and appendices, together with complete derivations of the theoretical results and additional implementation details to facilitate independent verification and reproduction.

REFERENCES

Amari, S.-i. (2016). *Information geometry and its applications*. Springer.

Bowman, S. and Dahl, G. (2021). What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ding, X., Pan, C., Li, Z., Zhang, J., Wang, S., and Wei, Z. (2026). Autojudger: An agent-driven framework for efficient benchmarking of mllms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15009–15034.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1383–1392.

Embretson, S. and Reise, S. (2004). Item response theory for psychologists.

Ethayarajh, K. and Jurafsky, D. (2020). Utility is in the eye of the user: A critique of nlp leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, volume 22, pages 700–725. Cambridge University Press.

Gao, T., Yao, X., and Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6894–6910.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., and Farhadi, A. (2016). A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., et al. (2021). Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pages 4110–4124.

Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., and Shan, Y. (2024). Seed-bench: Benchmarking multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308. IEEE.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Routledge.

Lord, F. M. and Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

Perlitz, Y., Bandel, E., Gera, A., Arviv, O., Dor, L. E., Shnarch, E., Slonim, N., Shmueli-Scheuer, M., and Choshen, L. (2024). Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2519–2536.

Polo, F. M., Weber, L., Choshen, L., Sun, Y., Xu, G., and Yurochkin, M. (2024). tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*.

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

Qian, Q., Huang, C., Xu, J., Lv, C., Wu, M., Liu, W., Wang, X., Wang, Z., Huang, Z., Tian, M., et al. (2026). Benchmarkˆ 2: Systematic evaluation of llm benchmarks. *arXiv preprint arXiv:2601.03986*.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* ERIC.

Reckase, M. D. (2009). Multidimensional item response theory models. *Multidimensional Item Response Theory*, pages 79–112.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., and Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Settles, B. (2009). Active learning literature survey.

Wang, S., Wang, C., Fu, W., Min, Y., Feng, M., Guan, I., Hu, X., He, C., Wang, C., Yang, K., et al. (2026). Rethinking llm evaluation: Can we evaluate llms with 200× less data? In *International Conference on Learning Representations*, volume 2026, pages 109852–109870.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. (2024). Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Xu, R., Wang, Z., Fan, R.-Z., and Liu, P. (2024). Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S., et al. (2024). Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. (2024). Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Zhou, L., Pacchiardi, L., Martínez-Plumed, F., Collins, K. M., Moros-Daval, Y., Zhang, S., Zhao, Q., Huang, Y., Sun, L., Prunty, J. E., et al. (2026). General scales unlock ai evaluation with explanatory and predictive power. *Nature*, 652(8108):58–67.

CONTENTS OF APPENDIX

A Default Notation **14**

B From Local KL Geometry to Item Capability Information **15**

B.1 Setup and Assumptions 15

B.2 Local KL Expansion and Fisher Information 16

B.3 Specialization to the Multidimensional 2PL Model 17

B.4 Summary 18

C Population-Relative Informativeness **18**

C.1 Principle of Population-Relative Informativeness 18

C.2 Transfer-Error Decomposition 19

C.3 Cross-Population Transfer Study 19

C.4 Design Implications 21

C.5 Summary 22

D Evaluation Metrics **22**

D.1 Group-wise Ranking Fidelity 22

D.2 Group-wise Subsampling Stability 23

D.3 Auxiliary Deviation and Ranking Fidelity 23

E Detailed Implementation of Capability-Direction Acquisition **24**

E.1 Text-Only Benchmark Implementation 24

E.2 Robustness to Text Encoder Choice 25

E.3 Multimodal Benchmark Adaptation 26

E.4 Summary 27

F Implementation Details for Controlled Structural Perturbations **27**

G Implementation Details for Fidelity Evaluation **28**

H Subset Selection Under Budget Constraints **29**

H.1 Four Subset Construction Protocols 29

H.2 CFI-Greedy Construction 30

H.3 Structural and Reliability Trends Across Budget Levels 30

H.4 Relation to Efficient Benchmarking and Subset Selection Literature 32

I Fine-Grained Per-Benchmark Diagnostics **33**

I.1 MMLU-Pro Subject-Group Diagnostics 33

I.2 MATH_{TEST} Topic-Group Diagnostics 34

I.3 GSM8K Latent-Group Diagnostics 36

A DEFAULT NOTATION

Benchmarks, Items, and Capability Spaces	
$\mathcal{B} = \{x_i\}_{i=1}^{n}$	A benchmark containing n evaluation items; each item is denoted by x_i
N	Total item-pool size
$\mathcal{F}$	Global (conceptual) capability space
$\mathcal{F}_{\mathcal{B}} \subseteq \mathcal{F}$	Benchmark-specific local capability subspace
$d = \dim(\mathcal{F}_{\mathcal{B}})$	Dimension / number of target capability directions
$\theta \in \mathbb{R}^d$	A model's latent capability state in $\mathcal{F}_{\mathcal{B}}$
$\Delta \in \mathbb{R}^d$	A small capability perturbation direction/magnitude
Item Responses and Information	
$q_i(y \mid \theta)$	Response outcome distribution of item x_i under capability state θ
$y \in \{0, 1\}$	Binary correctness indicator (0: incorrect, 1: correct)
$D_{\mathrm{KL}}(P\|Q)$	Kullback–Leibler divergence
$g_i(\Delta)$	Local KL change induced by moving from θ to $\theta + \Delta$
$\mathcal{I}_i(\theta)$	Fisher information matrix of item x_i at capability state θ
$\pi_i(\theta) = q_i(1 \mid \theta)$	2PL response probability for correctness
$\sigma(\cdot)$	Logistic sigmoid function
$a_i \in \mathbb{R}^d$	Discrimination vector of item x_i in 2PL
$\beta_i \in \mathbb{R}$	Difficulty parameter of item x_i in 2PL
$\kappa_i(\theta) = \pi_i(\theta)(1 - \pi_i(\theta))$	Uncertainty term in the rank-one Fisher factorization
$\gamma_i = \|a_i\|_2$	Discrimination strength
$u_i = a_i / \|a_i\|_2 \in \mathbb{R}^d$	Discrimination direction
Capability Grouping and Direction Estimation	
$\ell_i \in \{1, \ldots, d\}$	Capability label (group id) of item x_i
$\mathcal{G}_{\mathcal{B}} = \{G_j\}_{j=1}^{d}$	Partition of $\mathcal{B}$ into capability groups
G_j	Set of items whose capability label equals j
$E(\cdot)$	Frozen encoder used to embed items
$z_i \in \mathbb{R}^m$	ℓ_2-normalized embedding of item x_i
b_j	Normalized semantic prototype for capability group G_j
$\bar{z}_j$	Mean embedding within G_j
$B \in \mathbb{R}^{m \times d}$	Prototype matrix
ϵ	Small constant for numerical stability
$\widetilde{B} = B(B^\top B + \epsilon I_d)^{-1/2}$	Regularized Gram whitening of prototypes
$v_i \in \mathbb{R}^d$	Operational probing direction of item x_i
Information Strength and CFI	

α_i	Information strength / item weight for x_i
$\widehat{\mathcal{I}}_i = \alpha_i v_i v_i^\top$	Operational item information matrix
$\mathrm{CFI}(\mathcal{B}, \alpha)$	Capability Fisher Information of benchmark $\mathcal{B}$
Structural Diagnostics (Coverage / Uniformity / Redundancy)	
$\mathrm{Cov}(\mathcal{B}, \alpha)$	Coverage: smallest eigenvalue of CFI
$\mathrm{Unif}(\mathcal{B}, \alpha)$	Uniformity: det-to-trace ratio of CFI eigenvalues
$\mathrm{Red}(\mathcal{B})$	Redundancy: average max pairwise redundancy
$r_{ij} = (v_i^\top v_j)^2 \,\mathrm{sim}(i, j)$	Pairwise redundancy term
$\mathrm{sim}(i, j) = \max\{0, z_i^\top z_j\}$	Nonnegative semantic similarity between items x_i and x_j
$\det(\cdot)$	Determinant of a matrix
$\mathrm{tr}(\cdot)$	Trace of a matrix

B From Local KL Geometry to Item Capability Information

This appendix derives the local second-order relationship between KL divergence and Fisher information, and then specializes it to the multidimensional 2PL model introduced in Section 3.2. The resulting item information has the rank-one form $\mathcal{I}_i(\theta) = \kappa_i(\theta)\gamma_i^2 u_i u_i^\top$, which separates the capability direction from the corresponding information strength.

B.1 Setup and Assumptions

Let $\theta \in \mathbb{R}^d$ denote a capability state in the benchmark-specific capability space $\mathcal{F}_\mathcal{B}$. Each item x_i induces a conditional response distribution $q_i(y \mid \theta)$ over an observable response variable Y_i. For a continuous response variable, q_i is understood as a probability density.

We use the following regularity assumptions:

(A1) *Smoothness and local positivity.* For every response value y in the support of q_i, the function $q_i(y \mid \theta)$ is twice continuously differentiable in a neighborhood of θ and is strictly positive throughout that neighborhood.

(A2) *Fixed support and valid differentiation.* The support of $q_i(\cdot \mid \theta)$ does not depend on θ. Differentiation may be exchanged with summation, or with integration when the response variable is continuous.

(A3) *Finite and locally controlled derivatives.* The score has a finite second moment,

$$\mathbb{E}_{Y_i \sim q_i(\cdot \mid \theta)} \left[\left\| \nabla_\theta \log q_i(Y_i \mid \theta) \right\|_2^2 \right] < \infty,$$

and the relevant second derivatives are locally dominated by an integrable function. These conditions justify the Taylor expansion and the information identity used below.

Under the multidimensional 2PL model, the response is binary, $Y_i \in \{0, 1\}$, and the probability of a correct response is

$$\pi_i(\theta) \triangleq q_i(1 \mid \theta) = \sigma\big(a_i^\top \theta - \beta_i\big), \qquad q_i(0 \mid \theta) = 1 - \pi_i(\theta), \tag{17}$$

where $a_i \in \mathbb{R}^d$ is the discrimination vector, $\beta_i \in \mathbb{R}$ is the difficulty parameter, and $\sigma(t) = 1/(1 + \exp(-t))$ is the logistic function.

B.2 Local KL Expansion and Fisher Information

For item x_i, define the KL divergence between the response distributions at the nearby capability states θ and $\theta + \Delta$ as

$$\begin{aligned} g_i(\Delta) &\triangleq D_{\mathrm{KL}}\left(q_i(\cdot \mid \theta) \,\|\, q_i(\cdot \mid \theta + \Delta)\right) \\ &= \mathbb{E}_{Y_i \sim q_i(\cdot|\theta)}\left[\log \frac{q_i(Y_i \mid \theta)}{q_i(Y_i \mid \theta + \Delta)}\right]. \end{aligned} \tag{18}$$

A second-order Taylor expansion around $\Delta = 0$ gives

$$g_i(\Delta) = g_i(0) + \nabla_\Delta g_i(0)^\top \Delta + \frac{1}{2}\Delta^\top \nabla_\Delta^2 g_i(0)\Delta + o\big(\|\Delta\|_2^2\big). \tag{19}$$

We evaluate each term in this expansion below.

Zeroth-order term. Because the two distributions coincide at $\Delta = 0$, we have $g_i(0) = 0$.

First-order term. Differentiating with respect to Δ gives

$$\nabla_\Delta g_i(\Delta) = -\mathbb{E}_{Y_i \sim q_i(\cdot|\theta)}\left[\nabla_\Delta \log q_i(Y_i \mid \theta + \Delta)\right].$$

At $\Delta = 0$, this becomes

$$\nabla_\Delta g_i(0) = -\mathbb{E}_{Y_i \sim q_i(\cdot|\theta)}\left[\nabla_\theta \log q_i(Y_i \mid \theta)\right].$$

The expectation on the right is zero. For a discrete response variable,

$$\begin{aligned} \mathbb{E}\left[\nabla_\theta \log q_i(Y_i \mid \theta)\right] &= \sum_y q_i(y \mid \theta) \nabla_\theta \log q_i(y \mid \theta) \\ &= \sum_y \nabla_\theta q_i(y \mid \theta) \\ &= \nabla_\theta \sum_y q_i(y \mid \theta) \\ &= \nabla_\theta 1 = 0. \end{aligned}$$

The same argument applies with integration when the response variable is continuous. Therefore, $\nabla_\Delta g_i(0) = 0$.

