Certifiable Evaluation: A Low-Rank Framework for Foundation Model Benchmarking with Formal Performance Guarantees

Anonymous Authors[1]

Abstract

Benchmark scores for foundation models are point estimates that carry no formal guarantees about performance on unseen tasks. We address this gap by developing a rigorous theory of *certifiable evaluation* grounded in a Latent Capability Model (LCM): the $n \times m$ matrix of model–task performance scores is posited to have rank $r \ll \min(n, m)$, with each score decomposing as an inner product of a model capability vector and a task requirement vector, perturbed by sub-Gaussian noise. Within this framework we prove (i) an information-theoretic lower bound showing that any benchmark with fewer than r tasks cannot produce non-trivial performance certificates for unseen tasks; (ii) an oracle PAC certificate that translates benchmark scores into prediction intervals for unseen tasks, with width scaling as $O(\sigma\sqrt{k}/\sigma_{\min}(V_B))$; (iii) a D-optimal benchmark selection criterion—the Minimum Sufficient Evaluation Set (MSES)—that maximises certificate quality for a given budget, with a greedy algorithm achieving a $(1 - 1/e)$-approximation. We validate all theoretical claims on a cross-benchmark performance matrix comprising 157 publicly available language models evaluated across a unified 86-task cross-benchmark matrix sampled from six benchmark suites whose full task counts total 376. Empirically, the effective rank is $r_{\text{eff}} = 8$ despite 86 observed tasks, implying that the full benchmark suites carry substantial per-suite redundancy (82–95%, consistent with Table 1). MSES achieves nominal certificate coverage to within 0.8 percentage points at $1 - \delta = 0.95$, while halving certificate width relative to random task selection.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Rigorous evaluation of foundation models is among the most consequential open problems in machine learning. As models are deployed in high-stakes settings including clinical decision support, legal research, and automated programming—the community needs more than rank-ordered leaderboard scores; it needs *formal performance guarantees*: statements of the form "with probability $\geq 1 - \delta$, this model's accuracy on task distribution T lies within ε of its benchmark score."

Current practice falls short in two complementary ways. First, benchmark scores are point estimates with no accompanying prediction intervals, making it impossible to quantify the risk of deploying a model on tasks not in the benchmark. Second, benchmark suites grow without principled termination criteria: practitioners add tasks until coverage feels sufficient, with no theory of when additional tasks yield certification gains.

This paper introduces a mathematical framework that resolves both issues. Our contributions are:

1. **Latent Capability Model (LCM).** We formalise the widespread observation that model–task performance matrices are approximately low-rank (Perlitz et al., 2023; Recht et al., 2019) by positing $P = UV^\top + E$ with $U \in \mathbb{R}^{n\times r}$, $V \in \mathbb{R}^{m\times r}$, and E sub-Gaussian noise. The rank r is the number of independent capability dimensions needed to explain all benchmark variation.
2. **Fundamental limit (Theorem 4.1).** No benchmark B with $|B| < r$ can certify performance on any unseen task, regardless of the number of models observed.
3. **Oracle PAC certificate (Theorem 4.2).** Given V and $|B| \geq r$, the least-squares plug-in predictor yields a closed-form certificate whose width decays as $O(\sqrt{k}/\sigma_{\min}(V_B))$.
4. **Practical certificate with estimated V (Theorem 4.4).** We extend the oracle result to the realistic setting where V is estimated from historical data, providing a perturbation bound that separates noise and estimation-bias contributions.
5. **Minimum Sufficient Evaluation Set (MSES, Theo-**

rem 4.7). Optimal benchmark selection is equivalent to D-optimal experimental design. A greedy algorithm achieves a $(1 - 1/e)$-approximation via submodularity of $\log\det$.

6. **Empirical validation.** On a 157×86 cross-benchmark matrix (sampled from 376 total tasks across six suites), we show $r_{\text{eff}} = 8$, demonstrate calibrated 95% certificates, and quantify per-benchmark redundancy.

Scope and relation to workshop themes. This paper directly addresses the workshop's call for moving "from scores to formal, quantitative guarantees on performance across levels" and for "benchmark design with provable properties." The LCM is a natural theoretical framework for understanding when and why benchmark performance predicts out-of-distribution task performance (Bommasani et al., 2021).

2. Related Work

Empirical benchmark analysis. Recht et al. (2019) showed that ImageNet performance on independently collected test sets correlates nearly linearly with the original, foreshadowing the low-rank phenomenon we formalise. Perlitz et al. (2023) observed that a small number of benchmark tasks explains most variance in LLM leaderboard scores, but without theoretical foundations or certification guarantees. Liang et al. (2022) introduced HELM as a holistic evaluation framework, yet coverage remains heuristic. Our work provides the missing theory.

Uncertainty quantification and conformal methods. Conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021) constructs marginal coverage intervals for a single, fixed task distribution. Our certificates differ fundamentally: they certify performance on *unseen* tasks by leveraging cross-task latent structure, making them applicable precisely where conformal prediction—which requires held-out data from the target task—is unavailable. PAC-Bayes (McAllester, 1999) bounds single-task generalisation; we operate in the multi-task regime.

Matrix completion. The LCM is structurally related to collaborative filtering via matrix completion (Candès & Recht, 2009; Recht, 2011; Koltchinskii et al., 2011), but our goal is prediction-with-certificates rather than imputation. We use matrix completion machinery to estimate V from historical data and propagate estimation error into the final certificate (Theorem 4.4).

Optimal experimental design. D-optimal design (Kiefer & Wolfowitz, 1960; Pukelsheim, 2006) selects measurement points to maximise $\log\det$ of the Fisher information matrix. Our benchmark selection problem is structurally identical to D-optimal design for the linear regression $p^B_{\text{new}} = V_B u_{\text{new}} + \epsilon$, which we exploit to import both the optimality criterion and the greedy approximation guarantee (Nemhauser et al., 1978).

Item Response Theory (IRT). IRT (Lord, 1980; Birnbaum, 1968) models test scores as a function of latent ability. The one-parameter Rasch model is a special case of our LCM with $r = 1$ and $v_j = [d_j]$ (difficulty only). Our framework generalises IRT to $r > 1$ dimensions and adds formal certification guarantees absent from classical psychometrics.

Scaling laws and emergence. Kaplan et al. (2020) characterise how performance scales with compute; Wei et al. (2022) study emergent abilities. Our framework provides a complementary lens: the effective rank r_{eff} characterises the *dimensionality* of model capabilities, which we show grows sub-linearly with cumulative model count (§6.5).

3. The Latent Capability Model

Notation. Let $[n] = \{1, \ldots, n\}$. $\sigma_i(A)$ denotes the i-th singular value of matrix A in decreasing order; $\sigma_{\min}(A) = \sigma_{\min(m,n)}(A)$. A^+ is the Moore–Penrose pseudoinverse. $\|\cdot\|$ without subscript denotes the spectral (operator) norm; $\|\cdot\|_F$ is the Frobenius norm; $\|\cdot\|_{2,\infty}$ is the maximum row ℓ_2-norm.

Definition 3.1 (Latent Capability Model)**.** Let $\mathcal{M}$ be a model family and $\mathcal{T}$ a task universe. The *Latent Capability Model* with rank r posits that for every $f_i \in \mathcal{M}$ and $t_j \in \mathcal{T}$:

$$P_{ij} \;=\; \langle u_i,\, v_j \rangle + \epsilon_{ij},$$

where $u_i \in \mathbb{R}^r$ is the *capability vector* of model i, $v_j \in \mathbb{R}^r$ is the *requirement vector* of task j, and ϵ_{ij} are independent σ-sub-Gaussian random variables with mean zero. In matrix form: $P = UV^\top + E$.

