RANDOM IS ALMOST ENOUGH? STABLE AND LOW-COST BENCHMARK COMPRESSION VIA BESTRAND

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmark compression reduces the often prohibitive time and cost of model evaluation by evaluating models on a small benchmark subset that closely matches full-benchmark results. To quantify estimation error, we derive an accurate analytical approximation to the error distribution when estimating full-benchmark scores from a subset, showing that most subsets yield small errors, while a few produce much larger errors, resulting in a long-tailed distribution. This means random sampling alone can achieve surprisingly strong performance with high probability, but remains unstable due to rare large errors. We further show that this instability is inherent to modern foundation models and persists even as more models are evaluated. Based on these findings, we introduce BestRand, a stable and low-cost benchmark compression method that preserves the efficiency of random sampling while substantially reducing its high-error tail. We validate BestRand on 70+ text, multimodal, and agent benchmarks, and further extend it to VLA evaluation. Compared with competitive methods, BestRand reduces high-error risk by 50.2% and estimation error by 16.4% on average, while requiring only 51.3% of the construction cost. It further reduces OOD estimation error by 21.8% on average across score-range, model-family, and temporal shifts. Together, our theory provides practical guidance for understanding and predicting the accuracy and reliability of benchmark compression, while BestRand enables stable and accurate compression at low cost. All code and data will be released.

1 INTRODUCTION

Evaluating large language models (LLMs) and vision-language models (VLMs) on benchmarks can require substantial time and compute (Polo et al., 2024), as these benchmarks often contain many test examples and may also include redundant ones (Zhang et al., 2025b). Benchmark compression reduces this burden by evaluating models on a small subset to approximate full-benchmark results (Vivek et al., 2024). Yet the accuracy and reliability of this approximation are not well characterized.

To quantify estimation error, we derive an accurate analytical approximation to the error distribution when using a subset to estimate full-benchmark scores (Truong et al., 2025). Our theoretical and empirical results show that most subsets yield small errors, while a few produce much larger errors, resulting in a long-tailed distribution. This means random sampling alone can achieve surprisingly strong performance with high probability (Saranathan et al., 2025), but remains unstable due to rare large errors (Huang et al., 2025b). We further explain this high-probability effectiveness and show that the instability arises from the similarity among LLMs and VLMs and persists even as more models are evaluated. Beyond explaining these phenomena, our theory quantifies the relationship between subset size and estimation error (Zhang et al., 2025a), guiding the balance between evaluation accuracy and cost (Perlitz et al., 2024). Across all theoretical results, numerical experiments consistently agree with our theory. Figure 1 and Appendix F provide extensive numerical validation.

Motivated by our theory, we introduce BestRand, a stable and low-cost benchmark compression method that preserves the efficiency of random sampling while substantially reducing its high-error tail. BestRand generates multiple random subsets and uses a few probe models to search for the best candidate. Our theory identifies a mechanism by which probe-guided selection can reduce common-shift variability, motivating BestRand's design. Experiments demonstrate its improved MAE stability. We validate BestRand on 70+ text, multimodal, and agent benchmarks, and further extend it

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.

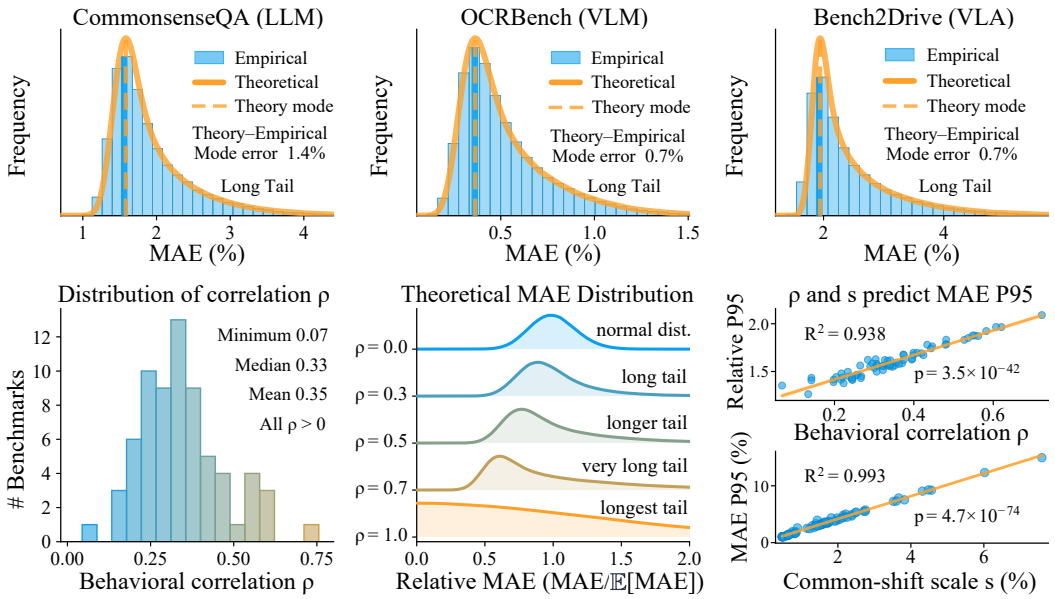

Figure 1: Empirical results are consistent with our theory. Top: empirical MAE distributions and modes closely match theoretical results on representative LLM, VLM, and VLA benchmarks. Bottom: ρ is generally positive; higher ρ theoretically yields longer tails, while ρ and s strongly predict empirical MAE P95, which captures long-tail risk. Additional validation is provided in Appendix F.

to VLA evaluation. Against competitive methods, BestRand reduces high-error risk by 50.2% and estimation error by 16.4% on average, while requiring only 51.3% of the construction cost. It further reduces OOD estimation error by 21.8% on average across score-range, model-family, and temporal shifts. Together, our theory provides a framework for understanding and guiding benchmark compression, while BestRand turns these insights into a stable and low-cost method.

We discuss related works in Appendix G and summarize our contributions as follows:

- **Theory.** We derive an accurate analytical approximation to the error distribution for benchmark compression, explaining why random sampling is effective with high probability yet remains unstable. We further identify inherent similarity among foundation models as a primary source of this instability. Extensive numerical experiments are highly consistent with our theory.
- **Method.** We propose BestRand, a stable and low-cost benchmark compression method that improves random sampling through lightweight probe-guided search, with theoretical support for improved stability. We validate BestRand on 70+ text, multimodal, and agent benchmarks, as well as VLA evaluation, showing strong accuracy, stability, and OOD generalization at low cost.

2 Theory-inspired Method: BestRand

In this section, we combine theoretical analysis with empirical evidence to explain why random sampling often achieves strong compression performance yet remains unstable. Building on these insights, we introduce BestRand to mitigate this instability while preserving the efficiency.