Second-order term. Differentiating again and evaluating at $\Delta = 0$ yields

$$\nabla_\Delta^2 g_i(0) = -\mathbb{E}_{Y_i \sim q_i(\cdot|\theta)}\left[\nabla_\theta^2 \log q_i(Y_i \mid \theta)\right].$$

Define the score function by $s_i(y; \theta) = \nabla_\theta \log q_i(y \mid \theta)$. The item-level Fisher information matrix is

$$\mathcal{I}_i(\theta) \triangleq \mathbb{E}_{Y_i \sim q_i(\cdot|\theta)}\left[s_i(Y_i; \theta) s_i(Y_i; \theta)^\top\right]. \tag{20}$$

For a discrete response variable, this can be written as

$$\mathcal{I}_i(\theta) = \sum_y q_i(y \mid \theta) \nabla_\theta \log q_i(y \mid \theta) \nabla_\theta \log q_i(y \mid \theta)^\top,$$

which agrees with Eq. 4.

Under Assumptions (A1)–(A3), the information identity gives

$$\mathcal{I}_i(\theta) = -\mathbb{E}_{Y_i \sim q_i(\cdot|\theta)}\left[\nabla_\theta^2 \log q_i(Y_i \mid \theta)\right]. \tag{21}$$

For completeness, consider the discrete case. The Hessian of the log-probability is

$$\nabla_\theta^2 \log q_i(y \mid \theta) = \frac{\nabla_\theta^2 q_i(y \mid \theta)}{q_i(y \mid \theta)} - \frac{\nabla_\theta q_i(y \mid \theta) \nabla_\theta q_i(y \mid \theta)^\top}{q_i(y \mid \theta)^2}.$$

Multiplying by $q_i(y \mid \theta)$ and summing over y gives

$$\begin{aligned}\mathbb{E}\left[\nabla_\theta^2 \log q_i(Y_i \mid \theta)\right] &= \sum_y \nabla_\theta^2 q_i(y \mid \theta) - \sum_y \frac{\nabla_\theta q_i(y \mid \theta) \nabla_\theta q_i(y \mid \theta)^\top}{q_i(y \mid \theta)} \\ &= \nabla_\theta^2 \sum_y q_i(y \mid \theta) - \mathcal{I}_i(\theta) \\ &= -\mathcal{I}_i(\theta).\end{aligned}$$

The continuous case follows by replacing sums with integrals.

It follows from Eq. 21 that $\nabla_\Delta^2 g_i(0) = \mathcal{I}_i(\theta)$. Substitution into Eq. 19 yields

$$D_{\mathrm{KL}}\left(q_i(\cdot \mid \theta) \,\|\, q_i(\cdot \mid \theta + \Delta)\right) = \frac{1}{2}\Delta^\top \mathcal{I}_i(\theta)\Delta + o\big(\|\Delta\|_2^2\big), \qquad \|\Delta\|_2 \to 0. \tag{22}$$

Thus, the item-level Fisher information matrix describes the local curvature of the KL divergence with respect to changes in capability state.

B.3 Specialization to the Multidimensional 2PL Model

We now specialize the general result to the multidimensional 2PL response model in Eq. 17.

Gradient of the correct-response probability. Let $t_i(\theta) = a_i^\top \theta - \beta_i$ and define

$$\kappa_i(\theta) \triangleq \pi_i(\theta)\big(1 - \pi_i(\theta)\big).$$

Because $\sigma'(t) = \sigma(t)(1 - \sigma(t))$, the gradient of the correct-response probability is

$$\begin{aligned}\nabla_\theta \pi_i(\theta) &= \sigma'(t_i(\theta))\, \nabla_\theta t_i(\theta) \\ &= \pi_i(\theta)\big(1 - \pi_i(\theta)\big) a_i \\ &= \kappa_i(\theta) a_i.\end{aligned}$$

Score function under the 2PL model. For a binary response $y \in \{0, 1\}$,

$$\log q_i(y \mid \theta) = y \log \pi_i(\theta) + (1 - y) \log\big(1 - \pi_i(\theta)\big).$$

Differentiating with respect to θ gives

$$\begin{aligned}s_i(y; \theta) &= \nabla_\theta \log q_i(y \mid \theta) \\ &= \left(\frac{y}{\pi_i(\theta)} - \frac{1 - y}{1 - \pi_i(\theta)}\right) \nabla_\theta \pi_i(\theta) \\ &= \frac{y - \pi_i(\theta)}{\pi_i(\theta)(1 - \pi_i(\theta))} \nabla_\theta \pi_i(\theta) \\ &= \big(y - \pi_i(\theta)\big) a_i.\end{aligned} \tag{23}$$

Rank-one Fisher information. Substituting Eq. 23 into the Fisher information definition gives

$$\begin{aligned}\mathcal{I}_i(\theta) &= \mathbb{E}\left[s_i(Y_i; \theta) s_i(Y_i; \theta)^\top\right] \\ &= \mathbb{E}\left[\big(Y_i - \pi_i(\theta)\big)^2\right] a_i a_i^\top \\ &= \mathrm{Var}(Y_i \mid \theta) a_i a_i^\top \\ &= \pi_i(\theta)\big(1 - \pi_i(\theta)\big) a_i a_i^\top \\ &= \kappa_i(\theta) a_i a_i^\top.\end{aligned} \tag{24}$$

For $a_i \neq 0$, let $\gamma_i = \|a_i\|_2$ and $u_i = a_i / \gamma_i$. Then $a_i = \gamma_i u_i$ and $\|u_i\|_2 = 1$. Eq. 24 becomes

$$\boxed{\mathcal{I}_i(\theta) = \underbrace{\kappa_i(\theta)\gamma_i^2}_{\text{information strength}} \underbrace{u_i u_i^\top}_{\text{capability subspace}}}. \tag{25}$$

For finite θ, the logistic model satisfies $0 < \pi_i(\theta) < 1$. Therefore, if $a_i \neq 0$, the information matrix has rank one. Its unique nonzero eigenvalue is $\kappa_i(\theta)\gamma_i^2$, and the corresponding eigenspace is spanned by u_i. The sign of u_i is immaterial because the outer product $u_i u_i^\top$ is unchanged under $u_i \mapsto -u_i$.

If $a_i = 0$, then $\mathcal{I}_i(\theta) = 0$, and no unit capability direction needs to be assigned. The decomposition in Eq. 25 therefore separates the one-dimensional capability subspace associated with the item from the amount of information it contributes locally.

This decomposition motivates the operational construction in Section 3.3. Because the latent direction u_i and the discrimination magnitude γ_i are unavailable without a calibrated response model, CFI uses the frame-consistent semantic direction v_i and the nonnegative strength proxy α_i. The resulting operational item information, $\widehat{\mathcal{I}}_i = \alpha_i v_i v_i^\top$, is defined formally in Eq. 11.

B.4 Summary

Under the stated regularity assumptions, the local curvature of the item-level KL divergence is given by the Fisher information matrix. For the multidimensional 2PL model, this matrix has the rank-one form $\mathcal{I}_i(\theta) = \kappa_i(\theta)\gamma_i^2 u_i u_i^\top$. This decomposition separates the item's capability subspace from its information strength and provides the theoretical motivation for the operational approximation $\widehat{\mathcal{I}}_i = \alpha_i v_i v_i^\top$ used by CFI.

C Population-Relative Informativeness

A key property of reference-weighted CFI is its dependence on the seed-model pool: different choices of reference models can induce different informativeness weights for the same items, changing which items are regarded as informative and, consequently, how measurement effort is allocated in benchmark construction. We interpret this dependence through the notion of *population-relative informativeness*: the informativeness of an evaluation depends on the model population for which it is used. Under this view, seed-model dependence reflects the population-relative nature of item informativeness rather than an implementation artifact.

C.1 Principle of Population-Relative Informativeness

> **The informativeness of a benchmark is not absolute, but depends on the capability distribution of the model population being evaluated.**

This population dependence is consistent with recent work on psychometric approaches to AI evaluation, which observes that extracted factors and parameters can depend substantially on the population of systems and benchmarks under consideration (Zhou et al., 2026). It is also consistent with the local nature of Fisher information. An item is informative only insofar as model responses remain sensitive to capability differences in the population of interest. Items that are uniformly easy or uniformly difficult provide little discrimination within that population, whereas items whose difficulty is better matched to the population's capability range provide stronger response variation.

Mechanism in LLM evaluation. Consider an item x_i with response probability $p_i(\theta)$. Its response-dependent information factor is

$$\kappa_i(\theta) = p_i(\theta)\big(1 - p_i(\theta)\big), \tag{26}$$

which is maximized when $p_i(\theta) = 1/2$ and decreases toward zero as $p_i(\theta)$ approaches either 0 or 1. Thus, the same item can be informative for one model population but uninformative for another: an item whose difficulty is well matched to a population can distinguish models within that population, whereas the same item may exhibit ceiling or floor effects for substantially stronger or weaker populations.

Implications of reference-model dependence. The resulting weights should be interpreted relative to the capability range represented by the reference models rather than as universal item properties. A pool spanning a broader range of relevant capability levels emphasizes discrimination over a broader reference population, whereas a narrowly concentrated one emphasizes discrimination within a more

local capability range. This dependence motivates reporting the composition of the reference pool and evaluating sensitivity to alternative choices of such models.

C.2 Transfer-Error Decomposition

Let R and T denote the reference and target model populations. For item i, define the population-level success probabilities

$$p_i^R = \mathbb{E}_{m\sim R}[y_{m,i}], \qquad p_i^T = \mathbb{E}_{m\sim T}[y_{m,i}], \tag{27}$$

and the corresponding operational informativeness weights

$$\alpha_i^R = f(p_i^R), \qquad \alpha_i^T = f(p_i^T), \qquad f(p) = p(1-p). \tag{28}$$

For a finite reference pool $\mathcal{M}_R = \{m_1, \dots, m_{n_R}\}$, we use

$$\hat{p}_i^R = \frac{1}{n_R} \sum_{m\in\mathcal{M}_R} y_{m,i}, \qquad \hat{\alpha}_i^R = f(\hat{p}_i^R). \tag{29}$$

The transfer error satisfies

$$|\hat{\alpha}_i^R - \alpha_i^T| \le \underbrace{|\hat{\alpha}_i^R - \alpha_i^R|}_{\text{finite-pool error}} + \underbrace{|\alpha_i^R - \alpha_i^T|}_{\text{population-shift error}} . \tag{30}$$

Because $f(p) = p(1-p)$ is 1-Lipschitz on $[0,1]$,

$$|\hat{\alpha}_i^R - \alpha_i^R| \le |\hat{p}_i^R - p_i^R|, \qquad |\alpha_i^R - \alpha_i^T| \le |p_i^R - p_i^T|. \tag{31}$$

If the reference models are idealized as n_R i.i.d. draws from R, Hoeffding's inequality and a union bound over the N items give

$$\max_{1\le i\le N} |\hat{p}_i^R - p_i^R| \le \sqrt{\frac{\log(2N/\delta)}{2n_R}} \qquad \text{with probability at least } 1-\delta. \tag{32}$$

Define the worst-case itemwise response-rate shift

$$D(R,T) = \max_i |p_i^R - p_i^T|. \tag{33}$$

Then, simultaneously for all items,

$$|\hat{\alpha}_i^R - \alpha_i^T| \le \sqrt{\frac{\log(2N/\delta)}{2n_R}} + D(R,T). \tag{34}$$

Thus, under the i.i.d. sampling idealization, increasing the reference pool size tightens the finite-pool term but does not remove mismatch between the reference and target populations.

Since the operational capability directions satisfy $\|v_i\|_2 = 1$, for any subset S,

$$\left\| \sum_{i\in S} \hat{\alpha}_i^R v_i v_i^\top - \sum_{i\in S} \alpha_i^T v_i v_i^\top \right\|_2 \le |S| \left(\sqrt{\frac{\log(2N/\delta)}{2n_R}} + D(R,T) \right). \tag{35}$$

C.3 Cross-Population Transfer Study

To probe the role of population mismatch in transfer, we compare mixed-population and scale-separated reference–target pairs.

C.3.1 Setup

Model partitions We construct two alternative reference–target partitions from the same model collection. The scale-separated partition forms a smaller-scale pool $\mathcal{W}$ and a larger-scale pool $\mathcal{S}$, while the mixed partition forms two disjoint pools $\mathcal{P}_A$ and $\mathcal{P}_B$, each spanning a broad range of model scales. Because model scale is only a coarse proxy for capability, these partitions are intended to induce different degrees of population alignment rather than an objective weak–strong separation. The resulting pools are listed in Table 2.