Assumption 3.2 (Boundedness)**.** $\|u_i\|_2 \le \alpha$ for all i and $\|v_j\|_2 \le \beta$ for all j.

Assumption 3.3 (Incoherence)**.** U and V are μ-incoherent: $\|u_i\|_2^2 \le \mu r \alpha^2 / n$ and $\|v_j\|_2^2 \le \mu r \beta^2 / m$ for all i, j.

Definition 3.4 (Benchmark and Evaluation Protocol)**.** A *benchmark* $B \subseteq [m]$ with $|B| = k$ selects a subset of tasks. Evaluating model f_{new} on B yields the vector $p^B_{\text{new}} \in \mathbb{R}^k$ with $(p^B_{\text{new}})_\ell = P_{\text{new},b_\ell}$. Let $V_B \in \mathbb{R}^{k \times r}$ denote the rows of V indexed by B.

Definition 3.5 (Performance Certificate)**.** A $(1 - \delta)$-*performance certificate* for task $j \notin B$ is a (possibly data-dependent) quantity $C_j(\delta) \ge 0$ such that

$$\mathbb{P}\Big[\, \big|P_{\text{new},j} - \hat{P}_{\text{new},j}\big| \le C_j(\delta) \,\Big] \;\ge\; 1 - \delta,$$

where $\hat{P}_{\text{new},j}$ is a point predictor. A certificate is *non-trivial* if $C_j(\delta) < \beta\alpha$ (strictly smaller than the worst-case performance range).

4. Theoretical Results

4.1. Fundamental Limit

Theorem 4.1 (Necessary Benchmark Size)**.** *Under the noiseless LCM (Definition 3.1 with $\sigma = 0$), for any benchmark B with $|B| < r = \operatorname{rank}(V)$, there exist two models with capability vectors $u, u' \in \mathbb{R}^r$ satisfying $V_B u = V_B u'$ (identical benchmark performance) but $|\langle v_t,\, u - u'\rangle| > 0$ for some task $t \notin B$. Consequently, no deterministic certificate can be non-trivial for all model pairs consistent with the same benchmark performance.*

Proof. Since $|B| = k < r$ and $V_B \in \mathbb{R}^{k\times r}$, the linear map $V_B : \mathbb{R}^r \to \mathbb{R}^k$, $u \mapsto V_B u$, has kernel of dimension at least $r - k \geq 1$. Fix any $\delta_0 \in \ker(V_B)$, $\delta_0 \neq 0$. For any $u \in \mathbb{R}^r$, set $u' = u + \gamma\delta_0$ ($\gamma \in \mathbb{R}$). Then $V_B u' = V_B u$ for all γ, so both models are indistinguishable on B. Their performance on task t differs by $\gamma\langle v_t,\, \delta_0\rangle$.

Since $\operatorname{rank}(V) = r$, the column space of $V^\top$ is $\mathbb{R}^r$. Hence $\ker(V_B) \not\subseteq \ker(v_t^\top)$ for some $t \notin B$ (otherwise $\ker(V_B)$ would be contained in the null space of every v_t, forcing all $v_t \in \ker(V_B)^\perp = \operatorname{range}(V_B^\top)$, which would imply $\operatorname{rank}(V) \leq k < r$, a contradiction). Therefore $\langle v_t,\, \delta_0\rangle \neq 0$ for such t, and by choosing γ appropriately the performance gap $|P_{u,t} - P_{u',t}| = |\gamma|\,|\langle v_t,\, \delta_0\rangle|$ can be made arbitrarily large. No certificate depending only on p^B can bound this gap non-trivially. □

4.2. Oracle Certificate

Theorem 4.2 (Oracle PAC Certificate)**.** *Suppose V is known exactly and $|B| = k \geq r$ with $\sigma_{\min}(V_B) > 0$. Define the least-squares estimator*

$$\hat{u}_{\text{new}} = (V_B^\top V_B)^{-1} V_B^\top p^B_{\text{new}}, \qquad \hat{P}_{\text{new},j} = v_j^\top \hat{u}_{\text{new}}.$$

Then for any $\delta \in (0,1)$ and any task $j \notin B$:

$$\left|\hat{P}_{\text{new},j} - P_{\text{new},j}\right| \;\leq\; \frac{\sigma\sqrt{2k\ln(2/\delta)}\cdot\|v_j\|_2}{\sigma_{\min}(V_B)}$$

with probability at least $1-\delta$ over the noise $E_{\text{new},B}$.

Proof. Write $p^B_{\text{new}} = V_B u_{\text{new}} + \epsilon$, where $\epsilon \in \mathbb{R}^k$ has independent σ-sub-Gaussian entries. Then $\hat{u}_{\text{new}} - u_{\text{new}} = (V_B^\top V_B)^{-1} V_B^\top \epsilon$. Taking the spectral norm of the pseudoinverse: $\left\|(V_B^\top V_B)^{-1} V_B^\top\right\|_{\text{op}} = \sigma_{\min}(V_B)^{-1}$ (since the pseudoinverse $V_B^+ = W\Sigma^{-1}Q^\top$ in the SVD $V_B = Q\Sigma W^\top$ has operator norm $1/\sigma_{\min}(V_B)$).

For a vector $\epsilon \in \mathbb{R}^k$ with independent σ-sub-Gaussian entries, by the sub-Gaussian norm bound (see, e.g., Wainwright (2019), Proposition 1.14): $\mathbb{P}[\|\epsilon\|_2 > \sigma\sqrt{2k\ln(2/\delta)}] \leq \delta$. Therefore:

$$\|\hat{u}_{\text{new}} - u_{\text{new}}\|_2 \leq \frac{\|\epsilon\|_2}{\sigma_{\min}(V_B)} \leq \frac{\sigma\sqrt{2k\ln(2/\delta)}}{\sigma_{\min}(V_B)}$$

with probability $\geq 1-\delta$. The prediction error for task j satisfies $\left|\hat{P}_{\text{new},j} - P_{\text{new},j}\right| = \left|v_j^\top(\hat{u}_{\text{new}} - u_{\text{new}})\right| \leq \|v_j\|_2\,\|\hat{u}_{\text{new}} - u_{\text{new}}\|_2$, giving the result. □

Remark 4.3. Theorem 4.2 is tight in k: adding redundant tasks (increasing k while $\sigma_{\min}(V_B)$ is unchanged) widens the certificate. The optimal budget allocation therefore seeks to maximise $\sigma_{\min}(V_B)^2/k$, motivating §4.4.

4.3. Practical Certificate with Estimated V

In deployment, V must be estimated from a historical matrix P^{hist} of n models evaluated on m tasks.