Preliminaries. Let $X \in [0,1]^{N\times M}$ be a fixed matrix for N items and M models, whose entry $X_{s,m}$ is the score of model m on item s. All randomness in our sampling analysis comes from drawing a subset $\mathcal{S}$ of k items uniformly without replacement, where $1 \le k \le N$. The full-benchmark score, subset estimate, and error of model m are

$$\mu_m = (1/N)\sum_{s=1}^{N} X_{s,m}, \qquad \hat{\mu}_m = (1/k)\sum_{s\in\mathcal{S}} X_{s,m}, \qquad E_m = \hat{\mu}_m - \mu_m. \tag{1}$$

Our primary quantity is the mean absolute error $\mathrm{MAE} = M^{-1}\sum_{m=1}^{M}|E_m|$ of the error vector $E = (E_1,\dots,E_M)^\top$. Subscripts $\mathcal{S}$ and N indicate subset and full-benchmark averages. We write Var for sampling variance, Var_N for empirical variance over items with denominator N, and Φ for the standard normal cumulative distribution function.

Define item difficulty as $g_s = M^{-1}\sum_{m=1}^{M} X_{s,m}$, with larger scores indicating easier items. Its full-benchmark mean and variance are $\bar{g}_N = N^{-1}\sum_{s=1}^{N} g_s$ and $\mathrm{Var}_N(g) = N^{-1}\sum_{s=1}^{N}(g_s - \bar{g}_N)^2$. The common shift $\bar{E} = M^{-1}\sum_{m=1}^{M} E_m = \bar{g}_{\mathcal{S}} - \bar{g}_N$, where $\bar{g}_{\mathcal{S}} = k^{-1}\sum_{s\in\mathcal{S}} g_s$, measures whether the subset is easier or harder on average. Its sampling standard deviation defines our central scale,

$$s := \sqrt{\mathrm{Var}(\bar{E})} = c(k)\sqrt{\mathrm{Var}_N(g)}, \qquad c(k) = \sqrt{(N-k)/[k(N-1)]}. \tag{2}$$

The finite-population factor $c(k)$ grows as fewer items are retained and satisfies $c(N) = 0$.

Assumption 1: Consider non-degenerate models whose scores vary across test items. For binary scores, this excludes models that answer all items incorrectly or all items correctly. Define the population score covariance matrix $C \in \mathbb{R}^{M\times M}$ and its relation to sampling errors by

$$C_{ml} = \frac{1}{N}\sum_{s=1}^{N}(X_{s,m} - \mu_m)(X_{s,l} - \mu_l), \qquad \mathrm{Cov}(E) = c(k)^2 C. \tag{3}$$

Assume $C_{mm} = \sigma^2 > 0$ and $C_{ml} = \rho\sigma^2$ for $m \neq l$, with $\rho \in [0, 1)$. Then the common-shift scale simplifies to $s = c(k)\,\sigma\sqrt{\rho + (1-\rho)/M}$. Empirically, σ^2 and ρ are read directly from C as the mean diagonal entry and the mean normalized off-diagonal entry, $\sigma^2 = M^{-1}\sum_m C_{mm}$ and $\rho = [M(M-1)]^{-1}\sum_{m\neq l} C_{ml}/\sigma^2$, so ρ is the average pairwise correlation of the model errors.

Assumption 2: Let $x_s = (x_{s,1}, \ldots, x_{s,p})^\top \in \mathbb{R}^p$ contain the scores of p probe models on item s, with means $\bar{x}_{\mathcal{S}} = k^{-1}\sum_{s\in\mathcal{S}} x_s$ and $\bar{x}_N = N^{-1}\sum_{s=1}^{N} x_s$. Regress difficulty on probe scores by least squares with intercept $\alpha \in \mathbb{R}$, coefficients $\beta \in \mathbb{R}^p$, and residual e_s. We assume $\mathrm{Var}_N(g) > 0$ and positive explained variance,

$$g_s = \alpha + \beta^\top x_s + e_s, \qquad R^2 := \mathrm{Var}_N(\beta^\top x)/\mathrm{Var}_N(g) > 0. \tag{4}$$

Derivations are provided in Appendix F, and empirical checks across numerous benchmarks in Appendix F.2 support these assumptions as useful approximations rather than exact descriptions of the data. The resulting theoretical predictions closely match empirical observations (e.g., Figure 1), further supporting the value of these simplifications for tractable analysis.

2.1 Why Random Sampling is almost enough for compression but unstable?

In this Section, Theorem 1 fixes the shape of the error distribution and its tail, and show why the random sampling is unstable. Theorem 2 helps explain why random sampling may be almost enough in many cases: for large M, the mode of the MAE distribution lies close to the lower endpoint of its limiting distribution. Proofs are in Appendix F.

Theorem 1. Under Assumption 1, when the effective sample size $\min\{k, N-k\}$ is large, the error vector is approximately zero-mean multivariate normal:

$$E \;\dot{\sim}\; \mathcal{N}\big(0,\, c(k)^2 C\big), \qquad \mathrm{MAE} = \frac{1}{M}\sum_{m=1}^{M}|E_m|. \tag{5}$$

Moreover, $\mathrm{MAE} \geq |\bar{E}|$ holds for every sampled subset, and the common shift $\bar{E} = M^{-1}\sum_m E_m$ is approximately $\mathcal{N}(0, s^2)$. Hence, for every threshold $t > 0$,

$$\Pr(\mathrm{MAE} > t) \;\geq\; \Pr(|\bar{E}| > t) \;\approx\; 2\,\Phi\Big(-\tfrac{t}{s}\Big), \tag{6}$$

where $s = \sqrt{\mathrm{Var}(\bar{E})}$ is the common-shift scale defined above.

Why random sampling is unstable. Instability means that a single random draw occasionally lands far from the true scores, which is exactly the right tail of the MAE distribution. Theorem 1 controls this tail through its lower bound $\Pr(\mathrm{MAE} > t) \geq 2\Phi(-t/s)$, whose only free quantity is the scale s. At any fixed threshold t, a larger s pushes $-t/s$ toward zero, and because Φ is increasing it raises the probability of a large error. A larger s therefore raises this lower bound on high-error risk, so it remains to ask what makes s large. Under Assumption 1, $s = c(k)\sigma\sqrt{\rho + (1-\rho)/M}$. For $M > 1$, the term under the square root increases with the behavioral correlation ρ, since its

derivative is $1 - 1/M > 0$. Stronger correlation therefore increases s and raises the lower bound on the probability of large errors. Increasing the number of evaluated models reduces the contribution $(1-\rho)/M$, but cannot eliminate the correlation-dependent component: for fixed $k < N$ and $\rho > 0$, $s \to c(k)\sigma\sqrt{\rho} > 0$ as $M \to \infty$. Thus, expanding the model panel alone cannot remove this source of instability. Numerical experiments across diverse benchmarks in Figure 1 provide empirical support for this analysis based on Theorem 1.