Table 2: Model pools used in the cross-population study. Panel (a) separates models by scale to form smaller-scale and larger-scale pools, while Panel (b) forms two mixed-scale pools, each spanning a broader range of model scales. Throughout the main experiments, $\mathcal{P}_A$ **contains the default reference models used to estimate reference-weighted item informativeness**, whereas $\mathcal{P}_B$ **contains the default held-out models used for reliability evaluation**.

Smaller-scale pool ($\mathcal{W}$)	**Larger-scale pool** ($\mathcal{S}$)
gemma-3-4b-it	GLM-4-9B
Ministral-3-3B-Instruct	gemma-3-12b-it
Ministral-3-8B-Instruct	gemma-3-27b-it
Mistral-7B-Instruct-v0.3	Ministral-3-14B-Instruct
Qwen3-4B	Qwen2.5-14B-Instruct
Qwen3-8B	Phi-4
Phi-4-mini-instruct	Qwen2.5-32B-Instruct
Llama-3.2-1B-Instruct	GLM-4-32B
Llama-3.2-3B-Instruct	Seed-OSS-36B-Instruct
Llama-3.1-8B-Instruct	Llama-3.3-70B-Instruct

(a) Scale-separated model pools

Mixed A ($\mathcal{P}_A$)	**Mixed B** ($\mathcal{P}_B$)
gemma-3-4b-it	Ministral-3-8B-Instruct
Ministral-3-3B-Instruct	Qwen3-8B
Mistral-7B-Instruct-v0.3	Phi-4-mini-instruct
Qwen3-4B	Llama-3.2-3B-Instruct
GLM-4-9B	Llama-3.1-8B-Instruct
Llama-3.2-1B-Instruct	gemma-3-12b-it
Qwen2.5-14B-Instruct	gemma-3-27b-it
Ministral-3-14B-Instruct	Phi-4
Qwen2.5-32B-Instruct	GLM-4-32B
Seed-OSS-36B-Instruct	Llama-3.3-70B-Instruct

(b) Mixed-capability model pools

Construction protocol For each finite reference pool $\mathcal{M}_R \in \{\mathcal{W}, \mathcal{S}, \mathcal{P}_A, \mathcal{P}_B\}$, we estimate the reference-weighted item informativeness $\hat{\alpha}_i^R$ according to Eq. 29. For each of the 50 construction seeds s, we then run CFI-Greedy with candidate pool size $p = 500$ and $\beta = 0$ to construct an optimized subset $S_R^{(s)}$, which is evaluated on the corresponding prespecified target pool. We consider four reference–target transfers: $\mathcal{P}_A \to \mathcal{P}_B$, $\mathcal{P}_B \to \mathcal{P}_A$, $\mathcal{W} \to \mathcal{S}$, and $\mathcal{S} \to \mathcal{W}$, yielding four reference–target pairs per benchmark.

Alignment metric We quantify reference–target similarity using the Spearman rank correlation between their empirical item-informativeness vectors:

$$A(R,T) = \text{Spearman}\left((\hat{\alpha}_i^R)_{i=1}^N, (\hat{\alpha}_i^T)_{i=1}^N\right). \tag{36}$$

where the target-pool weights are computed analogously to the reference-pool weights and are used only for post-hoc analysis, not for subset construction. Higher $A(R,T)$ indicates greater agreement in item-informativeness rankings between the two pools. Note that $A(R,T) = A(T,R)$; transfer direction enters through which pool is used for subset construction and which serves as the target.

C.3.2 Results

Table 3: **Cross-population transfer results (budget ratio $b = 0.1$, 50 construction seeds).** Score denotes the mean accuracy of target models on the constructed subset. For Score, τ_{cat}, and τ_{self}, we report the mean and standard deviation across construction seeds.

Benchmark	**Ref.**	**Target**	$A(R,T)$	**Score** (%)	τ_{cat}	τ_{self}
MMLU-Pro	$\mathcal{P}_A$	$\mathcal{P}_B$	0.55	63.5 ± 0.8	0.861 ± 0.019	0.908 ± 0.008
	$\mathcal{P}_B$	$\mathcal{P}_A$	0.55	50.6 ± 0.7	0.899 ± 0.015	0.928 ± 0.006
	$\mathcal{W}$	$\mathcal{S}$	0.20	74.4 ± 0.8	0.733 ± 0.027	0.832 ± 0.012
	$\mathcal{S}$	$\mathcal{W}$	0.20	35.1 ± 0.6	0.809 ± 0.017	0.884 ± 0.007
MATH$_{TEST}$	$\mathcal{P}_A$	$\mathcal{P}_B$	0.62	71.4 ± 1.0	0.875 ± 0.027	0.933 ± 0.008
	$\mathcal{P}_B$	$\mathcal{P}_A$	0.62	59.5 ± 0.9	0.870 ± 0.027	0.915 ± 0.010
	$\mathcal{W}$	$\mathcal{S}$	0.34	80.4 ± 0.8	0.837 ± 0.030	0.904 ± 0.010
	$\mathcal{S}$	$\mathcal{W}$	0.34	47.5 ± 0.9	0.901 ± 0.017	0.944 ± 0.007
GSM8K	$\mathcal{P}_A$	$\mathcal{P}_B$	0.45	89.8 ± 0.5	0.889 ± 0.028	0.921 ± 0.015
	$\mathcal{P}_B$	$\mathcal{P}_A$	0.45	66.4 ± 0.5	0.917 ± 0.021	0.943 ± 0.012
	$\mathcal{W}$	$\mathcal{S}$	0.37	92.3 ± 0.5	0.752 ± 0.055	0.831 ± 0.025
	$\mathcal{S}$	$\mathcal{W}$	0.37	61.8 ± 0.4	0.826 ± 0.030	0.885 ± 0.017

Observations. As shown in Table 3, the mixed-population transfers exhibit consistently higher informativeness-ranking alignment than the scale-separated transfers ($A = 0.45$–0.62 versus 0.20–

0.37). Averaged across benchmarks and transfer directions, this difference is accompanied by higher ranking fidelity: τ_{cat} decreases from 0.885 to 0.810, while τ_{self} decreases from 0.925 to 0.880.

The effect is not uniform across individual transfers. In particular, the $\mathcal{S} \to \mathcal{W}$ transfer on MATH$_{TEST}$ retains high ranking fidelity ($\tau_{\text{cat}} = 0.901$, $\tau_{\text{self}} = 0.944$) despite relatively low informativeness alignment. Thus, lower alignment is associated with weaker transfer on average, but is not sufficient to predict transfer quality in every case.

The $\mathcal{W} \to \mathcal{S}$ transfers yield consistently high mean scores (74.4–92.3), consistent with a tendency toward ceiling effects in the larger-scale target pool. The reverse direction does not exhibit a corresponding uniform floor, with mean scores ranging from 35.1 to 61.8.

Interpretation. These results are consistent with the hypothesis motivated by Eq. 30: transfer tends to be more reliable when the reference and target populations induce more similar informativeness profiles. However, $A(R, T)$ is a rank-based empirical diagnostic and does not estimate the worst-case response-rate shift $D(R, T)$. The experiment should therefore be interpreted as evidence of an aggregate association rather than a direct verification of the bound or a deterministic relationship between alignment and ranking fidelity. The high-fidelity $\mathcal{S} \to \mathcal{W}$ result on MATH$_{TEST}$ illustrates this distinction.

C.4 Design Implications

The population-relative nature of item informativeness implies that benchmark design should make explicit the capability range of the model population it is intended to evaluate. In practice, benchmark construction relies on a set of reference models to estimate item informativeness and construct an optimized subset, whereas deployment involves evaluation on a target model population. When the reference and target populations are better aligned, our results show more reliable ranking transfer on average; under substantial population mismatch, transfer quality may degrade, although the magnitude and direction of the effect are benchmark-dependent.

1. Specify the reference models used for construction. Benchmark builders should report the models used to estimate item informativeness and construct the optimized subset. At minimum, they should report the model families and scales and summarize the capability range represented by the reference pool. This information allows users to assess whether their target models are well aligned with the reference population used for benchmark construction.

2. Quantify alignment for the target populations of interest. Before release, benchmark curators should assess how well the item informativeness estimated from the reference models transfers to held-out target populations. A simple diagnostic is to compute informativeness-ranking alignment and verify that ranking fidelity remains stable across target populations. If alignment is low or scores exhibit pronounced ceiling or floor effects, the benchmark should be released with clear guidance on its intended use.

3. Provide stratified suites rather than a single one-size-fits-all benchmark. A single fixed benchmark is unlikely to be equally informative for all model populations. If the goal is to rank flagship models, the reference model pool should emphasize strong models. If the goal is to distinguish smaller models, a weaker reference model pool is more appropriate. If broad coverage is desired, the reference models should span weak, intermediate, and strong capability ranges. Because these trade-offs depend on the use case, multiple suites optimized for different reference populations may be more useful than a single universal benchmark.

4. Temper universality claims. Statements such as "this benchmark measures capability X" should be qualified. A more defensible claim is that the benchmark is designed and validated to distinguish models within the capability range represented by its reference population, while its informativeness may change when applied to substantially different model populations.

C.5 SUMMARY

Seed-model dependence is an explicit conditioning factor of reference-weighted CFI: the resulting item weights are relative to the capability range represented by the reference population. Our cross-population study shows that better-aligned mixed-population reference–target pairs exhibit higher informativeness alignment and stronger ranking transfer on average than scale-separated pairs, although the relationship is not deterministic and varies across benchmarks and transfer directions.

These results motivate treating population alignment as part of benchmark validation rather than assuming that item informativeness transfers universally. Benchmark builders should document the reference model pool, evaluate transfer on representative held-out populations, and consider population-specific or stratified benchmark suites when the intended model range is broad.

D EVALUATION METRICS

This appendix defines the held-out reliability metrics used in the main experiments and the accompanying analyses. The benchmark-internal structural metrics, including Coverage, Uniformity, and Redundancy, are defined in Section 3.4.

Let $\mathcal{B} = \{x_i\}_{i=1}^n$ denote the full benchmark, and let $S \subseteq \mathcal{B}$ be a selected subset. We use $\mathcal{M}_{\text{eval}}$ to denote the held-out models used to evaluate subset reliability. This set is distinct from the seed-model set $\mathcal{M}$ used to construct the reference-weighted strength proxy α_i in Section 3.3.2.

The benchmark items are partitioned into the capability groups $\mathcal{G}_{\mathcal{B}} = \{G_j\}_{j=1}^d$ defined in Eq. 8. For each group, let $S_j = S \cap G_j$ denote the items from that group included in the selected subset. When native benchmark annotations are available, the groups correspond to the associated categories or topics. Otherwise, they correspond to the benchmark-specific groups constructed during capability-space acquisition. The metrics below can therefore be applied to native categories, mathematical topics, or clustering-induced capability groups.

For each held-out model $m \in \mathcal{M}_{\text{eval}}$, let $y_{m,i} \in \{0, 1\}$ indicate whether model m answers item x_i correctly. For any non-empty item set $A \subseteq \mathcal{B}$, its model-specific accuracy is

$$\text{Acc}_m(A) = \frac{1}{|A|} \sum_{x_i \in A} y_{m,i}. \tag{37}$$

When $A = G_j$ or $A = S_j$, this quantity represents the group-level accuracy of model m.

Kendall's tau-b. We use Kendall's tau-b, denoted by τ_b, to compare model rankings while accounting for ties. Consider two score functions u_m and v_m defined on the same model set. For each unordered model pair (m_a, m_b), let $\Delta_{ab}^u = u_{m_a} - u_{m_b}$ and $\Delta_{ab}^v = v_{m_a} - v_{m_b}$.

A pair is concordant if $\Delta_{ab}^u \Delta_{ab}^v > 0$ and discordant if $\Delta_{ab}^u \Delta_{ab}^v < 0$. It is tied only under u if $\Delta_{ab}^u = 0$ and $\Delta_{ab}^v \neq 0$, and tied only under v if $\Delta_{ab}^v = 0$ and $\Delta_{ab}^u \neq 0$. Let n_c, n_d, n_u, and n_v denote the corresponding numbers of pairs. Kendall's tau-b is

$$\tau_b(u, v) = \frac{n_c - n_d}{\sqrt{(n_c + n_d + n_u)(n_c + n_d + n_v)}}. \tag{38}$$

Pairs tied under both score functions do not contribute to either the numerator or the denominator. For evaluation purposes, we assign $\tau_b(u, v) = 0$ when the denominator is zero. This convention is relevant when all held-out models receive identical scores, which is more likely for small subsets.

D.1 GROUP-WISE RANKING FIDELITY

Our primary held-out ranking metric is group-wise ranking fidelity, denoted by τ_{cat}. The subscript "cat" refers to aggregation over benchmark groups and does not require those groups to be predefined categories.