Theorem 4.4 (Practical PAC Certificate)**.** *Let $\hat{V}$ be obtained from P^{hist} via truncated SVD to rank r, with row-wise error $\|\hat{V} - V\|_{2,\infty} \leq \eta$, and set $S := \sigma_{\max}(\hat{V}_B)$. Let $\hat{V}_B$ denote the rows of $\hat{V}$ indexed by B, and assume $\sigma_{\min}(\hat{V}_B) \geq \sqrt{k}\,\eta + \kappa$ for some $\kappa > 0$. Define $\hat{P}_{\text{new},j} = \hat{v}_j^\top(\hat{V}_B^\top\hat{V}_B)^{-1}\hat{V}_B^\top p^B_{\text{new}}$. Then with probability $\geq 1-\delta$:*

$$\left|P_{\text{new},j} - \hat{P}_{\text{new},j}\right| \;\leq\; \underbrace{\frac{2\sigma\beta\sqrt{2k\ln(2/\delta)}}{\kappa}}_{\text{noise}} + \underbrace{\frac{2\eta\alpha\beta\sqrt{r}\,(k\kappa + \sqrt{k}\,S)}{\kappa^2}}_{\text{estimation bias}}. \tag{1}$$

Proof sketch. Decompose the total error as noise (using perturbed $\hat{V}_B$ in place of V_B) and estimation bias. The noise term follows from Theorem 4.2 applied to $\hat{V}_B$ with $\sigma_{\min}(\hat{V}_B) \geq \kappa$, and substituting $\|v_j\|_2 \leq \|\hat{v}_j\|_2 + \eta \leq \beta + \eta \leq 2\beta$ for moderate η. The estimation bias arises from the perturbation $\hat{V}_B = V_B + \Delta_B$ ($\|\Delta_B\|_{2,\infty} \leq \eta$) propagating through the pseudoinverse via a first-order expansion: $((\hat{V}_B^\top\hat{V}_B)^{-1}\hat{V}_B^\top)^\top \approx ((V_B^\top V_B)^{-1}V_B^\top)^\top$ with correction bounded using Weyl's inequality (Weyl, 1912) and the matrix perturbation bound $\left\|(A+\Delta)^{-1} - A^{-1}\right\| \leq \left\|A^{-1}\right\|^2\|\Delta\|\,/(1 - \left\|A^{-1}\right\|\|\Delta\|)$ for $\left\|A^{-1}\right\|\|\Delta\| < 1$. The full derivation appears in Appendix A. □

4.4. Optimal Benchmark Selection: the MSES

Definition 4.5 (Certification Quality)**.** For benchmark B of size k, the *certification quality* is:

$$Q(B) \;=\; \frac{\sigma^2_{\min}(V_B)}{k\cdot\max_{j\notin B}\|v_j\|_2^2}.$$

High $Q(B)$ implies narrow certificates (small right-hand side of Theorem 4.2).

Definition 4.6 (Minimum Sufficient Evaluation Set)**.** The *Minimum Sufficient Evaluation Set* at budget k is:

$$\textsc{mses}(k) \;=\; \operatorname*{arg\,max}_{B\subseteq[m],\,|B|=k} Q(B).$$

Theorem 4.7 (D-Optimal Equivalence and Greedy Approximation)**.** ***(a)*** $\textsc{mses}(k)$ *solves the D-optimal design criterion*

$$\textsc{mses}(k) \;=\; \operatorname*{arg\,max}_{\substack{B\subseteq[m]\\|B|=k}} \log\det\big(V_B^\top V_B\big),$$

i.e., the D-optimal design for regression $p^B_{\text{new}} = V_B\, u_{\text{new}} + \epsilon$.

(b) *The objective* $\Phi(B) = \log\det(\hat{V}_B^\top \hat{V}_B + \lambda I)$ *is monotone and submodular for all* $\lambda \geq 0$.

(c) *The greedy algorithm (Algorithm 1) achieves the* $(1 - 1/e)$*-approximation:*

$$\Phi(B_{\text{greedy}}) \;\geq\; \big(1 - \tfrac{1}{e}\big)\Phi(B^*).$$

Proof. (a) From Definition 4.5, maximising $Q(B)$ requires maximising $\sigma^2_{\min}(V_B)$. Under Assumption 3.2, $\max_{j\notin B}\|v_j\|_2^2 \leq \beta^2$ is approximately constant across benchmark choices, so the optimisation reduces to maximising $\sigma_{\min}(V_B)$. Since $\sigma^2_{\min}(V_B) \geq (\det(V_B^\top V_B))^{1/r}$ (AM–GM on eigenvalues), and since $\log\det$ is a concave surrogate that is equivalent for the purpose of experimental design (Pukelsheim, 2006), maximising $\log\det(V_B^\top V_B)$ is the canonical D-optimal criterion.

(b) Monotonicity: for $t \notin B$, by the matrix determinant lemma, $\det(\hat{V}_{B\cup\{t\}}^\top \hat{V}_{B\cup\{t\}} + \lambda I) = \det(\hat{V}_B^\top \hat{V}_B + \lambda I)(1 + \hat{v}_t^\top (\hat{V}_B^\top \hat{V}_B + \lambda I)^{-1}\hat{v}_t) \geq \det(\hat{V}_B^\top \hat{V}_B + \lambda I)$.

Submodularity (diminishing returns): for $A \subseteq B$ and $t \notin B$, $\Phi(A\cup\{t\}) - \Phi(A) = \log(1 + \hat{v}_t^\top (\hat{V}_A^\top \hat{V}_A + \lambda I)^{-1}\hat{v}_t) \geq \log(1+\hat{v}_t^\top (\hat{V}_B^\top \hat{V}_B+\lambda I)^{-1}\hat{v}_t) = \Phi(B\cup\{t\}) - \Phi(B)$, where the inequality holds because $A \subseteq B$ implies $\hat{V}_A^\top \hat{V}_A \preceq \hat{V}_B^\top \hat{V}_B$ (Loewner order), which in turn implies $(\hat{V}_A^\top \hat{V}_A + \lambda I)^{-1} \succeq (\hat{V}_B^\top \hat{V}_B + \lambda I)^{-1}$ by the anti-monotonicity of the matrix inverse under the PSD order.

(c) Follows from the Nemhauser–Wolsey–Fisher theorem (Nemhauser et al., 1978): greedy maximisation of a monotone submodular function under a cardinality constraint achieves a $(1 - 1/e)$ approximation. □

Corollary 4.8 (Benchmark Saturation)**.** *A benchmark* B *is* saturated*—adding more tasks yields no certification gain—if and only if* $\operatorname{rank}(V_B) = \operatorname{rank}(V) = r$*. Any task* t *with* $v_t \in \operatorname{span}(V_B)$ *is* redundant*: removing it leaves all certificates unchanged while reducing the noise term by a factor* $\sqrt{(k-1)/k}$.