For foundation models, such behavioral correlation is closely tied to how the models are built and trained. Language, vision-language, and action models often share similar or identical backbones and overlapping training corpora, while knowledge distillation can further align their behavior. These shared design and training choices can lead models to succeed and fail on the same items, sustaining a sizable ρ. In this sense, random sampling instability is rooted in the shared characteristics of foundation models, rather than merely in the finite number of models evaluated. The non-vanishing limit above describes a consequence of this intrinsic correlation: evaluating more similarly constructed models does not average away their shared errors.

Theorem 2. Under Assumption 1 and Theorem 1 with fixed $c(k) > 0$, $\sigma > 0$, and $0 < \rho < 1$, the lower endpoint of the limiting MAE distribution as $M \to \infty$ is $y_0 = \sqrt{2/\pi}\, c(k)\, \sigma\sqrt{1-\rho}$. This is not a lower bound for finite M, where the MAE distribution can have probability mass below y_0. The mode $\hat{y}(M)$ of the MAE distribution approaches y_0 at the universal rate

$$\frac{\hat{y}(M)}{y_0} - 1 = \frac{C_0}{\sqrt{M}} + O\Big(\tfrac{1}{M}\Big), \quad C_0 = \sqrt{\tfrac{\pi}{2} - 1}\, \arg\max_{u>0} \int_0^\infty v^{-1/2} \varphi(u - v)\, \mathrm{d}v \approx 0.578, \quad (7)$$

independent of $c(k), \sigma$ and ρ, with φ the standard normal density.

Why random sampling is almost enough. Theorem 2 gives an asymptotic description of the low-error region: for large M, the mode lies near the lower endpoint y_0 of the limiting distribution, with a leading relative gap of $C_0/\sqrt{M}$. This prediction is consistent with the empirical observation in Figure 1 and more experiments are shown in Appendix F.6, where the mode lies close to the lower end of the observed MAE distribution. Together, the asymptotic analysis and the observed concentration at small errors help explain why random sampling is often already effective. The long tail nevertheless leaves a small but non-negligible fraction of high-error draws, motivating BestRand to improve stability while preserving the low cost of random sampling.

2.2 How to build more stable and low-cost benchmark compression methods.

Theorem 2 leaves a clear target. The asymptotic analysis and empirical observations suggest that random sampling already favors low-error subsets, so a good compressor should preserve its low cost while reducing the risk of high-error draws. BestRand does this by drawing several random candidate subsets and selecting the one that a few cheap probe models judge most representative. It runs the target models on the selected subset alone and reuses the probe scores across all candidates, so its only overhead beyond random sampling is a one-time probe pass over a small pool.

The BestRand procedure. Given a benchmark of N items, a target size k, p probe models, a pool multiplier n, and a candidate count m, BestRand proceeds in four steps.

1. Draw a pool P of $\min(nk, N)$ items uniformly without replacement and score the p probe models on P, recording $x_s \in \mathbb{R}^p$ for each item s.
2. Form the reference $H = |P|^{-1} \sum_{s \in P} x_s$, the average probe score over the pool.
3. Repeat m times: draw a candidate k-subset S_j from P uniformly without replacement, and compute its probe mean $h_j = k^{-1} \sum_{s \in S_j} x_s$ and its imbalance $d_j = \|h_j - H\|^2$.
4. Return $S^* = S_{j^*}$ with $j^* = \arg\min_{j \le m} d_j$, and evaluate every target model on S^* alone.

The single probe pass costs $p \times |P|$ model-item evaluations, and the pool reference H replaces the full-set probe mean so the probes need not run on the whole benchmark when $nk < N$. The m candidate draws in step 3 reuse the cached probe scores and add no model calls, computing only averages of cached vectors, so they run in parallel as a single batched tensor operation on a GPU and make the search essentially instant. The selected subset S^* is reused across all target models.

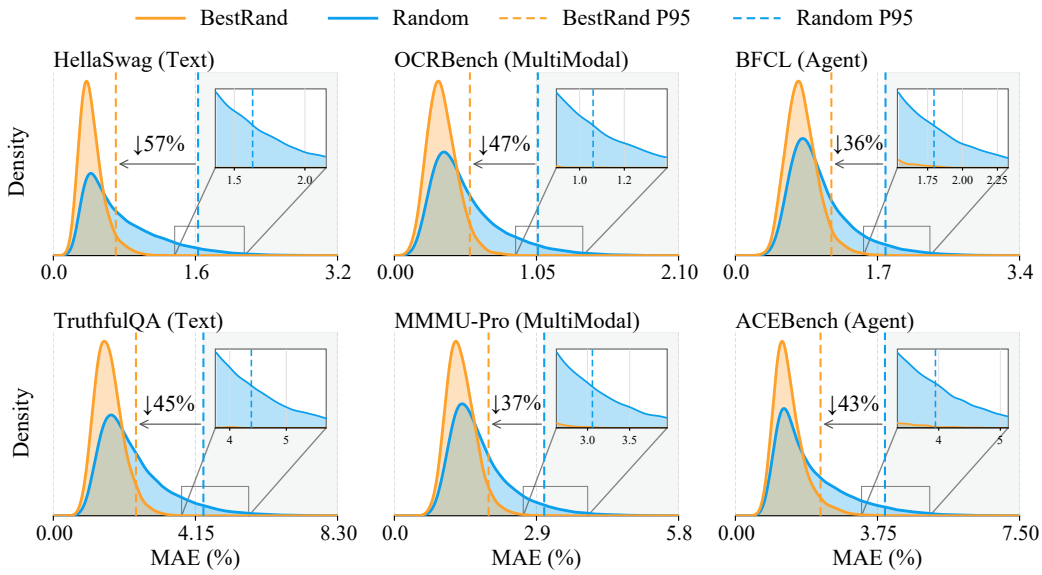

Figure 2: Error distributions of Random and BestRand on six representative benchmarks at 20% retention. Random exhibits a pronounced right tail, whereas BestRand concentrates errors near zero and suppresses high-error outcomes. Results on additional benchmarks are shown in Figure 8.