For each capability group G_j, the full benchmark and the selected subset produce the held-out model scores $u_m^{(j)} = \text{Acc}_m(G_j)$ and $v_m^{(j)} = \text{Acc}_m(S_j)$, respectively. Their group-level ranking agreement

is

$$\tau_j(S) = \tau_b\left(\{u_m^{(j)}\}_{m\in\mathcal{M}_{\text{eval}}}, \{v_m^{(j)}\}_{m\in\mathcal{M}_{\text{eval}}}\right). \tag{39}$$

We then average the group-level scores without weighting them by group size:

$$\tau_{\text{cat}}(S) = \frac{1}{d}\sum_{j=1}^{d}\tau_j(S). \tag{40}$$

The full benchmark serves as the empirical reference for this comparison. Because the aggregation is performed across groups, large groups do not dominate the reliability estimate. If $|S_j| < n_{\min}$, the selected subset is considered insufficient for a reliable group-level comparison, and we set $\tau_j(S) = 0$. We use $n_{\min} = 3$ throughout the experiments. The general zero-value convention for undefined Kendall correlations also applies when all held-out models receive identical group-level scores.

D.2 Group-wise Subsampling Stability

The second primary reliability metric is group-wise subsampling stability, denoted by τ_{self}. It measures whether a selected subset produces stable model rankings under repeated subsampling, without directly comparing those rankings with the full benchmark.

Given a selected subset S, we generate R subsamples $S^{(1)}, \ldots, S^{(R)}$ by sampling without replacement from S. Each subsample contains a fraction ρ of the items in S. For capability group G_j, let $S_j^{(r)} = S^{(r)} \cap G_j$. Whenever this set is non-empty, the corresponding score of model m is $u_{m,r}^{(j)} = \text{Acc}_m(S_j^{(r)})$.

For each group, the valid subsamples are those containing at least $n_{\min}$ items:

$$\mathcal{R}_j = \{r \in \{1, \ldots, R\} : |S_j^{(r)}| \geq n_{\min}\}.$$

When $|\mathcal{R}_j| \geq 2$, we define the stability of group G_j as the mean pairwise ranking agreement among its valid subsamples:

$$\text{Stab}_j(S) = \frac{2}{|\mathcal{R}_j|(|\mathcal{R}_j|-1)}\sum_{\substack{r,s\in\mathcal{R}_j \\ r<s}}\tau_b\left(\{u_{m,r}^{(j)}\}_{m\in\mathcal{M}_{\text{eval}}}, \{u_{m,s}^{(j)}\}_{m\in\mathcal{M}_{\text{eval}}}\right). \tag{41}$$

If fewer than two valid subsamples are available, we set $\text{Stab}_j(S) = 0$.

The final stability score is the unweighted average across capability groups:

$$\tau_{\text{self}}(S) = \frac{1}{d}\sum_{j=1}^{d}\text{Stab}_j(S). \tag{42}$$

Larger values indicate that the selected subset produces more stable held-out model rankings under subsampling. Unless otherwise specified, we use $R = 100$, $\rho = 0.8$, and $n_{\min} = 3$.

D.3 Auxiliary Deviation and Ranking Fidelity

In addition to the primary metrics τ_{cat} and τ_{self}, we report several auxiliary metrics. They provide complementary views of how well a selected subset reproduces model comparisons obtained from the full benchmark.

Overall mean absolute error. Overall mean absolute error measures the difference between each held-out model's full-benchmark accuracy and selected-subset accuracy:

$$\text{MAE}(S) = \frac{1}{|\mathcal{M}_{\text{eval}}|}\sum_{m\in\mathcal{M}_{\text{eval}}}\left|\text{Acc}_m(\mathcal{B}) - \text{Acc}_m(S)\right|. \tag{43}$$

Lower values indicate closer agreement between subset-level and full-benchmark performance.

Overall ranking fidelity. Overall ranking fidelity measures whether the selected subset preserves the held-out model ordering induced by the full benchmark:

$$\tau_{\text{overall}}(S) = \tau_b\left(\{\text{Acc}_m(\mathcal{B})\}_{m\in\mathcal{M}_{\text{eval}}}, \{\text{Acc}_m(S)\}_{m\in\mathcal{M}_{\text{eval}}}\right). \tag{44}$$

Tail ranking fidelity. Tail ranking fidelity focuses on lower-performing held-out models. Let $\rho_{\text{tail}} \in (0, 1]$ denote the fraction of models included in the tail. We rank the models by their full-benchmark accuracies and let $\mathcal{M}_{\text{tail}}(\rho_{\text{tail}})$ contain the bottom $\lceil \rho_{\text{tail}}|\mathcal{M}_{\text{eval}}| \rceil$ models. Ties at the selection boundary are resolved using a fixed deterministic rule.

Tail ranking fidelity is then defined as

$$\tau_{\text{tail}}(S; \rho_{\text{tail}}) = \tau_b\left(\{\text{Acc}_m(\mathcal{B})\}_{m\in\mathcal{M}_{\text{tail}}(\rho_{\text{tail}})}, \{\text{Acc}_m(S)\}_{m\in\mathcal{M}_{\text{tail}}(\rho_{\text{tail}})}\right). \tag{45}$$

We use $\rho_{\text{tail}} = 0.2$ in our experiments. Larger values indicate better preservation of the relative ordering among lower-performing models.

Unlike τ_{cat}, which assigns equal importance to each capability group, τ_{overall} aggregates all benchmark items and may therefore be influenced more strongly by large or easy regions. Tail ranking fidelity isolates comparisons among lower-performing models, but it may be noisier because it is computed over a smaller model set. Accordingly, we treat these measures as auxiliary diagnostics and use τ_{cat} and τ_{self} as the primary held-out reliability metrics.

E Detailed Implementation of Capability-Direction Acquisition

This appendix provides implementation details for constructing the benchmark-specific capability frame and the operational probing directions v_i introduced in Section 3.3. We describe the procedure for text-only benchmarks, including those without predefined capability labels, and its extension to multimodal benchmarks. The resulting v_i serves as a frame-consistent semantic proxy for the latent information direction u_i.

E.1 Text-Only Benchmark Implementation

For text-only benchmarks, capability-direction acquisition consists of four steps: semantic embedding, capability grouping and prototype construction, Gram whitening, and projection into the calibrated capability frame.

E.1.1 Semantic Embedding

We use a frozen semantic encoder, specifically `text-embedding-3-large`, to map the textual content of each item x_i to an embedding $E(x_i) \in \mathbb{R}^m$. The embedding dimension m is determined by the encoder and is not constrained by the CFI framework. Consistent with Section 3.3, each embedding is ℓ_2-normalized as

$$z_i = \frac{E(x_i)}{\|E(x_i)\|_2} \in \mathbb{R}^m.$$

E.1.2 Capability Grouping and Prototype Construction

We partition the benchmark items into d capability groups and construct one semantic prototype for each group.

Benchmarks with predefined capability labels. When the benchmark provides a native taxonomy, such as subject or domain labels, we use these labels directly. In this case, ℓ_i denotes the native capability label of item x_i, and d is the number of benchmark-defined categories.

Benchmarks without predefined capability labels. When no native taxonomy is available, we derive benchmark-specific capability groups from the item embeddings. Specifically, we apply PCA (Hotelling, 1933) to the normalized embeddings $\{z_i\}_{i=1}^n$, retain at most 64 principal components, and ℓ_2-normalize the resulting features. We then apply KMeans clustering in this reduced space.

The capability dimension is selected from $\mathcal{D} = \{3, 4, 5, 6, 7, 8\}$ according to the mean cosine silhouette score across 20 random seeds, with each KMeans run using `n_init=20`. After selecting the optimal dimension d^*, we retain the highest-scoring partition obtained for d^* and use its assignments as pseudo-category labels ℓ_i, setting $d = d^*$. Averaging silhouette scores across seeds reduces sensitivity to random initialization.

The PCA features are used only to identify the capability groups. The capability prototypes themselves are computed from the original normalized embeddings z_i. Neither CFI metrics nor model-response information is used to determine the groups or select d.

Unified prototype construction. For both labeled and label-free benchmarks, the labels ℓ_i induce the capability groups $\mathcal{G}_{\mathcal{B}} = \{G_j\}_{j=1}^{d}$, as defined in Eq. 8. Following Eq. 9, we compute

$$\bar{z}_j = \frac{1}{|G_j|} \sum_{x_i \in G_j} z_i, \qquad b_j = \frac{\bar{z}_j}{\|\bar{z}_j\|_2}, \qquad j = 1, \ldots, d.$$

The resulting prototype matrix is $B = [b_1, \ldots, b_d] \in \mathbb{R}^{m \times d}$. Each b_j therefore provides a semantic anchor for capability group j, regardless of whether the group is defined by a native label or induced by clustering.

E.1.3 Capability-Frame Calibration via Gram Whitening

Because the semantic prototypes need not be mutually orthogonal, we apply the regularized Gram whitening transformation in Eq. 10. Specifically, we form $G_\epsilon = B^\top B + \epsilon I_d$, using $\epsilon = 10^{-6}$ unless otherwise specified. Its symmetric inverse square root is computed by eigendecomposition:

$$G_\epsilon = U \Lambda U^\top, \qquad \widetilde{B} = B U \Lambda^{-1/2} U^\top.$$

This produces an approximately orthonormal capability frame while preserving the subspace spanned by the original prototypes.

E.1.4 Projection and Operational Probing Direction

For each normalized embedding z_i, we compute its coordinates in the calibrated capability frame and normalize them:

$$v_i^{(\mathrm{raw})} = \widetilde{B}^\top z_i, \qquad v_i = \frac{v_i^{(\mathrm{raw})}}{\|v_i^{(\mathrm{raw})}\|_2}.$$

The resulting $v_i \in \mathbb{R}^d$ is the operational probing direction used to construct CFI and compute the Coverage, Uniformity, and Redundancy diagnostics. Information scaling is handled separately by the strength proxy α_i defined in Section 3.3.2.

E.2 Robustness to Text Encoder Choice

We examine whether CFI preserves its relative assessment of candidate subsets when the text encoder used to construct the semantic geometry is changed. We compare three encoders: `text-embedding-3-large`, `bge-m3`, and `e5-large-v2`. For each budget ratio $b \in \{0.05, 0.10, 0.20\}$, we construct 200 candidate subsets using random sampling, stratified random sampling, capability collapse, and redundancy injection, with 50 random seeds per procedure. The subset indices are generated once and reused across all encoders, so that each encoder evaluates exactly the same subsets. This yields 600 candidate subsets across the three budgets. Full subset-construction details are provided in Appendix G.

For each encoder, we reconstruct the embedding and similarity geometry and recompute the corresponding CFI diagnostics. Because different encoders induce distinct semantic representations, their raw diagnostic values need not be directly comparable. We therefore assess cross-encoder robustness through ranking consistency. At each budget, we compute pairwise Spearman correlations between the diagnostic rankings induced by the three encoders and use Kendall's W to measure their joint concordance.

Table 4 shows substantial agreement across the three encoders. Across all diagnostics and budgets, pairwise Spearman correlations range from 0.782 to 0.967, while Kendall's W ranges from 0.882 to

Table 4: Cross-encoder ranking consistency of CFI diagnostics on MMLU-Pro. For each budget ratio b, all encoders evaluate the same 200 candidate subsets. OAI, BGE, and E5 denote `text-embedding-3-large`, `bge-m3`, and `e5-large-v2`, respectively. Pairwise Spearman's ρ and Kendall's W are computed independently at each value of b.

b	**Diagnostic**	$\rho_{\mathrm{OAI,BGE}}$	$\rho_{\mathrm{OAI,E5}}$	$\rho_{\mathrm{BGE,E5}}$	**Kendall's W**
0.05	Coverage	0.823	0.859	0.827	0.891
	Uniformity	0.944	0.929	0.950	0.961
	Redundancy	0.967	0.948	0.946	0.969
	Ref-weighted Coverage	0.847	0.872	0.859	0.906
	Ref-weighted Uniformity	0.952	0.943	0.957	0.967
0.1	Coverage	0.812	0.870	0.821	0.890
	Uniformity	0.938	0.919	0.952	0.957
	Redundancy	0.949	0.905	0.904	0.946
	Ref-weighted Coverage	0.845	0.872	0.852	0.904
	Ref-weighted Uniformity	0.946	0.927	0.943	0.959
0.2	Coverage	0.782	0.882	0.806	0.882
	Uniformity	0.939	0.915	0.956	0.958
	Redundancy	0.937	0.896	0.879	0.936
	Ref-weighted Coverage	0.818	0.878	0.846	0.898
	Ref-weighted Uniformity	0.956	0.928	0.950	0.963

0.969. Uniformity is particularly stable: its pairwise correlations remain between 0.915 and 0.956, with Kendall's W between 0.957 and 0.961. Redundancy also shows strong agreement, although its concordance decreases somewhat as the subset budget increases, with W declining from 0.969 at $b = 0.05$ to 0.936 at $b = 0.20$.