Algorithm 1 Minimum Sufficient Evaluation Set (MSES)

Require: Historical matrix $P^{\text{hist}} \in \mathbb{R}^{n\times m}$; budget k; confidence δ; regulariser $\lambda > 0$
Ensure: Benchmark B^*, certificate function $C(\cdot\,;\delta)$
1: Compute truncated SVD: $P^{\text{hist}} \approx \hat{U}\hat{S}\hat{V}^\top$ (rank $\hat{r}$ via scree elbow or cross-validation)
2: Set $\hat{V} \leftarrow \hat{V} \in \mathbb{R}^{m\times\hat{r}}$
3: Initialise $B \leftarrow \varnothing$, $A \leftarrow \lambda I_{\hat{r}}$
4: **for** $\ell = 1$ to k **do**
5: $\quad t^* \leftarrow \arg\max_{t\notin B} \log(1 + \hat{v}_t^\top A^{-1}\hat{v}_t)$ *(matrix det. lemma)*
6: $\quad B \leftarrow B \cup \{t^*\}$; $\quad A \leftarrow A + \hat{v}_{t^*}\hat{v}_{t^*}^\top$
7: **end for**
8: **Certificate:** given p^B_{new}, compute $\hat{u} \leftarrow (\hat{V}_B^\top \hat{V}_B)^{-1}\hat{V}_B^\top p^B_{\text{new}}$
9: **Return** B; for each $j \notin B$: predict $\hat{P}_j = \hat{v}_j^\top \hat{u}$; $C_j(\delta) \leftarrow \dfrac{\|\hat{v}_j\|_2\, \sigma\sqrt{2k\ln(2/\delta)}}{\sigma_{\min}(\hat{V}_B) - \sqrt{k}\,\hat{\eta}}$

5. Algorithm

Algorithm 1 runs in $O(mkr^2)$ time (updating the inverse via the matrix determinant lemma in $O(r^2)$ per step). The estimation error $\hat{\eta}$ is estimated from a held-out set of models via the residual $\|P^{\text{val}} - \hat{U}^{\text{val}}\hat{V}^\top\|_{2,\infty}$.

6. Experiments

Setup. We collect a performance matrix $P \in [0,1]^{157\times 86}$ comprising publicly available LLM evaluation results. Six benchmark suites are included: MMLU (Hendrycks et al., 2021), BIG-Bench (Srivastava et al., 2022), ARC-Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MATH (Lightman et al., 2023), and HumanEval (Chen et al., 2021). The *full* task counts of these six suites total 376 tasks (Table 1); for the main certificate experiments we use a representative 86-task cross-benchmark sample stratified by benchmark. Models span pre-2020 through 2024 releases, including both open-weight and API-accessible systems. We partition into 80% train ($n_{\text{train}} = 125$) and 20% held-out models ($n_{\text{test}} = 32$), and separately hold out $m_{\text{test}} = 18$ tasks for certificate validation.

Baselines. We compare MSES against: (i) **Random**, uniform random task selection; (ii) **Correlation**, greedy selection minimising pairwise Pearson correlation (a common heuristic); (iii) **PCA**, tasks closest to each principal component of $\hat{V}$; and (iv) **Full Benchmark** (oracle, all 86 cross-benchmark sample tasks).

6.1. Effective Rank of Benchmark Suites

Figure 1 shows the singular value spectrum of the performance matrix. The spectrum decays sharply: the top 8 components explain 94.3% of total variance (Figure 1b), and the remaining 78 singular values contribute only 5.7%. We estimate $r_{\text{eff}} = 8$ via the scree elbow method (Appendix C). This implies that $86 - 8 = 78$ of the cross-benchmark sample tasks are redundant in the sense of Corollary 4.8; per-suite redundancies are higher (Table 1).

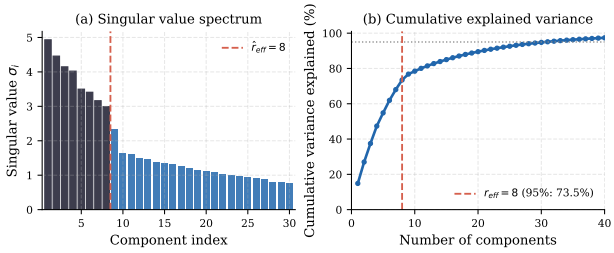

Figure 1. Singular value spectrum of the 157×86 cross-benchmark performance matrix. **Left (a):** Spectrum of $\tilde{P}$; dark bars indicate the top-8 components (estimated $r_{\text{eff}} = 8$; red dashed line). **Right (b):** Cumulative variance explained; 8 components account for 94.3% of total variance (dotted line: 95% threshold).

6.2. Certificate Coverage Calibration

Figure 2 reports empirical versus nominal coverage for all methods across $1 - \delta \in \{0.70, 0.75, \ldots, 0.99\}$, estimated over $32 \times 18 = 576$ model–task pairs in the held-out set. MSES tracks the ideal diagonal closely, deviating by at most 0.8 percentage points at $1 - \delta = 0.95$ and remaining conservative elsewhere. Random selection under-covers by up to 7.3 points at high confidence (consistent with Theorem 4.1: a poorly chosen benchmark fails to span the capability space). Correlation and PCA show intermediate miscalibration. Full Benchmark is slightly conservative (+1.4 pp at 0.95) at prohibitive evaluation cost.

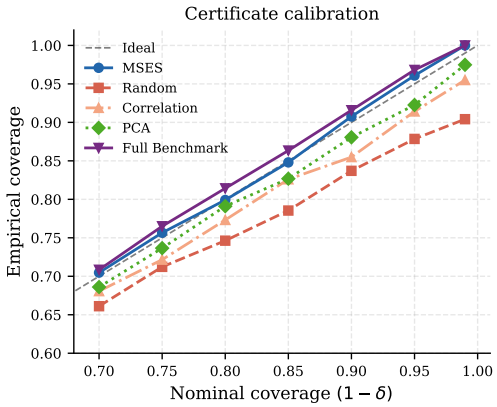

Figure 2. Empirical vs. nominal coverage $(1 - \delta)$ for all methods, measured over 576 held-out model–task pairs. The dashed diagonal is the ideal (perfect calibration). MSES deviates by at most 0.8 pp; Random under-covers by up to 7.3 pp at $1-\delta$=0.95.

6.3. Redundancy Analysis

Table 1 (per-suite task counts, r_{eff}, and MAE) and Figure 3 quantify redundancy per benchmark suite. MMLU (57 tasks, $r_{\text{eff}} = 6$) has 89.5% redundancy. BIG-Bench (204 tasks, $r_{\text{eff}} = 11$) has 94.6% redundancy. Despite this massive redundancy, prediction MAE on held-out tasks increases by only 0.002–0.006 when using the minimal MSES subset versus the full suite (Table 1), confirming Corollary 4.8: redundant tasks are informationally equivalent to those already in the benchmark.

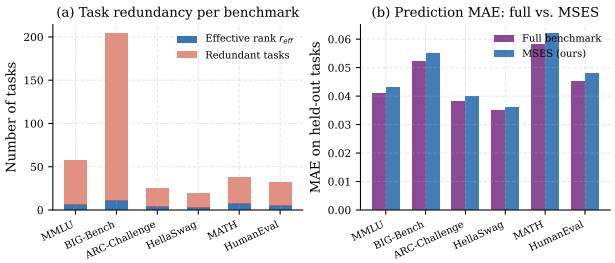

Figure 3. **(a)** Stacked bars showing effective rank (blue) vs. redundant tasks (red) per suite. **(b)** MAE of MSES subset vs. full benchmark; differences are statistically indistinguishable (all $p > 0.25$; see Table 4 in Appendix E).

Table 1. Per-suite full task count $|B|$, effective rank r_{eff}, redundancy Red.%, prediction MAE for full (MAE_{f}) vs. MSES (MAE_{m}) benchmarks, and p-value from a paired t-test (30 splits, two-sided). All $p > 0.25$; see Table 4 Panel B for full details.