Theorem 3. Under Assumptions 1 and 2, least-squares orthogonality splits the common shift into a probe-explained part and a residual part,

$$\bar{E} = \bar{E}_{\parallel} + \bar{E}_{\perp}, \qquad \bar{E}_{\parallel} = \beta^{\top}(\bar{x}_{\mathcal{S}} - \bar{x}_N), \quad \bar{E}_{\perp} = \bar{e}_{\mathcal{S}} = \tfrac{1}{k}\sum_{s\in\mathcal{S}} e_s, \tag{8}$$

which are uncorrelated with variances $\mathrm{Var}(\bar{E}_{\parallel}) = R^2 s^2$ and $\mathrm{Var}(\bar{E}_{\perp}) = (1-R^2)s^2$, so R^2 is the fraction of the common-shift variance that the probes explain. Draw m candidates independently, each uniformly among the k-subsets of the full benchmark, and select the candidate with the smallest full-reference probe imbalance, breaking ties by a fixed index rule:

$$S^* = \arg\min_{j\le m} \left\|\bar{x}_{S_j} - \bar{x}_N\right\|, \qquad D^* = \left\|\bar{x}_{S^*} - \bar{x}_N\right\|, \tag{9}$$

Within a joint Gaussian model of probe and residual imbalances, the selected common-shift variance satisfies

$$s^2_{S^*} = \mathrm{Var}\big(\bar{E}_{\parallel}(S^*)\big) + (1-R^2)\, s^2, \qquad 0 \le \mathrm{Var}\big(\bar{E}_{\parallel}(S^*)\big) \le \|\beta\|^2\, \mathbb{E}\big[D^{*2}\big], \tag{10}$$

Thus, if $\mathbb{E}[D^{*2}] \to 0$, the common-shift scale converges to $\sqrt{1-R^2}\, s$ within this model.

Why BestRand is better than random sampling. Theorem 3 **motivates** BestRand by identifying the probe-explained component of the common shift as a target for selection. By the Cauchy–Schwarz inequality, $|\bar{E}_{\parallel}| \le \|\beta\|\, \|\bar{x}_{\mathcal{S}} - \bar{x}_N\|$. Matching the full-benchmark probe mean therefore controls this component without knowing β or evaluating the target models, motivating BestRand's probe-based selection criterion. To implement this criterion at low cost, BestRand uses the pool mean H to approximate $\bar{x}_N$ and selects the candidate that best matches H. The connection follows from $\|\bar{x}_{S^*} - \bar{x}_N\| \le \|\bar{x}_{S^*} - H\| + \|H - \bar{x}_N\|$. The first term is the imbalance minimized by the candidate search, while the second is the reference error controlled by the pool size. Choosing a sufficiently large n makes H close to $\bar{x}_N$ in mean square, with exact equality when the pool covers the full benchmark. Increasing m then improves or preserves the best match to this reference without additional model evaluations. Thus, n controls reference accuracy and m controls search quality.

This design aims to reduce the common-shift scale. Within the selection model of Theorem 3, the probe-explained variance is bounded by $\|\beta\|^2\mathbb{E}[D^{*2}]$, while the residual variance remains $(1-R^2)s^2$. If $\mathbb{E}[D^{*2}] \to 0$ and $s > 0$, then

$$s^2_{S^*} = \mathrm{Var}\big(\bar{E}_{\parallel}(S^*)\big) + (1-R^2)\, s^2 \;\longrightarrow\; (1-R^2)\, s^2 < s^2. \tag{11}$$

Table 1: Compression accuracy across 70 benchmarks at six retention ratios. Results are grouped into 44 text, 16 multimodal, and 10 agent benchmarks. Each entry reports average MAE (%) and the standard deviation across benchmarks. The final column averages all benchmarks and retention ratios. Lower is better. Best mean values are **bold** and second-best mean values are underlined.

	Average MAE (%) ↓ mean ± std.						
Method	**2%**	**5%**	**10%**	**15%**	**20%**	**50%**	**Mean**
Text 44 benchmarks							
SparseEval (Linear)	19.88 ±9.48	21.08 ±6.99	22.25 ±7.72	21.18 ±10.10	21.44 ±9.60	20.04 ±7.95	20.98 ±8.67
SparseEval (MLP)	12.53 ±5.85	11.94 ±5.16	11.70 ±5.38	11.49 ±5.32	11.74 ±5.51	11.80 ±5.80	11.87 ±5.46
Anchor Points	7.68 ±3.24	7.23 ±2.66	6.78 ±2.73	6.32 ±2.66	6.41 ±2.73	4.72 ±2.58	6.52 ±2.90
TinyBenchmarks	7.45 ±4.49	4.94 ±3.34	3.26 ±2.22	2.56 ±1.82	2.26 ±1.70	1.45 ±1.55	3.65 ±3.37
ZipBench	6.48 ±3.97	4.35 ±2.65	3.03 ±1.97	2.40 ±1.48	2.02 ±1.20	1.10 ±0.68	3.23 ±2.85
Random	7.37 ±5.18	4.60 ±3.20	3.17 ±2.19	2.51 ±1.73	2.11 ±1.45	1.05 ±0.72	3.47 ±3.46
BestRand (ours)	**6.27** ±4.08	**3.92** ±2.49	**2.67** ±1.69	**2.09** ±1.31	**1.73** ±1.07	**0.85** ±0.52	**2.92** ±2.80
Multimodal 16 benchmarks							
SparseEval (Linear)	22.59 ±10.58	21.92 ±6.65	25.21 ±14.38	25.97 ±10.38	22.76 ±6.94	21.02 ±8.38	23.25 ±9.81
SparseEval (MLP)	13.61 ±6.37	14.59 ±8.14	14.00 ±7.11	14.12 ±7.52	12.68 ±7.05	13.00 ±6.70	13.67 ±7.01
Anchor Points	8.18 ±4.13	6.83 ±2.66	6.66 ±2.23	5.84 ±2.15	6.61 ±2.46	5.07 ±2.10	6.53 ±2.81
TinyBenchmarks	7.58 ±3.31	4.21 ±2.11	2.88 ±1.54	2.27 ±1.09	2.11 ±1.01	1.36 ±0.62	3.40 ±2.74
ZipBench	**6.10** ±2.89	4.15 ±2.24	3.28 ±1.96	2.54 ±1.66	1.99 ±1.09	1.09 ±0.54	3.19 ±2.45
Random	6.89 ±3.47	4.30 ±2.15	2.94 ±1.46	2.34 ±1.16	1.96 ±0.98	0.98 ±0.49	3.24 ±2.66
BestRand (ours)	6.16 ±3.21	**3.83** ±2.00	**2.61** ±1.36	**2.06** ±1.08	**1.71** ±0.90	**0.84** ±0.45	**2.87** ±2.43
Agent 10 benchmarks							
SparseEval (Linear)	20.90 ±6.89	25.68 ±6.57	24.74 ±6.99	22.33 ±7.91	19.77 ±4.42	22.79 ±6.04	22.70 ±6.60
SparseEval (MLP)	12.32 ±3.41	12.76 ±2.74	12.65 ±3.37	11.65 ±2.72	10.66 ±2.93	10.40 ±2.55	11.74 ±2.99
Anchor Points	12.35 ±5.84	9.56 ±4.16	7.84 ±2.65	7.25 ±2.50	7.19 ±2.35	5.34 ±2.06	8.26 ±4.04
TinyBenchmarks	13.10 ±8.57	7.91 ±5.73	5.96 ±4.37	4.58 ±2.80	3.80 ±1.98	2.27 ±1.28	6.27 ±5.81
ZipBench	12.17 ±7.83	8.07 ±5.68	5.53 ±3.82	4.31 ±2.98	3.90 ±2.74	2.08 ±1.38	6.01 ±5.51
Random	13.32 ±8.77	8.22 ±5.55	5.48 ±3.56	4.38 ±2.88	3.65 ±2.36	1.82 ±1.18	6.14 ±5.91
BestRand (ours)	**11.94** ±8.25	**7.30** ±5.14	**4.82** ±3.25	**3.82** ±2.61	**3.14** ±2.13	**1.56** ±1.06	**5.43** ±5.44