Coverage is more sensitive to encoder choice than the other diagnostics. Its pairwise correlations range from 0.782 to 0.882, and its Kendall's W ranges from 0.882 to 0.891. Reference weighting modestly increases this agreement at every budget, raising the corresponding W values to 0.898–0.906. Reference-weighted Uniformity remains highly consistent as well, with W between 0.959 and 0.967.

These results indicate that encoder choice affects the numerical semantic geometry more than it affects the relative ordering of candidate subsets. The robustness claim here therefore concerns cross-encoder ranking consistency rather than invariance of the raw CFI diagnostic values.

E.3 Multimodal Benchmark Adaptation

We adapt the capability-direction construction procedure to benchmarks that contain both textual and visual information. The main difference from the text-only setting lies in the construction of the item embeddings. Once the two modalities have been combined into a shared representation, the remaining steps follow the same procedure described above.

E.3.1 Multimodal Feature Extraction and Fusion

We use a frozen `openai/clip-vit-large-patch14` model to encode the text and image associated with each item. Since the text and vision encoders of CLIP map their inputs into a shared representation space, their outputs can be combined directly. Let $e_i^{(\mathrm{text})}$ and $e_i^{(\mathrm{img})}$ denote the text and image embeddings for item x_i. We normalize each embedding separately:

$$z_i^{(\mathrm{text})} = \frac{e_i^{(\mathrm{text})}}{\|e_i^{(\mathrm{text})}\|_2}, \qquad z_i^{(\mathrm{img})} = \frac{e_i^{(\mathrm{img})}}{\|e_i^{(\mathrm{img})}\|_2}.$$

The two normalized embeddings are combined using

$$h_i = \omega_{\mathrm{img}} z_i^{(\mathrm{img})} + \omega_{\mathrm{text}} z_i^{(\mathrm{text})}, \qquad \omega_{\mathrm{img}} + \omega_{\mathrm{text}} = 1.$$

We then normalize the fused representation:

$$z_i = \frac{h_i}{\|h_i\|_2}.$$

We assign equal weight to the two modalities, setting $\omega_{\text{img}} = \omega_{\text{text}} = 0.5$. The symbol ω is reserved for modality-fusion weights and is distinct from the item-level information-strength proxy α_i used in CFI.

E.3.2 Capability-Frame Construction and Projection

Given the fused representations $\{z_i\}_{i=1}^n$, we construct the capability frame using the same procedure as in the text-only setting. Items are first assigned to capability groups using either native labels or the clustering procedure described above. Group prototypes are then constructed and calibrated through regularized Gram whitening. Finally, each fused item representation is projected into the calibrated frame and normalized. The resulting vector v_i is used as the operational probing direction for the multimodal item.

E.4 Summary

The implementation described in this appendix provides a common procedure for constructing operational probing directions from benchmark content. For text-only benchmarks, each item is represented by a normalized semantic embedding. For the multimodal benchmarks considered here, normalized text and image embeddings are combined in CLIP's shared representation space. The subsequent construction is identical in both cases.

Capability groups are defined using native benchmark labels when available. Otherwise, they are inferred by applying KMeans clustering to PCA-reduced embedding features. The normalized group centroids form the initial capability prototypes. Regularized Gram whitening calibrates these prototypes into an approximately orthonormal frame, and projection into this frame yields the operational probing direction v_i for each item. The direction v_i describes the item's position in the capability space, whereas its information strength is represented separately by α_i.

The robustness analysis also indicates that the ranking of candidate subsets is broadly stable across different text encoders. The resulting construction therefore offers a practical, benchmark-specific approximation to the latent information direction u_i and can be applied consistently to both textual and multimodal benchmarks.

F Implementation Details for Controlled Structural Perturbations

Notation and setup. Let $\mathcal{B} = \{x_i\}_{i=1}^N$ denote the item pool of MMLU-Pro (so $N = 12032$ in our experiments). Each item x_i is assigned a capability label $\ell_i \in \{1, \dots, d\}$ induced by the capability geometry (for MMLU-Pro, $d = 14$). These labels induce capability groups $\mathcal{G}_{\mathcal{B}} = \{G_j\}_{j=1}^d$, where $G_j = \{x_i \in \mathcal{B} : \ell_i = j\}$. Given a target subset size k, we construct perturbed subsets $\mathcal{S} \subseteq \{1, \dots, N\}$ to induce controlled structural deficiencies. Unless otherwise stated, randomness is controlled via deterministic seeds.

1) Directional item removal (blind-spot perturbation). Directional removal targets a drop label ℓ^* and progressively removes a fraction of items from that label. For each random seed, we first construct a base subset $\mathcal{S}_0$ by stratified sampling across capability labels. We then remove

$$n_{\text{drop}} = \text{round}\Big(\big|\{i \in \mathcal{S}_0 : \ell_i = \ell^*\}\big| \cdot r\Big)$$

items uniformly at random from $\{i \in \mathcal{S}_0 : \ell_i = \ell^*\}$. If the resulting set size is smaller than k, the remainder is filled by sampling uniformly from items whose labels are not ℓ^* until $|\mathcal{S}| = k$.

2) Capability-space collapse. Capability-space collapse reduces coverage across capability labels by restricting items to a small set of active labels. We sample an active label set $\mathcal{A} \subseteq \{1, \dots, d\}$ with

$|\mathcal{A}| = n_{\text{active}}$, and then sample k items uniformly from the union $\{i : \ell_i \in \mathcal{A}\}$. If the available pool is smaller than k, sampling is performed with replacement.

3) Redundancy injection (local overlap perturbation). Redundancy injection increases local duplication using a precomputed item similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$. Its entries are defined consistently with the redundancy metric in the main text: $\mathbf{S}_{ij} = \text{sim}(i, j) = \max\{0, z_i^\top z_j\}$. For redundancy level $\ell \in \{\text{low}, \text{medium}, \text{high}\}$, we construct a subset via seed selection followed by optional nearest-neighbor expansion under $\mathbf{S}$. For *low*, we select k items under a farthest-to-set criterion with no neighbor expansion. For *medium* and *high*, we select $\max(1, \lfloor 0.70k \rfloor)$ and $\max(1, \lfloor 0.50k \rfloor)$ items, respectively, and expand each item by adding one nearest neighbor under $\mathbf{S}$. If the constructed set contains fewer than k items, remaining items are filled uniformly at random. The redundancy level determines the initial item count and whether neighbor expansion is applied.

4) Difficulty restriction. Difficulty restriction samples items from a specified accuracy bin. We compute per-item accuracies $\text{acc}(i)$ using a designated reference model and define

- *Easy*: $\text{acc}(i) > 0.7$;
- *Medium*: $0.3 \leq \text{acc}(i) \leq 0.7$;
- *Hard*: $\text{acc}(i) < 0.3$.

Given a bin name and target subset size k, we sample uniformly from the corresponding pool; if the pool size is smaller than k, sampling is performed with replacement.

Hyperparameter values used in Fig. 3. We set the subset size to

$$k = \max\left(1, \lceil b \cdot N \rceil\right),$$

where budget ratio $b = 0.1$ and $N = 12032$ for MMLU-Pro. All perturbations are repeated for random seeds $s \in \{0, \ldots, 49\}$. For directional removal, we vary the drop ratio as $r \in \{0.0, 0.25, 0.50, 0.75, 1.0\}$. For capability collapse, we set n_{active} as:

- *balanced*: $n_{\text{active}} = d = 14$;
- *one direction*: $n_{\text{active}} = 1$;
- *two direction*: $n_{\text{active}} = 2$;
- *partial*: $n_{\text{active}} = \max(1, \lfloor d/3 \rfloor) = 4$.

For redundancy injection, we use the three levels in Fig. 3(c). For difficulty restriction, each panel corresponds to its respective difficulty bin. In uniform-strength and reference-weighted regimes, item information strengths follow the main text definition of α_i.

G Implementation Details for Fidelity Evaluation

Goal. We test whether structural differences captured by CFI are associated with evaluation fidelity on a disjoint model pool. To separate these structural differences from variation in subset size, we fix the item budget and construct a diverse collection of candidate subsets without optimizing either the CFI diagnostics or the fidelity outcomes. For each subset, we compute uniform-strength Coverage and Uniformity together with Redundancy, and examine their association with ranking and score fidelity.

Subset budget. Let $N = 12032$ denote the total number of items in the MMLU-Pro pool, i.e., $\mathcal{B} = \{x_i\}_{i=1}^N$. We fix the budget ratio at $b = 0.10$, which corresponds to a subset size of $k = 1204$. All candidate subsets therefore contain the same number of items.

Sampling procedures. For the fixed subset size $k = 1204$, we generate candidate subsets $\mathcal{S} \subseteq \{1, \ldots, N\}$ using four procedures:

1. **Random sampling.** Items are drawn uniformly at random from $\mathcal{B}$ without replacement.

2. **Category-balanced sampling.** Items are sampled in a stratified manner across MMLU-Pro category labels $\ell_i \in \{1, \ldots, d\}$ by cycling through labels and selecting one item per step (without replacement when possible), until reaching size k.
3. **Capability-collapse sampling.** We first sample an active set of capability labels $\mathcal{A} \subseteq \{1, \ldots, d\}$ and then sample all k items from the union $\{i : \ell_i \in \mathcal{A}\}$ (with replacement if the union is insufficient).
4. **Redundancy-injected sampling.** We use a precomputed similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ to favor locally similar items. Concretely, $\mathbf{S}$ is the item-to-item similarity used in redundancy injection, i.e., $\mathbf{S}_{ij} = \mathrm{sim}(i, j)$ as defined in the main text. We select seed items and iteratively add highly similar items under $\mathbf{S}$ until reaching size k; any remaining slots are filled uniformly at random.

Random seeds and number of subsets. For each sampling procedure, we repeat the construction with 50 deterministic random seeds, yielding $4 \times 50 = 200$ candidate subsets in total.

Computing CFI diagnostics. For each sampled subset $\mathcal{S}$, we compute uniform-strength Coverage and Uniformity, together with the weight-independent Redundancy diagnostic, using the same formulas as in Section 3.4. Diagnostics are computed from the subset-induced structural quantities in the benchmark-specific capability space.

Fidelity evaluation and correlation reporting. For each candidate subset, we evaluate ranking and score fidelity on the disjoint model pool described in Section 4.1, using τ_{cat}, τ_{self}, and MAE. For each diagnostic $D \in \{\mathrm{Cov}, \mathrm{Unif}, \mathrm{Red}\}$ and each fidelity measure $m \in \{\tau_{\mathrm{cat}}, \tau_{\mathrm{self}}, \mathrm{MAE}\}$, we compute the Spearman rank correlation $\rho(D, m)$ across all 200 fixed-budget candidate subsets. Fig. 4 reports these correlations as point estimates.

H Subset Selection Under Budget Constraints

Overview. In this section, we study how different subset construction protocols behave under a fixed budget ratio. Across four protocols (Random, Diversity, COV ONLY, and COV+RED), we analyze both (i) structural diagnostics in the benchmark-specific capability space (Coverage, Uniformity, Redundancy) and (ii) held-out evaluation fidelity under disjoint model pools.

Common evaluation setup. All protocols use the same item capability directions and weights, and all subsets are evaluated using the same held-out model pool (disjoint from any reference models used for reference-weighting).

H.1 Four Subset Construction Protocols

We compare four subset construction methods on three benchmarks (MMLU-Pro, MATH$_{TEST}$, GSM8K) across budget ratios $b \in \{0.01, 0.02, 0.05, 0.10, 0.20\}$ of the full item pool. For each budget ratio, we repeat subset construction with multiple random seeds to estimate variability.

Sampling procedures. For each budget ratio b and subset size $k = \max(1, \lceil b \cdot N \rceil)$, we generate candidate subsets $\mathcal{S} \subseteq \{1, \ldots, N\}$ of size k using four procedures. Here, N denotes the total number of items in the corresponding benchmark item pool.