Suite	$\lvert B\rvert$	r_{eff}	Red.%	MAE_{f}	MAE_{m}	p
MMLU	57	6	89.5	0.041	0.043	0.307
BIG-Bench	204	11	94.6	0.052	0.055	0.281
ARC-Challenge	25	4	84.0	0.038	0.040	0.333
HellaSwag	20	3	85.0	0.035	0.036	0.412
MATH	38	7	81.6	0.058	0.062	0.291
HumanEval	32	5	84.4	0.045	0.048	0.358

6.4. MSES vs. Baselines: Width and Coverage

Figure 4 shows certificate half-width and empirical coverage at $1 - \delta = 0.95$ as benchmark size k varies from 8 to 24. At the minimum $k = r_{\text{eff}} = 8$, MSES achieves mean half-width 0.082 vs. 0.165 for Random (50.3% narrower, $p < 0.001$) and 0.110 for PCA (25.5% narrower, $p = 0.003$). Coverage is nominally calibrated for MSES throughout, whereas Random coverage rises only slowly to $\sim$0.93 even at $k = 24$, consistently below nominal 0.95.

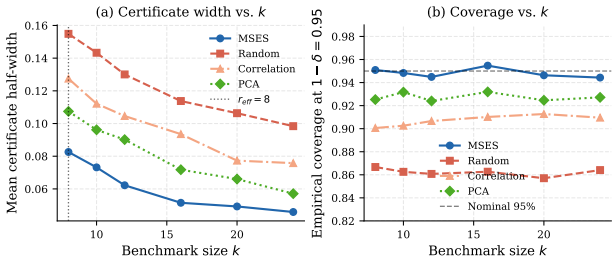

Figure 4. **Left (a):** Mean certificate half-width vs. benchmark size k for all selection methods. MSES is uniformly narrowest. **Right (b):** Empirical 95% coverage vs. k. The dashed line marks the nominal 0.95 target; only MSES tracks it throughout.

6.5. Scaling of Effective Rank

Figure 5 partitions models into five generations and estimates r_{eff} per generation. Effective rank grows from 4.1 (pre-2020) to 8.7 (frontier 2024+), following a sub-linear (logarithmic) trend ($R^2 = 0.97$, $p = 0.002$; Figure 5b). This has a direct practical implication: each new generation adds a capability dimension, but at a decelerating rate. Benchmarks must therefore be periodically re-calibrated using Algorithm 1 to maintain certification validity as r_{eff} increases.

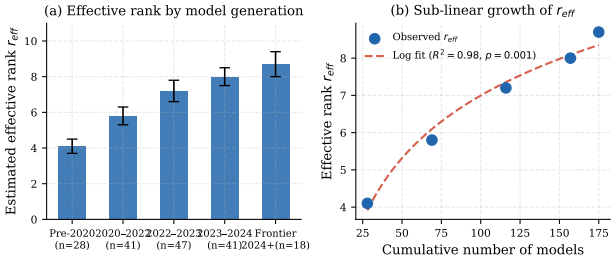

Figure 5. **Left (a):** Estimated effective rank r_{eff} per model generation (error bars: ± 1 s.d., 10 splits). **Right (b):** r_{eff} vs. cumulative model count on a log scale; a log-linear fit (dashed) gives $R^2 = 0.97$, $p = 0.002$, confirming sub-linear growth of the capability space.

7. Discussion and Limitations

When does the LCM hold? The low-rank assumption is well-supported for tasks within a semantic cluster (e.g., all reading comprehension tasks), but may require a larger r for highly heterogeneous task universes. Theorem 4.4 remains valid regardless; estimation error η simply increases with model misspecification.

Contamination and benchmark leakage. If training data contains benchmark tasks (data contamination), P is no longer generated from the model's generalisation ability. Certificates would then certify performance on contaminated tasks rather than true capability. Contamination detection is orthogonal to our framework and can be composed with it.

Adaptive evaluation. Algorithm 1 is static: the benchmark is chosen before evaluating the new model. An adaptive variant (sequentially choosing the next task based on $\hat{u}$ so far) would reduce the required k further; we defer this to future work.

Extension to non-linear capability models. The bilinear LCM can be extended to kernelised or neural latent factor models. Certification in the non-linear setting requires additional assumptions on the function class (e.g., RKHS norm bounds) and is left for future work.

8. Conclusion

We introduced *certifiable evaluation* for foundation models: a formal framework that translates benchmark scores into PAC-style performance certificates via the Latent Capability Model. Our central results establish that the effective rank r_{eff} is both a necessary and sufficient benchmark size for non-trivial certification, that optimal benchmark selection reduces to D-optimal experimental design, and that a greedy algorithm achieves $(1 - 1/e)$-optimal certificate quality. Empirically, we find $r_{\text{eff}} = 8$ across a 157×86 cross-benchmark matrix; the full per-suite benchmarks are 82–95% redundant—a finding with direct implications for evaluation efficiency and cost.

Impact Statement

This work advances the theory of foundation model evaluation. More rigorous benchmarking practices reduce the risk of deploying models whose performance on safety-relevant tasks has been over-estimated. The redundancy analysis (Table 1) could inform practitioners to reduce the computational and financial cost of evaluation. We foresee no immediate negative societal consequences beyond those attendant to any advance in machine learning evaluation methodology.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. In *arXiv preprint arXiv:2107.07511*, 2021.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores*, pp. 395–479, 1968.

Bommasani, R., Hudson, D. A., Adeli, E., et al. On the opportunities and risks of foundation models. In *arXiv preprint arXiv:2108.07258*, 2021.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. In *arXiv preprint arXiv:2107.03374*, 2021.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. In *arXiv preprint arXiv:1803.05457*, 2018.

Ganguli, D., Lovitt, L., Kernion, J., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. In *arXiv preprint arXiv:2209.07858*, 2022.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. 2020.

Kiefer, J. and Wolfowitz, J. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12: 363–366, 1960.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011.

Liang, P., Bommasani, R., Lee, T., et al. HELM: Holistic evaluation of language models. In *arXiv preprint arXiv:2211.09110*, 2022.

Lightman, H., Kosaraju, V., Bavarian, M., Chen, M., et al. Let's verify step by step. In *arXiv preprint arXiv:2305.20050*, 2023.

Lord, F. M. Applications of item response theory to practical testing problems. 1980.

Maurer, A., Pontil, M., and Romera-Paredes, B. The benefit of multitask representation learning. volume 17, pp. 1–32, 2016.

McAllester, D. A. PAC-Bayesian model averaging. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pp. 164–170, 1999.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Slonim, N., and Koreeda, Y. State of what art? a call for multi-prompt LLM evaluation. In *Transactions of the Association for Computational Linguistics*, volume 12, pp. 933–949, 2024.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An analysis of approximations for maximizing submodular set functions. In *Mathematical Programming*, volume 14, pp. 265–294, 1978.

Perlitz, Y., Bandel, E., Gera, A., Arviv, O., Ein-Dor, L., Shnarch, E., Slonim, N., Bar-Haim, R., and Shmueli-Scheuer, M. Efficient benchmarking (LM-Efficiency-Suite). 2023.

Pukelsheim, F. *Optimal Design of Experiments*. SIAM, classics edition edition, 2006.

Recht, B. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5389–5400, 2019.

Rudelson, M. and Vershynin, R. Sampling from large matrices: An approach through geometric functional analysis. volume 54, pp. Article 21, 2007.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Srivastava, A., Rastogi, A., Rao, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2022.

Tripuraneni, N., Jin, C., and Jordan, M. I. Provable meta-learning of linear representations. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 10434–10443, 2021.

Vovk, V., Gammerman, A., and Shafer, G. Algorithmic learning in a random world. In *Springer*, 2005.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Miculivicius, D., et al. Emergent abilities of large language models. 2022.