A larger explained fraction R^2 yields a greater potential reduction in the shared source of high-error risk identified by Theorem 1. This limit describes the idealized benefit of probe matching, rather than a guaranteed contraction for finite-pool selection, where reference error and nonzero attainable imbalance remain. Our numerical results are consistent with this analysis. Figure 2 shows that BestRand concentrates MAE near zero and suppresses the upper tail relative to random sampling across benchmarks, supporting the practical value of probe-guided selection for improving stability.

3 Experiments

We evaluate BestRand on 70 text, multimodal, and agent benchmarks, focusing on estimation accuracy and stability. We compare average errors across methods and compression budgets, then examine error distributions and high-error risk, with VLA results reported in the Appendix C. We first introduce the benchmarks, then describe the baselines, evaluation protocol, and metrics.

Benchmarks and budgets. Our evaluation spans 70 benchmarks, comprising 44 text, 16 multimodal, and 10 agent benchmarks. These include widely used text benchmarks such as MMLU (Hendrycks et al., 2020), GSM8K (Cobbe et al., 2021), and HumanEval (Chen et al., 2021), multimodal benchmarks such as MMMU (Yue et al., 2024), MathVista (Lu et al., 2024), and MME (Fu et al., 2025a), and agent benchmarks such as OSWorld (Xie et al., 2024) and tau-bench (Yao et al., 2024). To compare methods comprehensively across compression levels, we evaluate six retention ratios ranging from 2% to 50%. Details of all benchmarks are provided in Appendix A.2.

Baselines and evaluation protocol. We compare against six representative baselines. **Random** uniformly samples a subset and averages its scores. **SparseEval** (Zhang et al., 2026) uses a linear or

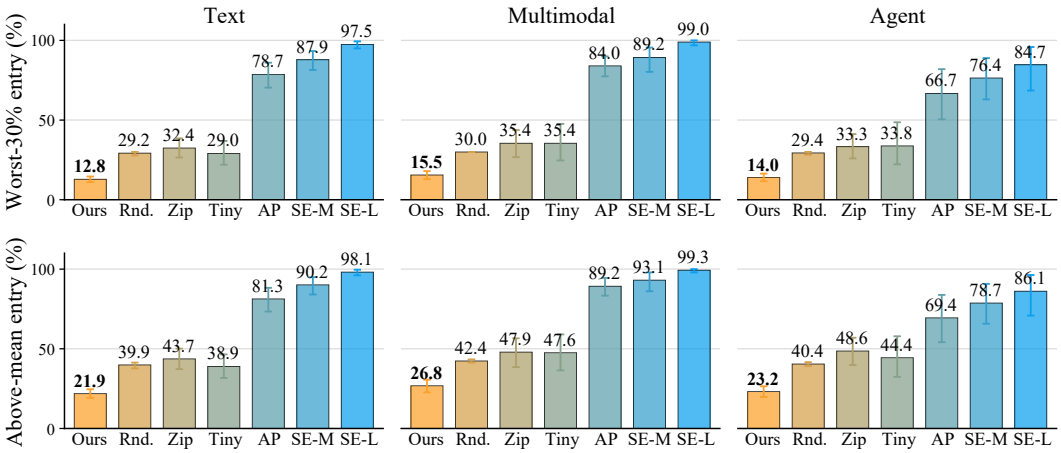

Figure 3: High-error risk across text, multimodal, and agent benchmarks. Worst-30% entry and worse-than-average entry use Random's 70th-percentile and mean errors as thresholds, respectively. Whiskers show 95% bootstrap confidence intervals. BestRand achieves the lowest risk in all settings. Rand, Zip, Tiny, AP, SE-M, and SE-L denote Random, ZipBench, TinyBenchmarks, Anchor Points, SparseEval MLP, and SparseEval Linear, respectively. Detailed results are shown in Figure 6.

MLP predictor to estimate full-benchmark scores from subset scores. **Anchor Points**, **TinyBenchmarks**, and **ZipBench** (Vivek et al., 2024; Polo et al., 2024; Huang et al., 2026) construct item representations and cluster them to select a representative subset. All methods use the same model split and item budget. Random and BestRand use 10,000 seeds, while other methods use three. BestRand uses $p = 5$, $m = 128$, $n = 4$, random probe selection, and L_2 matching. It probes only the candidate pool and thus incurs a lower construction cost. Section 4 further compares four additional baselines. Appendices A.1 and A.3 provide details on the data splits and baselines, respectively.

Evaluation metrics. We use mean absolute error (MAE) as our primary metric. We also evaluate ranking fidelity using Spearman's rank correlation coefficient (SP). Since rank correlations are generally strong for leading methods in both our experiments and prior work (Yauney et al., 2026), we focus on MAE in the main results and report detailed SP results in Appendix E, Tables 9–11.

We then compare BestRand with existing methods, showing its high accuracy, low cost, and stability.

Main results. As shown in Table 1, Random performs strongly on average. BestRand consistently improves on this baseline, reducing both mean MAE and its variation across all settings. Against three competitive baselines, namely Random, ZipBench, and TinyBenchmarks, BestRand reduces MAE by 16.4% on average, achieving the best overall performance. By probing only a candidate pool, BestRand uses just 8.0% of data-dependent baselines' construction records at 2% retention and 51.3% on average. BestRand therefore achieves the lowest error at roughly half the construction cost. Beyond average performance, we next compare the stability and high-error risk of all methods.