1. **Random sampling.** Follows the same random sampling rule in Appendix G.
2. **Diversity (Stratified Random; Capability-balanced sampling).** Uses the same capability-balanced sampling rule as in Appendix G.
3. **COV ONLY.** Greedily maximizes weakest-direction coverage without any redundancy penalty:

$$S^* \in \arg\max_{|S|=k} \widehat{\mathrm{Cov}}(S; \alpha) \triangleq \arg\max_{|S|=k} \lambda_{\min}\left(\sum_{i \in S} \alpha_i \, v_i v_i^\top\right). \tag{46}$$

4. **COV+RED.** Greedily maximizes weakest-direction coverage with a local-overlap (redundancy) penalty:

$$S^* \in \arg\max_{|S|=k} \Big\{ \widehat{\mathrm{Cov}}(S;\alpha) - \beta \cdot \mathrm{Red}(S) \Big\}, \tag{47}$$

where $\beta > 0$ controls the strength of the local duplication penalty. The redundancy term $\mathrm{Red}(S)$ is the one defined in the main text (Eq. 15).

H.2 CFI-Greedy Construction

Algorithm overview. Both COV ONLY and COV+RED use the same greedy procedure to construct a subset of size k. The only difference is whether the redundancy penalty is applied during selection. Setting $\beta = 0$ yields COV ONLY, with $J^{(i)} = \widehat{\mathrm{Cov}}^{(i)}$. Setting $\beta > 0$ yields COV+RED, which trades off weakest-direction coverage against local redundancy.

Algorithm 1: CFI-Greedy for COV ONLY and COV+RED

Require: Item directions $\{v_i\}_{i=1}^N \subset \mathbb{R}^d$, weights $\{\alpha_i\}_{i=1}^N$, subset size k, candidate pool size p, redundancy coefficient β
Ensure: Selected subset S_k

$S_0 \leftarrow \emptyset$, $\mathrm{CFI}_0 \leftarrow \mathbf{0}_{d\times d}$
$C_{\max} \leftarrow \lambda_{\min}\big(\sum_{i\in\mathcal{B}} \alpha_i v_i v_i^\top\big)$
for $t = 0$ **to** $k-1$ **do**
  Sample $\mathcal{P}_t \subseteq \mathcal{B} \setminus S_t$ with $|\mathcal{P}_t| = \min\{p, |\mathcal{B} \setminus S_t|\}$
  for each candidate $i \in \mathcal{P}_t$ **do**
    $\mathrm{CFI}^{(i)} \leftarrow \mathrm{CFI}_t + \alpha_i v_i v_i^\top$
    $\widehat{\mathrm{Cov}}^{(i)} \leftarrow \lambda_{\min}(\mathrm{CFI}^{(i)})/C_{\max}$
    $R^{(i)} \leftarrow \mathrm{Red}(S_t \cup \{i\})$
    $J^{(i)} \leftarrow \widehat{\mathrm{Cov}}^{(i)} - \beta R^{(i)}$
  end for
  $i^\star \leftarrow \arg\max_{i\in\mathcal{P}_t} J^{(i)}$
  $S_{t+1} \leftarrow S_t \cup \{i^\star\}$
  $\mathrm{CFI}_{t+1} \leftarrow \mathrm{CFI}_t + \alpha_{i^\star} v_{i^\star} v_{i^\star}^\top$
end for
return S_k

Implementation hyperparameters. To keep the greedy search scalable, each step evaluates a randomly sampled candidate pool containing at most $p = 500$ remaining items. For COV+RED, we set $\beta = 0.2$ without tuning it against held-out model rankings.

Computational complexity. At each greedy step, we evaluate at most p candidates. Computing the coverage score via eigendecomposition of a $d \times d$ candidate CFI matrix costs $O(d^3)$ per candidate. With precomputed pairwise overlap scores r_{ij}, incremental redundancy evaluation costs at most $O(k)$ per candidate. The overall selection complexity is therefore

$$O\big(kp(d^3 + k)\big),$$

which reduces to $O(kpd^3)$ when eigendecomposition dominates.

H.3 Structural and Reliability Trends Across Budget Levels

Coverage trends. Coverage is the smallest eigenvalue of the subset-induced CFI operator (Eq. 13). Across budgets:

- **Random:** Coverage increases nearly proportionally with budget.
- **Diversity:** Slightly improves Coverage over Random, reflecting coarse category balance; however, it does not target weak capability directions.

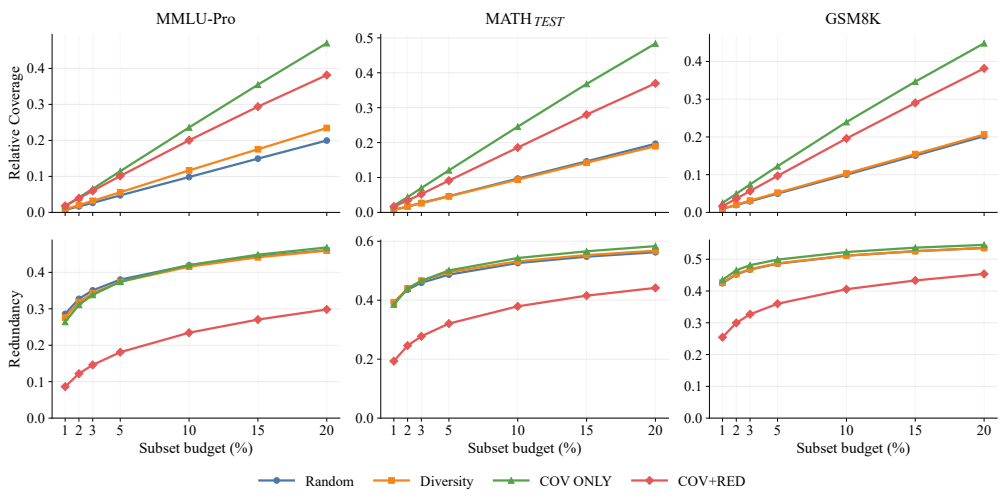


Figure 6: **Structural metrics under budget constraints.**

- **COV ONLY:** Coverage grows steeply at small budgets and saturates later. At moderate budget ratios, COV ONLY reaches substantially higher relative coverage than budget-proportional sampling.
- **COV+RED:** Tracks COV ONLY closely with only mild degradation, indicating that redundancy control mainly affects local duplication rather than weakest-direction support.

Redundancy trends. Redundancy measures local pairwise overlap using the redundancy term in Eq. 15.

- **Random & Diversity:** Redundancy changes only mildly and increases gradually with budget.
- **COV ONLY:** Redundancy is not effectively reduced by the coverage-only greedy rule; it stays comparable to (and may be slightly higher than) Random/Diversity.
- **COV+RED:** Redundancy is systematically lower than COV ONLY, demonstrating the effectiveness of the redundancy penalty in improving budget efficiency.

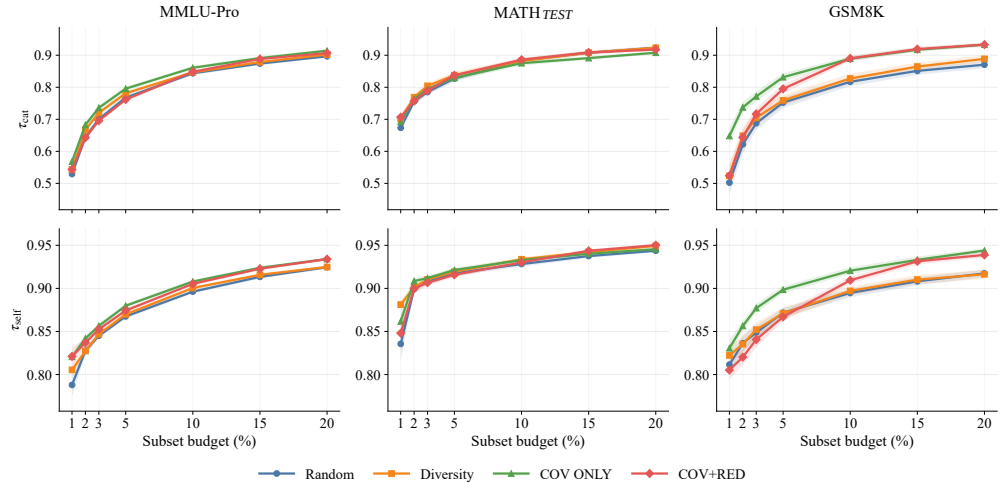


Figure 7: **Held-out reliability metrics under budget constraints.**

Held-out ranking fidelity (τ_{cat}). Category-wise ranking fidelity evaluates how well a selected subset preserves model rankings within benchmark-defined groups.

- **Random & Diversity:** achieve moderate τ_{cat} at budget ratio $b = 0.10$.
- **COV ONLY:** reaches substantially higher τ_{cat} at the same budgets.
- **COV+RED:** remains competitive, with a small drop relative to COV ONLY attributable to the redundancy constraint.

Bootstrap stability (τ_{self}). Bootstrap stability reflects how consistently model rankings emerge under resampling of the selected subset.

- **Random & Diversity:** yield the lowest stability.
- **COV ONLY & COV+RED:** yield higher stability, indicating that CFI-guided subsets concentrate measurement information on more discriminative directions.

Key takeaway. Overall, the results show a consistent correspondence between structural improvement and held-out reliability:

1. **Coverage is the primary driver:** achieving higher weakest-direction support yields the largest fidelity gains.
2. **Redundancy improves efficiency:** redundancy-aware selection reduces local duplication with limited impact on fidelity.
3. **Stability improves under structured selection:** CFI-guided subsets provide more robust model comparisons under resampling.

H.4 Relation to Efficient Benchmarking and Subset Selection Literature

Prior work on efficient benchmarking has pursued several related but distinct objectives. One prominent line treats subset selection as a *preservation* problem:

> **Given a full benchmark $\mathcal{B}$, it selects a small subset whose scores or model rankings approximate those obtained from $\mathcal{B}$.**

For example, TINYBENCHMARKS uses item response theory and representative anchor items to estimate full-benchmark performance from substantially fewer examples (Polo et al., 2024). ESSENCEBENCH similarly optimizes compact subsets for reconstructing full-benchmark scores while explicitly filtering textual and behavioral redundancy (Wang et al., 2026). These methods primarily ask how evaluation cost can be reduced while retaining the conclusions of the original benchmark.

Other work moves beyond static compression. AUTOJUDGER decomposes model ability, updates an estimated capability profile, and adaptively selects informative questions for the model being evaluated (Ding et al., 2026). BENCHMARK^2 instead evaluates the quality of benchmarks themselves through cross-benchmark ranking consistency, discriminability, and capability alignment, and uses these criteria to filter problematic items and construct reduced benchmarks (Qian et al., 2026). These approaches share our broader motivation that the full benchmark should not automatically be treated as an unquestioned gold standard.

Our CFI-guided construction is most closely related to this latter diagnosis-and-refinement perspective, but differs in its structural objective. We characterize a benchmark through how its measurement information is distributed over the intended capability subspace $\mathcal{F}_{\mathcal{B}}$. Rather than directly minimizing error relative to full-benchmark scores or rankings, we optimize

$$\lambda_{\min}\left(\sum_{i\in S}\alpha_i v_i v_i^{\top}\right),$$

which targets the least-informed capability direction. This criterion corresponds to the E-optimality principle in optimal experimental design (Pukelsheim, 2006). The redundancy-aware variant additionally penalizes $\mathrm{Red}(S)$ to discourage locally duplicative capability probes. Consequently, the selected subset may intentionally alter the measurement structure of the original benchmark when that structure is imbalanced, rather than merely reproduce its aggregate outcomes.

CFI-guided selection is also static rather than model-adaptive: the item weights and capability directions are estimated from a reference model pool, after which a common subset is constructed for evaluating held-out models. It therefore complements adaptive methods such as AUTOJUDGER and benchmark-quality frameworks such as BENCHMARK^2, while providing an explicit geometric criterion for weakest-direction coverage.

I FINE-GRAINED PER-BENCHMARK DIAGNOSTICS

The main text reports aggregate structural and reliability metrics. In this appendix, we further analyze reference-weighted CFI along benchmark-specific diagnostic categories, indexed by c. When native annotations are available, these categories follow the benchmark metadata: MMLU-Pro uses subject categories, and MATH$_{TEST}$ uses mathematical topics. For GSM8K, which does not provide native fine-grained labels, the latent clusters induced during capability-basis construction serve as the diagnostic categories.

These partitions are benchmark-internal and are not intended to define a shared taxonomy across benchmarks. Rather, they provide semantic anchors for examining how measurement information is distributed within the internal structure of each benchmark.