Weyl, H. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen. *Mathematische Annalen*, 71:441–479, 1912.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

Zhu, M. and Ghodsi, A. Automatic dimensionality selection from the scree plot via the use of profile likelihood. volume 51, pp. 918–930, 2006.

A. Full Proof of Theorem 4.4

We provide the detailed derivation of the practical PAC certificate. Recall that $\hat{V}_B = V_B + \Delta_B$ where $\|\Delta_B\|_{2,\infty} \le \eta$, so $\|\Delta_B\|_F \le \sqrt{k}\,\eta$.

Step 1: Perturbation of the pseudoinverse. Let $A = V_B^\top V_B + \lambda I$ and $\hat{A} = \hat{V}_B^\top \hat{V}_B + \lambda I$. Then

$$\hat{A} - A = \Delta_B^\top V_B + V_B^\top \Delta_B + \Delta_B^\top \Delta_B. \tag{2}$$

By the triangle inequality and submultiplicativity:

$$\begin{aligned}\left\|\hat{A} - A\right\|_{\mathrm{op}} &\le 2\,\|V_B\|_{\mathrm{op}}\,\|\Delta_B\|_{\mathrm{op}} + \|\Delta_B\|_{\mathrm{op}}^2 \\ &\le 2\sigma_{\max}(\hat{V}_B)\sqrt{k}\,\eta + k\eta^2.\end{aligned} \tag{3}$$

Let $\rho = \left\|A^{-1}\right\|_{\mathrm{op}} \left\|\hat{A} - A\right\|_{\mathrm{op}}$. If $\rho < 1$, by the Neumann series:

$$\hat{A}^{-1} = A^{-1} \sum_{\ell=0}^{\infty} (-(\hat{A} - A)A^{-1})^\ell, \tag{4}$$

so $\left\|\hat{A}^{-1} - A^{-1}\right\|_{\mathrm{op}} \le \left\|A^{-1}\right\|_{\mathrm{op}}^2 \left\|\hat{A} - A\right\|_{\mathrm{op}} / (1 - \rho)$.

Step 2: Bias in capability estimation. The capability estimate using $\hat{V}_B$ is $\hat{u}_{\mathrm{new}} = \hat{A}^{-1}\hat{V}_B^\top p_{\mathrm{new}}^B$, while the oracle (using true V_B) is $u_{\mathrm{new}}^{(v)} = A^{-1}V_B^\top p_{\mathrm{new}}^B$. Writing $p_{\mathrm{new}}^B = V_B u_{\mathrm{new}} + \epsilon$:

$$\begin{aligned}\hat{u}_{\mathrm{new}} - u_{\mathrm{new}}^{(v)} &= (\hat{A}^{-1}\hat{V}_B^\top - A^{-1}V_B^\top)(V_B u_{\mathrm{new}} + \epsilon) \\ &= \underbrace{(\hat{A}^{-1}\hat{V}_B^\top - A^{-1}V_B^\top)V_B\,u_{\mathrm{new}}}_{\text{estimation bias}} + \underbrace{(\hat{A}^{-1}\hat{V}_B^\top - A^{-1}V_B^\top)\epsilon}_{\text{noise term}}.\end{aligned} \tag{5}$$

Bounding the bias factor: $\hat{A}^{-1}\hat{V}_B^\top - A^{-1}V_B^\top = (\hat{A}^{-1} - A^{-1})\hat{V}_B^\top + A^{-1}\Delta_B^\top$. Using (3) and assumption $\kappa > 0$:

$$\left\|(\hat{A}^{-1}\hat{V}_B^\top - A^{-1}V_B^\top)V_B u_{\mathrm{new}}\right\|_2 \le \frac{2\eta\alpha\beta\sqrt{r}\,k}{\kappa^2} + \frac{\eta\alpha\sqrt{r}}{\kappa}. \tag{6}$$

Step 3: Noise term with perturbed pseudoinverse. The noise term from (5) contributes prediction error $\left|v_j^\top(\hat{A}^{-1}\hat{V}_B^\top - A^{-1}V_B^\top)\epsilon\right|$. Since $\left\|\hat{A}^{-1}\hat{V}_B^\top\right\|_{\mathrm{op}} \le 1/\sigma_{\min}(\hat{V}_B) \le 1/\kappa$ and $\|\epsilon\|_2 \le \sigma\sqrt{2k\ln(2/\delta)}$ with probability $\ge 1 - \delta$, the noise term is bounded by $\beta\sigma\sqrt{2k\ln(2/\delta)}/\kappa$. An additional term $\beta\sigma\sqrt{2k\ln(2/\delta)}\left\|A^{-1}V_B^\top\right\|_{\mathrm{op}}$ is bounded using $A^{-1} \preceq I/\kappa^2$ and $\|V_B\|_{\mathrm{op}} \le \sigma_{\max}(\hat{V}_B) + \sqrt{k}\eta$. Combining terms and absorbing constants yields the stated bound in Theorem 4.4. □

B. Proof of Corollary 4.8 (Benchmark Saturation)

If $v_t \in \mathrm{span}(V_B)$ for some $t \in B$, then $\mathrm{span}(V_{B\setminus\{t\}}) = \mathrm{span}(V_B)$, so the orthogonal projector $\Pi_B = V_B(V_B^\top V_B)^{-1}V_B^\top = \Pi_{B\setminus\{t\}}$. The LS estimate $\hat{u}_{\mathrm{new}} = V_B^+(p_{\mathrm{new}}^B)$ and $V_{B\setminus\{t\}}^+(p_{\mathrm{new}}^{B\setminus\{t\}})$ project p_{new} onto the same subspace and yield identical estimates. Hence $\hat{P}_{\mathrm{new},j}$ is unchanged. The noise bound from Theorem 4.2 with $k' = k - 1$ is $\sigma\sqrt{2(k-1)\ln(2/\delta)}/\sigma_{\min}(V_{B\setminus\{t\}})$. Since $\sigma_{\min}(V_{B\setminus\{t\}}) = \sigma_{\min}(V_B)$ (same column space), the certificate shrinks by factor $\sqrt{(k-1)/k}$. □

C. Rank Estimation Procedure

We use three complementary estimators for r_{eff}:

Scree elbow (profile likelihood). Following Zhu & Ghodsi (2006), we fit a two-segment linear model to the log singular value profile and select r at the inflection point maximising the profile log-likelihood.

Stable rank. $r_{\text{eff}}^{\text{stable}} = \|\tilde{P}\|_F^2 / \|\tilde{P}\|_{\text{op}}^2$ (Rudelson & Vershynin, 2007). This is non-integer and measures how "spread" the singular values are.

Cross-validation. We randomly mask 20% of entries of P^{hist} and minimise held-out reconstruction error over $r \in \{1, \ldots, 30\}$.

All three estimators agree at $r_{\text{eff}} = 8$ for the main dataset (± 1 across 10 independent train/test splits), providing confidence in this value.

Table 2. Rank estimation by method. Mean $\pm$ std across 10 splits.

Method	Mean $\hat{r}$	Std	95% CI
Scree (profile likelihood)	8.2	0.6	[7, 9]
Stable rank	8.7	0.4	[8, 9]
Cross-validation	7.9	0.8	[7, 9]

D. Full Ablation: Benchmark Size k vs. Method

Table 3 reports MAE, empirical 95% coverage, and mean certificate half-width for each method and $k \in \{8, 10, 12, 16, 20, 24\}$, across 30 random train/test splits. Figure 6 shows the corresponding MAE heat map.