Stability and high-error risk. Figure 2 shows that BestRand concentrates errors near zero and suppresses Random's upper tail, yielding more stable estimates. Figure 3 quantifies the risk of falling into the high-error region of Random's long-tailed distribution. We measure the probability that each method's error exceeds Random's 70th percentile or mean, corresponding to the worst 30% of random subsets and worse-than-average random performance, respectively. Under both thresholds, BestRand achieves the lowest risk across benchmark categories and reduces high-error probability by 50.2% versus three strong baselines. Additional results are reported in Appendix E. Together, these results demonstrate that BestRand enables accurate and stable benchmark compression.

4 Analysis and Ablation Study

This section analyzes the key factors affecting BestRand and evaluates its effectiveness, efficiency, and robustness across practical settings. Additional analyses are provided in Appendix C.

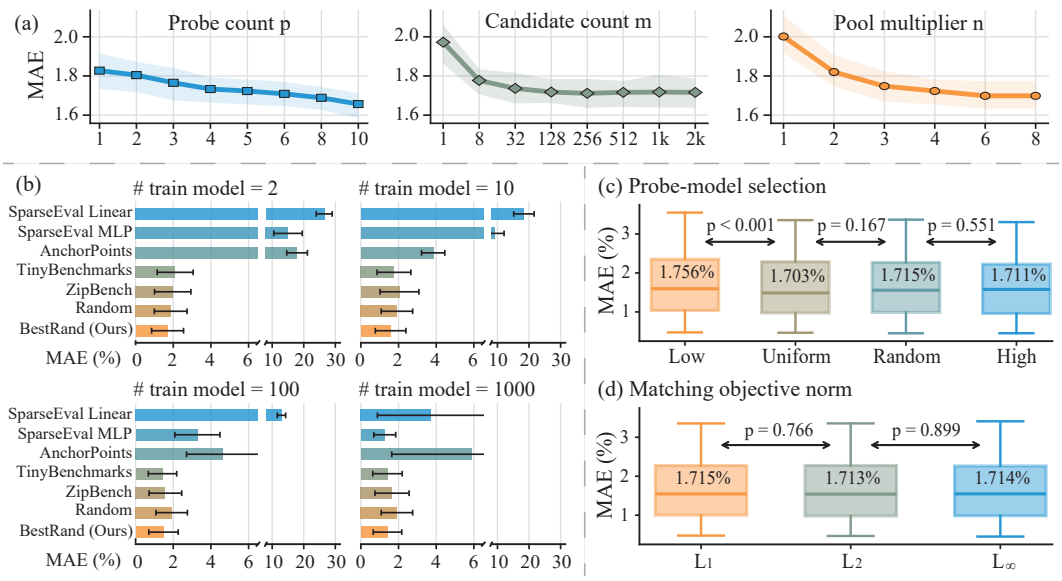

Figure 4: BestRand ablations at 20% retention. (a) Effects of the number of probe models, the number of candidate subsets, and the candidate-pool size. (b) Performance comparison with more construction models. (c) Effect of probe-model selection. (d) Effect of the matching norm.

(1) Effects of Probe Count, Candidate Count, and Pool Multiplier on Performance.

We first examine how the probe count p, candidate count m, and pool multiplier n affect compression accuracy. As shown in Figure 4(a), more probes improve the reliability of candidate comparison, more candidates increase the chance of finding a representative subset, and a larger pool provides a more accurate reference and greater candidate diversity. Performance improves with all three quantities, but the gains from increasing m and n quickly diminish. We therefore use $p = 5$, $m = 128$, and $n = 4$ by default. This setting achieves strong accuracy at low construction cost.

(2) Does BestRand Remain Competitive as More Construction Models Become Available?

Because our main experiments use only five construction models, we next test whether BestRand remains competitive when more models are available. Figure 4(b) shows that BestRand performs best with only a few construction models. Learning-based methods improve as they gain more response records, but BestRand remains comparable to the strongest alternatives even with hundreds or thousands of models. Thus, BestRand offers the greatest gains when construction data are limited while retaining competitive accuracy when they are abundant. The low-data setting is also more practical because collecting complete benchmark responses from many models is often costly.

(3) Sensitivity to Probe-Model Selection and Matching Norm.

We next assess the sensitivity of BestRand to probe selection and matching norm. We compare probes sampled uniformly across the score range, at random, or from its low and high ends. Using two-sided Wilcoxon signed-rank tests across benchmarks, Figure 4(c) shows no significant differences except between Uniform and Low, where Uniform performs better. This suggests that probe selection should avoid using only low-scoring models. Figure 4(d) shows similar results for L_1, L_2, and L_∞, with no significant differences according to the Wilcoxon signed-rank test. BestRand is therefore robust to both choices, and we use random selection with L_2 matching by default.

(4) Does BestRand Provide a Better Accuracy–Cost Trade-off than Existing Methods?

Methods differ in how many historical model responses they use. Response-free methods require no historical model-item scores but generally yield much higher MAE. Among response-based methods, the same number of construction models can also incur different costs because BestRand evaluates each model only on its candidate pool, whereas several baselines require full-benchmark responses. We therefore define normalized construction cost as $\text{Cost} = R/N$, where R is the total

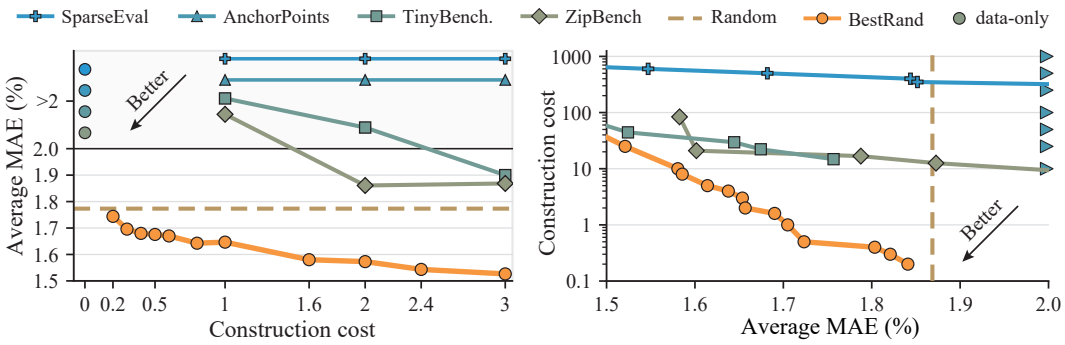

Figure 5: Low-cost compression at 20% retention. (a) At the same construction cost, BestRand achieves substantially lower average MAE across 12 text benchmarks. Response-free methods appear at Cost $= 0$ but generally have much larger errors. BestRand remains accurate at Cost < 1, while learnable baselines require Cost ≥ 1. (b) Over the displayed accuracy range, BestRand requires substantially less construction cost to reach the same average MAE on six text benchmarks.