We reuse the capability-space construction from Appendix E. Let b_c denote the raw semantic anchor associated with category c, constructed in the same manner as the capability prototypes used to form the raw prototype matrix B. Let $\widetilde{B}$ denote the calibrated basis obtained after regularized Gram whitening. We define the corresponding diagnostic direction as the normalized coordinate representation of this anchor in the calibrated capability space:

$$v_c = \frac{\widetilde{B}^\top b_c}{\|\widetilde{B}^\top b_c\|_2}. \tag{48}$$

This construction preserves the semantic reference carried by b_c while expressing it in the same calibrated coordinate system as the operational item directions v_i. Accordingly, v_c should be interpreted as the calibrated representation of the semantic anchor, rather than as the original embedding-space prototype or as a single axis of the whitened basis.

Given a subset S, we define its directional information mass along category c as

$$q_c(S;\alpha) = v_c^\top \mathrm{CFI}(S;\alpha) v_c = \sum_{i\in S} \alpha_i \left(v_i^\top v_c\right)^2. \tag{49}$$

To measure how much of the full-benchmark information is retained along this direction, we report the normalized recovery ratio

$$\widetilde{q}_c(S;\alpha) = \frac{q_c(S;\alpha)}{q_c(F;\alpha)}, \tag{50}$$

where F denotes the full benchmark item set. Thus, $\widetilde{q}_c(S;\alpha)$ measures the fraction of the directional information available in the full benchmark that is preserved by S.

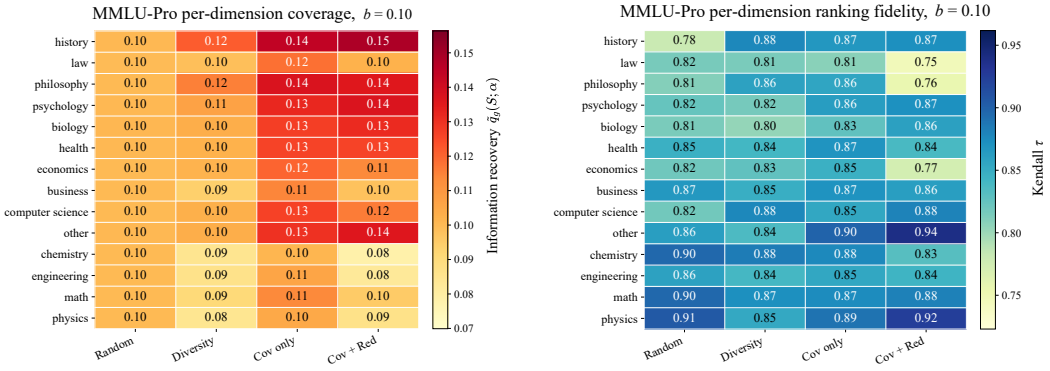
Because Gram whitening can mix the original semantic prototypes, $\widetilde{q}_c(S;\alpha)$ should not be interpreted as recovery of an isolated, axis-aligned ability dimension. Instead, it quantifies retention along the calibrated diagnostic direction associated with the corresponding semantic anchor, consistent with the benchmark-internal interpretation of CFI.

I.1 MMLU-PRO SUBJECT-GROUP DIAGNOSTICS

Setup. MMLU-Pro provides native subject categories, which we use as diagnostic groups. For each subject group g, we decompose the reference-weighted CFI along the corresponding subject direction and evaluate both information recovery $\widetilde{q}_g(S;\alpha)$ and within-subject ranking fidelity under a budget ratio $b = 0.1$.

Full-benchmark subject profile. The main text presents MMLU-Pro as a representative benchmark-level case study and reports its full-benchmark subject profile. Fig. 2(a) compares the number of items in each category with its full-benchmark information mass $q_g(F;\alpha)$. Although

categories with more items often provide more information, the two quantities are not identical. This indicates that category size alone is not a sufficient proxy for effective discriminative support: a large category may still contribute only moderate reference-weighted CFI mass, while a smaller category may provide relatively strong directional support.



(a) Information recovery (b) Ranking fidelity

Figure 8: Fine-grained MMLU-Pro diagnostics at the budget ratio $b = 0.10$. Panel (a) reports per-category information recovery $\widetilde{q}_g(S;\alpha)$, the fraction of full-benchmark directional information mass retained by the selected subset. Panel (b) reports per-category ranking fidelity, measured by the Kendall correlation between the selected-subset model ranking and the full-benchmark model ranking within the same category.

Subject-wise information recovery and ranking fidelity. Fig. 8(a) reports normalized directional information recovery $\widetilde{q}_g(S;\alpha)$ under a budget ratio $b = 0.1$. Random selection recovers roughly the budget-proportional amount of information in most categories, while diversity-based selection remains close to this baseline. In contrast, the CFI-guided methods recover substantially more directional information across MMLU-Pro categories. This shows that the aggregate coverage gains reported in the main text are not driven by a single category, but are distributed across the benchmark's native subject dimensions. Among the two CFI variants, COV ONLY typically achieves the strongest information recovery, while COV+RED often retains most of this gain, with the margin varying across categories; this aligns with its redundancy control objective.

Fig. 8(b) reports category-wise ranking fidelity, measured by the Kendall correlation between the selected-subset model ranking and the full-benchmark model ranking within the same category. Compared with directional information recovery, the ranking-fidelity pattern is more heterogeneous. This suggests that preserving category-wise rankings depends not only on recovered CFI mass, but also on category-specific factors such as model separability, category size, and within-category difficulty structure. We therefore interpret the ranking-fidelity heatmap as a complementary diagnostic rather than a direct monotonic consequence of directional information recovery.

Overall, the MMLU-Pro diagnostics show that CFI-guided selection improves information recovery across essentially all native subject categories in MMLU-Pro, rather than concentrating gains in a single domain. At the same time, category-wise ranking fidelity remains more heterogeneous, indicating that ranking preservation depends on both structural information recovery and category-specific evaluation factors.

I.2 MATH$_{TEST}$ TOPIC-GROUP DIAGNOSTICS

Setup. MATH$_{TEST}$ provides topic annotations, including Algebra, Geometry, Number Theory, Prealgebra, Precalculus, Counting & Probability, and Intermediate Algebra. We use these topics as diagnostic groups and decompose the reference-weighted CFI along the corresponding topic directions.

Full-benchmark topic profile. Fig. 9 compares topic-level item counts with full-benchmark information mass $q_g(F;\alpha)$. The topic sizes vary substantially, with Algebra and Intermediate

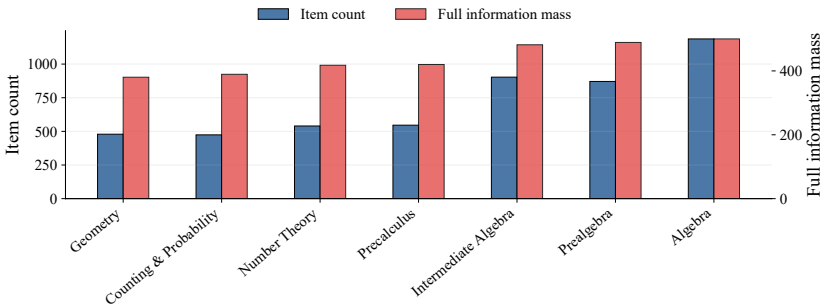
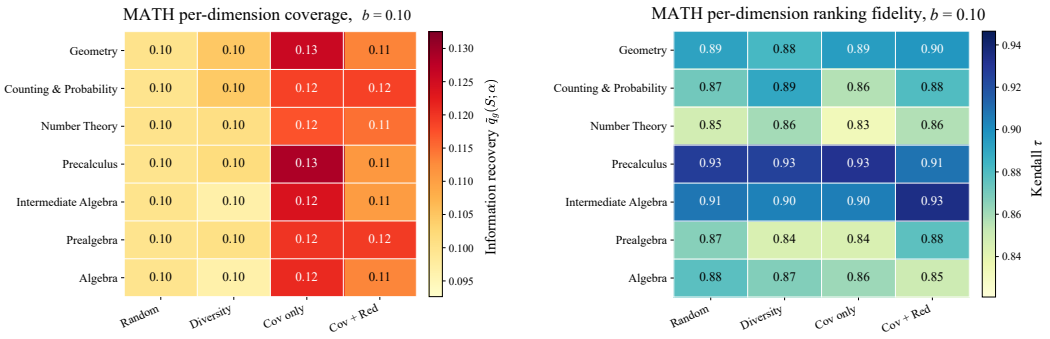


Figure 9: MATH$_{TEST}$ topic size and full-benchmark information mass. Full-benchmark information mass $q_g(F;\alpha)$ measures the reference-weighted CFI support along each topic direction. Topic size and effective information support are related but not identical, showing that item count alone is not a sufficient proxy for discriminative coverage.

Algebra containing many more items than several other topics. However, full-benchmark information mass varies less sharply than item count. For example, some smaller topics still contribute substantial reference-weighted CFI mass. This indicates that topic size alone does not determine effective information support; item difficulty, discriminativeness, and capability-direction geometry also affect how much useful information a topic contributes.



(a) Information recovery (b) Ranking fidelity

Figure 10: Fine-grained MATH$_{TEST}$ topic diagnostics at the budget ratio $b = 0.10$. Panel (a) reports per-topic information recovery $\widetilde{q}_g(S;\alpha)$, the fraction of full-benchmark directional information mass retained by the selected subset. Panel (b) reports per-topic ranking fidelity, measured by the Kendall correlation between the selected-subset model ranking and the full-benchmark model ranking within the same topic.

Topic-wise information recovery and ranking fidelity. Fig. 10(a) reports normalized directional information recovery $\widetilde{q}_g(S;\alpha)$ under a budget ratio $b = 0.1$. Random selection recovers roughly the budget-proportional amount of information, and diversity-based selection remains close to this baseline. In contrast, the CFI-guided methods recover substantially more information across all seven MATH$_{TEST}$ topics. In particular, COV ONLY consistently recovers about 20%–30% more directional information than the budget-proportional baseline across all topics, while COV+RED preserves a substantial fraction of this gain by explicitly penalizing local redundancy.

Fig. 10(b) reports topic-wise ranking fidelity. Unlike information recovery, ranking fidelity is high for most methods and topics, indicating a ceiling effect at the budget ratio $b = 0.10$. This suggests that MATH$_{TEST}$ topic-wise model rankings are already relatively stable under small subsets. Consequently, topic-wise ranking fidelity is less discriminative among selection methods in this setting, while directional information recovery more clearly exposes the structural advantage of CFI-guided selection.

Overall, the MATH$_{TEST}$ diagnostics show a particularly clear separation between structural information recovery and ranking fidelity. CFI-guided selection substantially increases topic-wise information

recovery under a fixed budget, whereas topic-wise ranking fidelity is already near saturation for most methods. This supports the view that CFI captures structural information allocation beyond what is visible from ranking-fidelity metrics alone.

I.3 GSM8K Latent-Group Diagnostics

Setup. Unlike MMLU-Pro and MATH$_{TEST}$, GSM8K does not provide native subject or topic labels that can be directly used for fine-grained category diagnostics. As described in Appendix E, we therefore construct a benchmark-internal latent capability frame from item embeddings. In this subsection, the latent groups are used only as diagnostic directions rather than as ground-truth semantic skill labels. We do not assign semantic names to the clusters; instead, we use them to test whether information-aware subset selection can preserve effective benchmark information in a setting without human-defined categories.

This makes GSM8K a label-free stress test for CFI. The key question is not what each individual cluster means semantically, but whether a selected subset preserves information across the latent directions induced by the benchmark itself. We therefore examine three aspects: the full-benchmark latent-group profile, latent-group information recovery under different subset selection methods, and robustness to the number of latent groups.

Latent-group selection. Fig. 11 reports the silhouette scores for candidate latent dimensions $d \in \{3, 4, 5, 6, 7, 8\}$. The highest score is obtained at $d = 3$, which we use as the default GSM8K latent frame. The scores for $d = 4$ and $d = 5$ remain close, while clustering quality drops more noticeably for $d \geq 6$. Importantly, this choice is made using the clustering diagnostic alone, before evaluating subset-selection performance. We further test fixed choices of d below to ensure that the conclusions are not tied to this single clustering resolution.

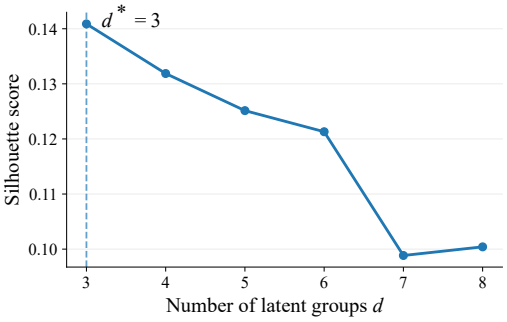


Figure 11: Silhouette scores for GSM8K latent-group construction. The default latent dimension is selected as $d^\star = 3$, which achieves the highest clustering score among the tested values.