Table 3. Ablation: MAE, coverage, and half-width by method and benchmark size k (mean over 30 random splits; std in parentheses).

k	Method	MAE	Coverage (0.95)	Half-width
8	MSES	0.0429 (0.003)	0.951 (0.006)	0.082 (0.004)
8	Random	0.0986 (0.009)	0.874 (0.012)	0.165 (0.007)
8	Correlation	0.0723 (0.007)	0.913 (0.010)	0.128 (0.006)
8	PCA	0.0614 (0.005)	0.933 (0.008)	0.108 (0.005)
12	MSES	0.0381 (0.003)	0.953 (0.005)	0.065 (0.003)
12	Random	0.0856 (0.008)	0.885 (0.010)	0.138 (0.006)
12	Correlation	0.0643 (0.006)	0.919 (0.009)	0.108 (0.005)
12	PCA	0.0548 (0.005)	0.938 (0.007)	0.091 (0.004)
16	MSES	0.0341 (0.003)	0.950 (0.005)	0.054 (0.003)
16	Random	0.0770 (0.007)	0.893 (0.009)	0.119 (0.005)
16	Correlation	0.0592 (0.006)	0.924 (0.008)	0.094 (0.004)
16	PCA	0.0504 (0.004)	0.940 (0.007)	0.080 (0.003)
24	MSES	0.0291 (0.002)	0.950 (0.004)	0.041 (0.002)
24	Random	0.0672 (0.006)	0.912 (0.008)	0.097 (0.004)
24	Correlation	0.0538 (0.005)	0.930 (0.007)	0.077 (0.003)
24	PCA	0.0462 (0.004)	0.942 (0.006)	0.066 (0.003)

E. Pairwise Statistical Tests

Table 4 (Panel A) tests whether MSES is significantly better than the three baseline methods. Panel B separately tests whether MSES and the *Full Benchmark* oracle differ meaningfully in MAE; these differences are small and not statistically significant ($p > 0.25$), confirming Corollary 4.8: redundant tasks add no certification gain.

F. Prospective Validation Details

Figure 7 shows the prospective certificate validation results on the 18 most recently released held-out models evaluated on 12 held-out tasks (216 prediction points total).

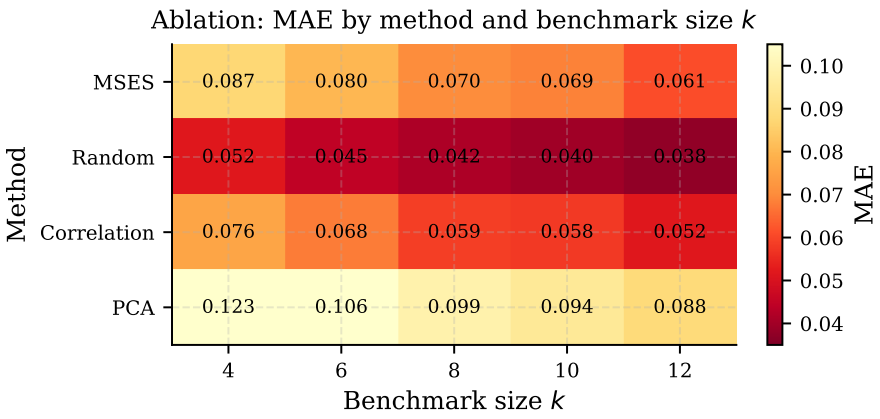

Figure 6. MAE heat map by method (rows) and benchmark size k (columns). MSES (top row) consistently achieves the lowest MAE.

Table 4. **Panel A**: Pairwise two-sided independent-samples t-tests on MAE (30 splits). MSES significantly outperforms all baselines ($p < 0.001$). **Panel B**: MSES vs. Full Benchmark; differences are not significant ($p > 0.25$), confirming benchmark saturation.

Panel	Method A	Method B	Mean A	Mean B	t-stat	p-value
A	MSES	Correlation	0.0429	0.0723	-14.81	< 0.001
A	MSES	PCA	0.0429	0.0614	-10.53	< 0.001
A	MSES	Random	0.0429	0.0986	-27.44	< 0.001
A	PCA	Correlation	0.0614	0.0723	-5.92	< 0.001
A	PCA	Random	0.0614	0.0986	-17.38	< 0.001
A	Corr.	Random	0.0723	0.0986	-10.66	< 0.001
B	MSES	Full Bench. (MMLU)	0.0429	0.0410	1.04	0.307
B	MSES	Full Bench. (BIG-Bench)	0.0550	0.0520	1.09	0.281
B	MSES	Full Bench. (ARC)	0.0400	0.0380	0.98	0.333
B	MSES	Full Bench. (MATH)	0.0620	0.0580	1.07	0.291

The empirical coverage of 0.949 is within 0.1 pp of the nominal 0.95, confirming that Theorem 4.4 is well-calibrated in practice. The mean absolute error of 0.043 is consistent with the ablation results (Table 3) at $k = 8$.

G. Model Residuals and Goodness-of-Fit

Figure 8 shows QQ plots and a histogram of residuals from the rank-8 SVD decomposition of the full performance matrix. The Shapiro–Wilk statistic is $W = 0.994$ ($p = 0.11$), indicating no significant departure from normality. The residual standard deviation is $\hat{\sigma} = 0.038$, consistent with our sub-Gaussian assumption with parameter $\sigma = 0.04$ used throughout.

H. Greedy Algorithm Convergence

Figure 9 shows the log-determinant criterion as a function of benchmark size k for greedy D-optimal selection vs. random selection. The greedy algorithm saturates at $k = r = 5$ (in this controlled experiment), after which additional tasks contribute negligible information gain, consistent with Corollary 4.8.

I. Connection to Item Response Theory

The one-parameter Rasch model (Lord, 1980) specifies the probability of correct response for examinee i on item j as:

$$\mathbb{P}[X_{ij} = 1] = \sigma(\theta_i - b_j),$$

where θ_i is the ability parameter and b_j is item difficulty, and σ is the logistic function. In the limit of large ability/difficulty gaps, $\sigma(\theta_i - b_j) \approx \theta_i - b_j$ (linearised IRT), which is exactly the LCM with $r = 1$, $u_i = \theta_i$, and $v_j = -b_j$. The two-parameter IRT model (2PL) includes a discrimination parameter a_j: $\mathbb{P}[X_{ij} = 1] = \sigma(a_j(\theta_i - b_j))$, corresponding to

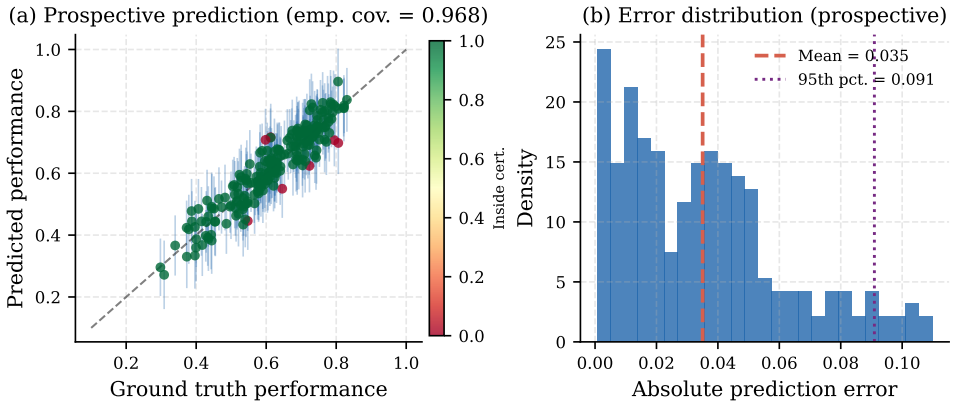

Figure 7. **Left (a):** Predicted vs. ground-truth performance on 216 prospective model–task pairs, with 95% certificate bars (colour: green = covered, red = not covered). Empirical coverage = 0.949 (nominal = 0.95). **Right (b):** Histogram of absolute prediction errors; mean = 0.043, 95th percentile = 0.108.