Table 2: OOD performance under score-range, model-family, and temporal shifts. Arrows in the column headers point from the construction group to the test group. BestRand achieves the lowest average MAE in every setting. The final column averages the three OOD types equally.

	Score Range OOD			Family OOD	Temporal OOD	
Method	**Mid→Tails**	**Low→High**	**High→Low**	**Held-out Families**	**Past→Future**	**Mean**
SparseEval (Linear)	18.50 ±3.24	21.50 ±5.74	18.31 ±4.58	22.45 ±3.63	24.32 ±6.65	22.07 ±2.98
SparseEval (MLP)	16.13 ±7.04	16.43 ±7.57	21.59 ±9.64	11.87 ±5.89	11.71 ±6.05	13.88 ±6.12
Anchor Points	4.94 ±1.62	5.63 ±3.03	6.96 ±4.52	7.26 ±1.93	8.23 ±2.64	7.11 ±1.57
TinyBenchmarks	2.34 ±1.17	2.46 ±1.35	3.03 ±1.93	2.20 ±0.93	2.27 ±1.01	2.36 ±0.89
ZipBench	2.03 ±0.94	2.43 ±1.62	2.30 ±1.06	2.16 ±1.28	2.15 ±1.28	2.19 ±1.21
Random	1.81 ±0.88	2.02 ±1.06	1.78 ±0.88	1.97 ±0.95	2.01 ±1.00	1.95 ±0.95
BestRand (ours)	**1.60** ±0.83	**1.74** ±1.01	**1.61** ±0.81	**1.70** ±0.89	**1.73** ±0.91	**1.69** ±0.88

number of observed model-item scores and N is the benchmark size. Thus, Cost $= 1$ corresponds to one full evaluation of one model. Under this common budget, Figure 5 shows that BestRand achieves lower MAE at the same cost and requires less cost to achieve the same MAE.

(5) OOD Generalization across Score-Range, Model-Family, and Temporal Shifts.

Table 2 evaluates BestRand under three types of OOD shifts. Score-range OOD covers Mid→Tails, which uses middle-scoring models for construction and low- and high-scoring models for testing, as well as Low→High and High→Low. Family OOD uses disjoint model families for construction and testing. Temporal OOD uses older models for construction and newer models for testing. Methods that learn from historical performance records perform worse than Random in every setting, showing that fitting these records can hurt under distribution shifts. In contrast, BestRand achieves the lowest error in all five settings and reduces average OOD error by 21.8% relative to competitive baselines.

5 Conclusion

In this paper, we developed a theoretical framework that characterizes the error distribution of benchmark compression and provides a principled basis for understanding and predicting its accuracy and reliability. Extensive empirical results closely match our theoretical results. Guided by our theory, we introduced BestRand, which improves the stability of random sampling at low construction cost. Across 70+ text, multimodal, and agent benchmarks, together with VLA evaluation, BestRand improved accuracy and stability over strong baselines and remained robust to distribution shifts. More broadly, our findings suggest that improving benchmark compression may not require complex selection, but rather understanding when and why simple sampling fails and correcting it at low cost.

AI USE STATEMENT

AI assistants were used in this work exclusively for language polishing and code implementation. All research ideas and scientific conclusions are original contributions of the authors. The authors reviewed all LLM-assisted outputs and take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

To support reproducibility, we document all key methodological, experimental, and theoretical details. Section 2 presents the complete formulation and procedure of BestRand. Sections 3 and 4 specify the evaluation protocol, metrics, default settings, ablations, construction cost, and OOD evaluation. Appendix A gives the model splits, benchmark sources and metrics, baseline details, and hyperparameters. Appendices E and C report detailed results and describe the protocols for the VLA, multi-metric, and data-only experiments. Appendix F contains the full proofs, derivations, assumption checks, and numerical validation. We will make all code and data publicly available.

REFERENCES

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025.

Andrew M Bean, Nabeel Seedat, Shengzhuang Chen, and Jonathan Richard Schwarz. Scales++: Compute efficient evaluation subset selection with cognitive scales embeddings. *arXiv preprint arXiv:2510.26384*, 2025.

Roger Beaty, Vijeta Deshpande, Clin KY Lai, Anna Attuch, Namrata Shivagunde, Swastik Roy, Rajkumar Pujari, Paul V DiStefano, Sherin Muckatira, Claire E Stevenson, et al. Agc-bench: Measuring artificial general creativity. *arXiv preprint arXiv:2607.01152*, 2026.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Manish Bhatt, Sahana Chennabasappa, Yue Li, Cyrus Nikolaidis, Daniel Song, Shengye Wan, Faizan Ahmad, Cornelius Aschermann, Yaohui Chen, Dhaval Kapil, et al. Cyberseceval 2: A wide-ranging cybersecurity evaluation suite for large language models. *arXiv preprint arXiv:2404.13161*, 2024.

Sam Bowyer, Laurence Aitchison, and Desi R Ivanova. Position: Don't use the clt in llm evals with fewer than a few hundred datapoints. *arXiv preprint arXiv:2503.01747*, 2025.

Alex Brooker and Tim Hughes. Pre-flight: A benchmark for evaluating large language models on aviation operational knowledge. *arXiv preprint arXiv:2607.01829*, 2026.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10708–10717. IEEE, 2022.

Chen Chen, Xinlong Hao, Weiwen Liu, Xu Huang, Xingshan Zeng, Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Yuefeng Huang, et al. Acebench: Who wins the match point in tool usage? *arXiv preprint arXiv:2501.12851*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7889–7901, 2023.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

W.G. Cochran. *Sampling Techniques*. Wiley publication in applied statistics. Wiley, 1977. ISBN 9788126515240. URL https://books.google.com/books?id=xbNn41DUrNwC.

Devleena Das, Rajeev Patwari, Vikram Kumar Bukka, Nithin Kumar Guggilla, Elliott Delaye, and Ashish Sirasao. Mince: Shrinking llm evaluation datasets via few-model monte carlo calibration. *arXiv preprint arXiv:2606.22826*, 2026.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E Primack, Summer Yue, and Chen Xing. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, 2025.

Jean-Claude Deville and Yves Tillé. Efficient balanced sampling: the cube method. *Biometrika*, 91 (4):893–912, 2004.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, pp. 10323–10337. PMLR, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025a. doi: 10.52202/085713-4899. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d79a27cf2772fe00be7f341efc0eb517-Paper-Datasets_and_Benchmarks_Track.pdf.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*,

volume 38, Main Conference. Curran Associates, Inc., 2025b. doi: 10.52202/085713-3253. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/8c2e6bb15be1894b8fb4e0f9bcad1739-Paper-Datasets_and_Benchmarks_Track.pdf.