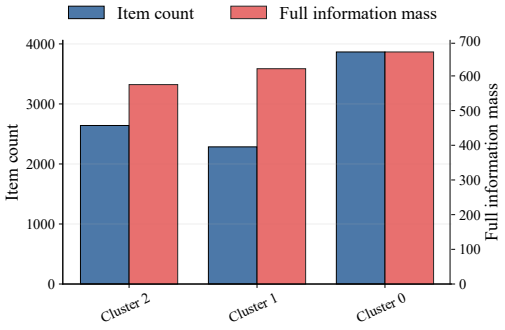


Figure 12: GSM8K latent-group profile under the default $d^\star = 3$ capability frame. The blue bars show item counts, and the red bars show full-benchmark information mass $q_g(F; \alpha)$.

Fig. 12 compares the item count and full-benchmark information mass $q_g(F; \alpha)$ of the three latent groups. The purpose of this profile is not to interpret each cluster as a named reasoning skill, but to examine whether item frequency and effective information support coincide in GSM8K.

The result shows that they are related but not identical. While some clusters with more items also tend to contribute larger information mass, the ordering across clusters is not fully determined by item count. In particular, at least one cluster with fewer items still contributes disproportionately large full-benchmark information mass. Thus, even in GSM8K, the distribution of effective reference-weighted information is not fully determined by the number of problems in each latent group.

This motivates the use of information recovery rather than count preservation as the fine-grained diagnostic. A subset may preserve the rough latent-group proportions but still fail to preserve the information contributed by each latent direction.

Latent-group information recovery. Fig. 13 reports the normalized directional information recovery $\widetilde{q}_g(S; \alpha)$ under a budget ratio $b = 0.1$. Since $\widetilde{q}_g(S; \alpha)$ measures the fraction of full-benchmark

information recovered along latent direction g, a purely budget-proportional subset would recover roughly 0.10 of the information in each group.

This is precisely what we observe for Random selection: its recovery is approximately 0.10 across all three latent groups. Diversity-based selection remains very close to Random, with only slight deviations in one group.

In contrast, the CFI-guided methods recover far more information than their item budget would suggest. COV ONLY achieves about 0.15–0.16 recovery across the three latent groups, while COV+RED achieves about 0.14–0.15. Thus, with only 10% of the items, the selected subsets retain a substantially larger fraction of the full-benchmark directional information along each latent diagnostic direction.

The main conclusion is therefore not that one unnamed cluster is more important than another. Rather, the result shows that CFI can identify information-rich subsets even when the benchmark does not provide native category labels. The gain is also distributed across all latent directions instead of being concentrated in a single cluster. The small decrease from COV ONLY to COV+RED is expected: the redundancy-aware objective trades a modest amount of directional information recovery for lower local overlap.

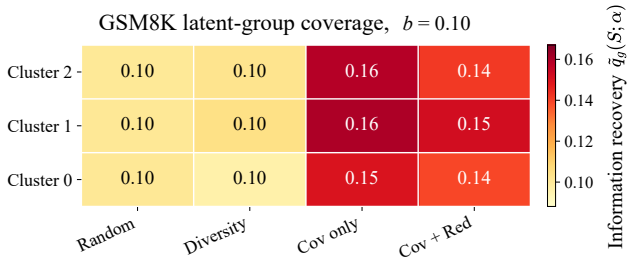


Figure 13: Latent-group information recovery on GSM8K at the budget ratio $b = 0.10$. Each entry reports $\widetilde{q}_g(S;\alpha)$, the fraction of full-benchmark directional information recovered along a latent-group direction.

Robustness to the number of latent groups. Because GSM8K latent groups are induced by unsupervised clustering rather than native annotations, it is important to verify that the conclusions do not depend on the particular choice $d^\star = 3$. We therefore repeat the construction and evaluation under fixed latent dimensions $d \in \{3, 4, 5, 6, 7, 8\}$, using the same 10% budget and candidate pool size $B = 500$.

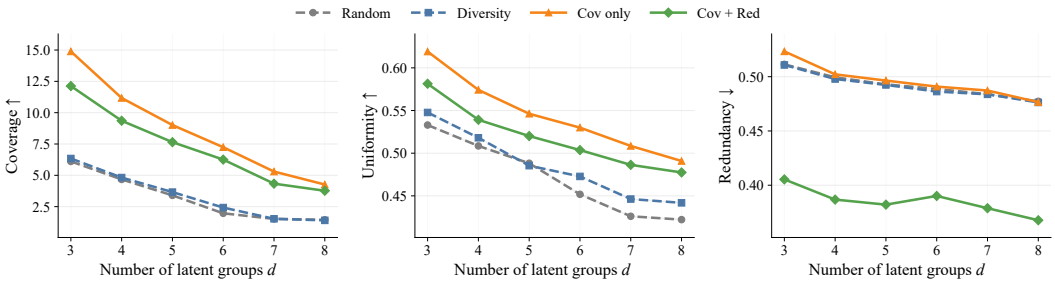


Figure 14: Robustness of GSM8K structural metrics to the number of latent groups d. All methods are evaluated at the budget ratio $b = 0.10$ with candidate pool size $p = 500$.

Fig. 14 shows the structural metrics under different fixed values of d. The absolute value of Coverage decreases as d increases, which is expected: weakest-direction coverage becomes harder to maintain in a higher-dimensional capability frame. Therefore, the important comparison is not the raw Coverage value across different d, but the relative ordering of methods under the same fixed frame.

Under this comparison, the CFI-guided methods are consistently stronger than Random and Diversity across all tested values of d. Both COV ONLY and COV+RED achieve much higher Coverage, and they also maintain stronger Uniformity in most settings. The redundancy plot further separates the

two CFI variants: COV+RED consistently yields the lowest Redundancy across all values of d. This confirms that the redundancy penalty is not only effective under the default latent frame, but remains effective under finer GSM8K latent decompositions.

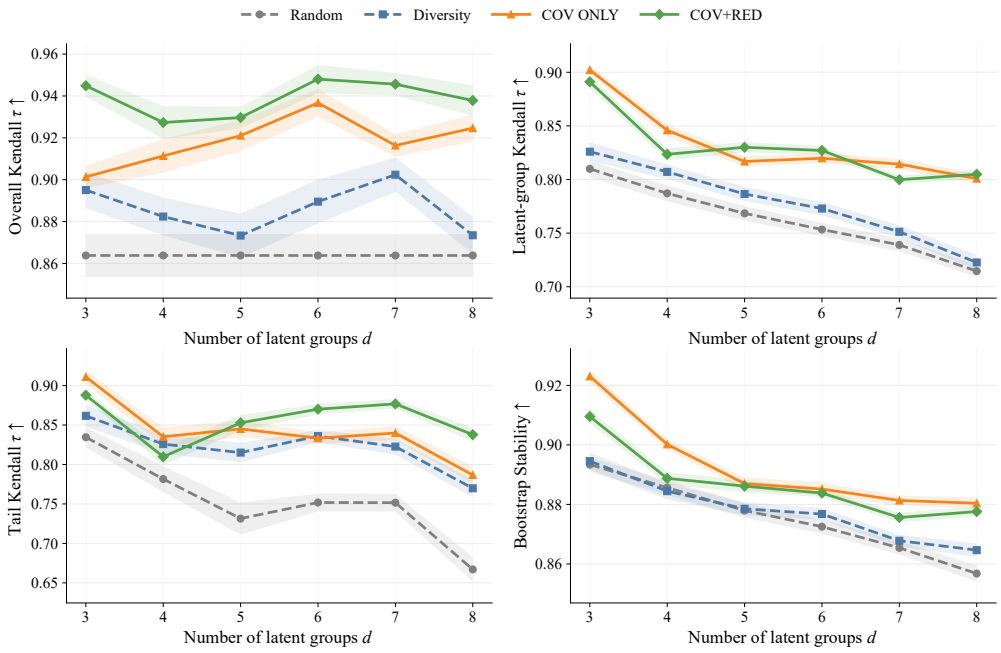


Figure 15: Robustness of GSM8K held-out reliability metrics to the number of latent groups d. CFI-guided subsets remain more reliable than Random and Diversity across the tested latent capability frames. Overall and tail Kendall correlations are auxiliary diagnostics; group-level Kendall correlation and bootstrap stability are the most diagnostic metrics for this analysis. Bands denote standard errors over 50 random seeds.

Fig. 15 reports held-out reliability metrics under the same fixed-d settings. The overall Kendall correlation is close to saturation for the CFI-guided methods, which nearly preserve the full-benchmark global model ranking under a budget ratio $b = 0.1$. Random and Diversity remain lower, around 0.87–0.90 depending on d.

The latent-group Kendall correlation is more diagnostic. As d increases, the latent groups become finer and contain fewer items, making group-level ranking estimates noisier. Consequently, all methods show some degradation with larger d. However, COV ONLY and COV+RED remain substantially higher than Random and Diversity throughout the tested range. This indicates that CFI-guided subsets preserve group-level model comparisons more reliably, even when the latent diagnostic frame is made more fine-grained.

Tail ranking fidelity is noisier, as expected, because it focuses on a smaller portion of the held-out model pool. We therefore treat it as an auxiliary diagnostic rather than a primary conclusion. Nevertheless, the broad pattern remains consistent: CFI-guided selection is generally more reliable than Random selection, especially for smaller and moderate values of d. Bootstrap stability shows the clearest robustness pattern: COV ONLY and COV+RED remain consistently above the baselines across all values of d, indicating that their model-comparison conclusions are more stable under resampling.

The GSM8K diagnostics show that CFI-guided selection remains useful even without native benchmark categories. The latent groups should not be interpreted as human-defined reasoning skills, but they provide a benchmark-internal diagnostic frame for testing information preservation. Under this frame, Random and Diversity recover only budget-proportional information, while CFI-guided methods achieve about 0.14–0.16 of the full-benchmark directional information along each latent diagnostic direction with only 10% of the items. The fixed-d analysis further shows that the structural and reliability advantages of CFI-guided selection are stable across multiple latent clustering

resolutions. Thus, GSM8K supports the broader claim that CFI does not require a native taxonomy to identify compact, information-preserving benchmark subsets.

Table 5: Open-Source Text Models Used in Experiments

Model	Developer	Params (B)	HF ID
Reference Models			
gemma-3-4b-it	Google	4	google/gemma-3-4b-it
Ministral-3-3B-Instruct	Mistral AI	3	mistralai/Ministral-3-3B-Instruct
Mistral-7B-Instruct-v0.3	Mistral AI	7	mistralai/Mistral-7B-Instruct-v0.3
Qwen3-4B	Qwen Team	4	Qwen/Qwen3-4B
GLM-4-9B	Z.ai	9	zai-org/glm-4-9b-0412
Llama-3.2-1B-Instruct	Meta AI	1	meta-llama/Llama-3.2-1B-Instruct
Qwen2.5-14B-Instruct	Qwen Team	14	Qwen/Qwen2.5-14B-Instruct
Ministral-3-14B-Instruct	Mistral AI	14	mistralai/Ministral-3-14B-Instruct
Qwen2.5-32B-Instruct	Qwen Team	32	Qwen/Qwen2.5-32B-Instruct
Seed-OSS-36B-Instruct	ByteDance	36	ByteDance-Seed/Seed-OSS-36B-Instruct
Held-out Models			
Ministral-3-8B-Instruct	Mistral AI	8	mistralai/Ministral-3-8B-Instruct
Qwen3-8B	Qwen Team	8	Qwen/Qwen3-8B
Phi-4-mini-instruct	Microsoft	3.8	microsoft/Phi-4-mini-instruct
Llama-3.2-3B-Instruct	Meta AI	3	meta-llama/Llama-3.2-3B-Instruct
Llama-3.1-8B-Instruct	Meta AI	8	meta-llama/Llama-3.1-8B-Instruct
gemma-3-12b-it	Google	12	google/gemma-3-12b-it
gemma-3-27b-it	Google	27	google/gemma-3-27b-it
Phi-4	Microsoft	14	microsoft/phi-4
GLM-4-32B	Z.ai	32	zai-org/glm-4-32b-0414
Llama-3.3-70B-Instruct	Meta AI	70	meta-llama/Llama-3.3-70B-Instruct

Note: For text-only benchmarks, seed and held-out cohorts are as listed above. For multimodal benchmarks, we use the 17 AutoJudger (Ding et al., 2026) test set models as the seed cohort to compute the reference-weighted CFI.