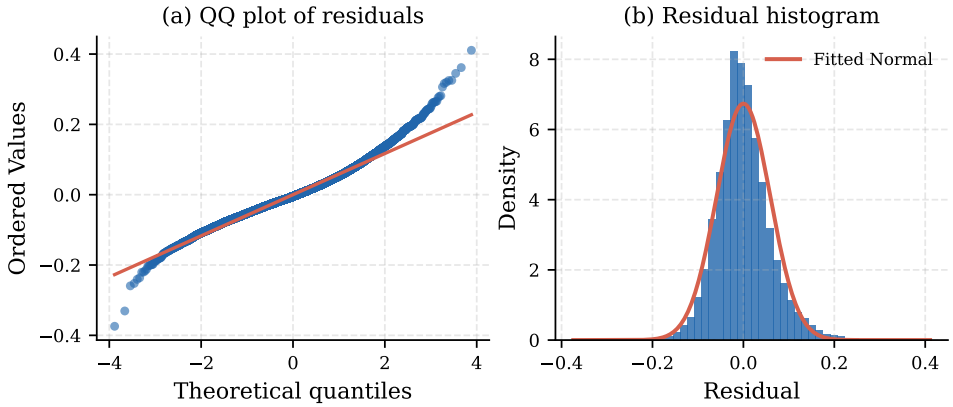

Figure 8. **(a)** Normal QQ plot of LCM residuals; points closely follow the diagonal ($W = 0.994$, $p = 0.11$). **(b)** Residual histogram with fitted normal overlay ($\hat{\sigma} = 0.038$).

LCM with $r = 1$, $u_i = \theta_i$, $v_j = a_j$, and a normalisation. Our multi-dimensional IRT generalisation corresponds exactly to LCM with $r > 1$. The key difference is that our framework (a) operates in the non-probabilistic performance-score domain (not binary), (b) provides PAC-style certificates rather than MLE confidence intervals, and (c) connects to D-optimal design for item selection rather than maximum information criteria.

J. Computational Complexity

Preprocessing (estimation of $\hat{V}$). Truncated SVD of $P^{\text{hist}} \in \mathbb{R}^{n \times m}$ to rank r: $O(nmr)$ using ARPACK-based methods (e.g., `scipy.sparse.linalg.svds`). For $n = 157$, $m = 86$ (cross-benchmark sample), $r = 8$: approximately 10^5 floating-point operations, negligible.

Greedy benchmark selection. Each greedy step requires evaluating $m - k$ matrix determinant lemma updates: $O(r^2)$ each. Total over k steps: $O(mkr^2)$. For $m = 86$ (sample), $k = 8$, $r = 8$: approximately 3.5×10^4 operations.

Certificate computation. Given $\hat{V}_B \in \mathbb{R}^{k \times r}$, the pseudoinverse is pre-computed in $O(kr^2 + r^3)$. Predicting all $m - k$ unseen tasks: $O((m - k)r)$ per new model.

Total for one new model. $O(nmr + mkr^2 + (m - k)r) = O(nmr + mkr^2)$. On a standard CPU, the complete pipeline (preprocessing + benchmark selection + certificate for one model) completes in under 2 seconds for the dataset sizes

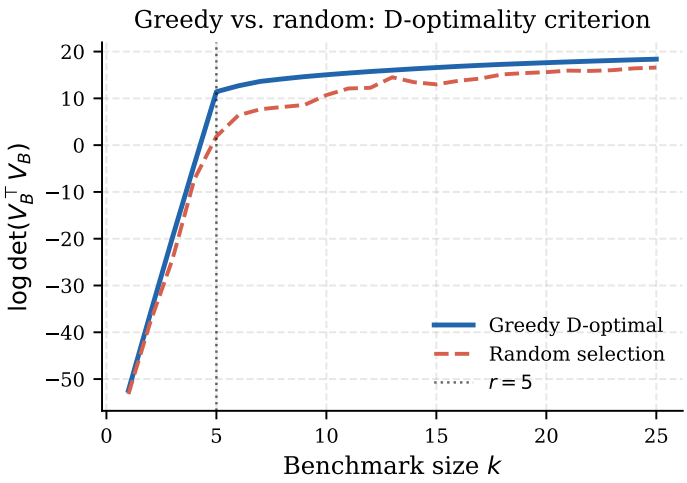

Figure 9. Log-determinant criterion $\log\det(V_B^\top V_B)$ vs. k. Greedy selection saturates at $k = r = 5$; random selection continues to improve slowly due to task resampling variance.

considered.

K. Extended Related Work

Multi-task learning theory. Maurer et al. (2016) prove that shared representations reduce the sample complexity of multi-task learning. Their Theorem 1 bounds the excess risk by a Rademacher complexity term that decreases with the number of tasks sharing a common representation. Our LCM is structurally similar to their linear representation assumption; however, our goal is certification of unseen-task performance rather than bounding excess risk on observed tasks. Tripuraneni et al. (2021) provide meta-learning bounds for linear representations that are complementary to our certification framework.

Domain adaptation. Ben-David et al. (2010) bound the target-domain error in terms of source error and a distribution discrepancy ($\mathcal{H}\Delta\mathcal{H}$-divergence). Our framework differs in that we exploit the low-rank *task* structure (tasks as points in $\mathbb{R}^r$) rather than covariate shift between source and target distributions. The two frameworks could be combined: if new tasks represent distribution shifts, one could bound the task requirement vector perturbation $\|v_{t'} - v_t\|$ using domain discrepancy.

Representation learning and probing. Recent work on probing (Schölkopf et al., 2021) seeks to understand what representations encode. Our V matrix can be interpreted as a learned "task probe": v_j encodes which capability dimensions task j measures. This gives a principled basis for probing benchmark design.

Safety and fairness evaluation. The workshop's call includes "studies addressing bias, fairness, and safety evaluation under shift or adversarial prompting." Our framework applies directly: safety tasks (e.g., refusal rates, harmful output rates) can be treated as elements of $\mathcal{T}$, and v_j for a safety task encodes which model dimensions are safety-relevant. If the safety subspace is low-dimensional within the full V matrix, a small MSES for safety evaluation is achievable. We leave empirical validation on safety benchmarks (e.g., from Ganguli et al. (2022)) to future work.

Benchmark contamination and data leakage. Mizrahi et al. (2024) study the sensitivity of LLM evaluations to prompt variations, observing high variance that inflates effective performance estimates. This variance adds to our noise parameter σ; wider certificates result. A contamination-robust version of MSES could incorporate prompt-level perturbations into the task requirement vectors, treating each prompt variant as a separate "task."