Simon Gerstenecker, Andreas Geiger, and Katrin Renz. Fail2drive: Benchmarking closed-loop driving generalization. *arXiv preprint arXiv:2604.08535*, 2026.

Jaroslav Hájek. Limiting distributions in simple random sampling from a finite population. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének közlemenyei*, 5(3):361–374, 1960.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah Smith, Hanna Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 17073–17114. Curran Associates, Inc., 2025. doi: 10.52202/085713-0576. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/18b3b2947174ae9b1a8fc91ff89f4eb6-Paper-Conference.pdf.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Zhongzhan Huang, Guoming Ling, Yupei Lin, Yandong Chen, Shanshan Zhong, Hefeng Wu, and Liang Lin. Routereval: A comprehensive benchmark for routing llms to explore model-level scaling up in llms. *arXiv preprint arXiv:2503.10657*, 2025a.

Zhongzhan Huang, Guoming Ling, Shanshan Zhong, Hefeng Wu, and Liang Lin. Minilongbench: the low-cost long context understanding benchmark for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11442–11460, 2025b.

Zhongzhan Huang, Junxin Li, Guoming Ling, Yupei Lin, Shanshan Zhong, and Hefeng Wu. Zipbench: Low-cost framework for compressing comprehensive benchmarks of large language models, 2026. URL https://arxiv.org/abs/2609.12475.

Naman Jain, Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/94074dd5a072d28ff75a76dabed43767-Paper-Conference.pdf.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Siddharth Joshi, Haoli Yin, Rishabh Adiga, Ricardo Monti, Aldo Carranza, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, et al. Datbench: Discriminative, faithful, and efficient vlm evaluations. *arXiv preprint arXiv:2601.02316*, 2026.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Yuanzhu Peter Chen, et al. Big-bench extra hard. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26473–26501, 2025.

Denica Kjorvezir, Marko Djukanović, Ana Gjorgjevikj, Gjorgjina Cenikj, and Tome Eftimov. Consistent and distinctive: Llm benchmark efficiency via maximum independent set prompt selection on similarity graphs. *arXiv preprint arXiv:2606.01400*, 2026.

Harsha Kokel, Michael Katz, Kavitha Srinivas, and Shirin Sohrabi. Acpbench hard: Unrestrained reasoning about action, change, and planning. In *International Conference on Learning Representations*, volume 2026, pp. 98938–98966, 2026.

Michael Krumdick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. Bizbench: A quantitative reasoning benchmark for business and finance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8309–8332, 2024.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 785–794, 2017.

Junlong Li, Wenshuo Zhao, Jian Zhao, Weihao Zeng, Haoze Wu, Xiaochen Wang, Rui Ge, Yuxuan Cao, Yuzhen Huang, Wei Liu, et al. The tool decathlon: Benchmarking language agents for diverse, realistic, and long-horizon task execution. In *International Conference on Learning Representations*, volume 2026, pp. 53359–53394, 2026.

Xinran Li and Peng Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769, 2017.

Xinran Li, Peng Ding, and Donald B Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162, 2018.

Yinsheng Li, Zhen Dong, and Yi Shao. Drafterbench: Benchmarking large language models for tasks automation in civil engineering. *arXiv preprint arXiv:2507.11527*, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

George Ling, Lijin Yang, Hao Yang, and Zhongzhan Huang. Blue: Toward better language use in efficient vision-language-action models for autonomous driving. *arXiv preprint arXiv:2606.08684*, 2026.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? *arXiv preprint arXiv:2404.05955*, 2024a.

Tianyang Liu, Canwen Xu, and Julian McAuley. Repobench: Benchmarking repository-level code auto-completion systems. In *International Conference on Learning Representations*, volume 2024, pp. 47832–47850, 2024b.

Yifan Liu, Wenkuan Zhao, Shanshan Zhong, Jinghui Qin, Mingfu Liang, Zhongzhan Huang, and Wushao Wen. Associam: A benchmark for evaluating association thinking while circumventing ambiguity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5203–5219, 2025.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

MAA. MAA Invitational Competitions. https://maa.org/maa-invitational-competitions/, 2026. Mathematical Association of America.

Mike Merrill, Alexander Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Shin, Thomas Walshe, E Kelly Buchanan, et al. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. In *International Conference on Learning Representations*, volume 2026, pp. 40903–40986, 2026.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 94743–94765, 2025.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Kari Lock Morgan and Donald B. Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2), April 2012. ISSN 0090-5364. doi: 10.1214/12-aos1008. URL http://dx.doi.org/10.1214/12-AOS1008.

Nexusflow.ai. Nexusbench: Fc and agent benchmarking suite, 2024. URL https://github.com/nexusflowai/NexusBench.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24838–24848. IEEE, 2025.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 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)*, pp. 2519–2536, 2024.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Antonio Purificato, Maria Sofia Bucarelli, Andrea Bacciu, Amin Mantrach, and Fabrizio Silvestri. Select, label, evaluate: Active testing in nlp. *arXiv preprint arXiv:2603.21840*, 2026.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. Infobench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13025–13048, 2024.

Nils Reimers and Iryna Gurevych. 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)*, pp. 3982–3992, 2019.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Gayathri Saranathan, Cong Xu, Mahammad Parwez Alam, Tarun Kumar, Martin Foltin, Soon Yee Wong, and Suparna Bhattacharya. Sublime: Subset selection via rank correlation prediction for data-efficient llm evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30572–30593, 2025.

Tommy Sha and Stella Zhao. Benchscope: How many independent signals does your benchmark provide? *arXiv preprint arXiv:2603.29357*, 2026.

Wentang Song, Jinqiang Li, Kele Huang, Junhui Lin, Shengxiang Wu, and Zhongshi Xie. Less is more: Not all samples are effective for evaluation. *arXiv preprint arXiv:2601.03272*, 2025a.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025b.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. In *International Conference on Learning Representations*, volume 2024, pp. 14670–14728, 2024.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Investigating prior knowledge for challenging chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155, 2020.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations*, volume 2025, pp. 63277–63303, 2025.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

Minyang Tian, Luyu Gao, Shizhuo D Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. Scicode: A research coding benchmark curated by scientists. *Advances in Neural Information Processing Systems*, 37:30624–30650, 2024.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024.

Sang Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. *arXiv preprint arXiv:2503.13335*, 2025.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36:38975–38987, 2023.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1601, 2024.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, et al. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*, 2023.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024a.

Shaobo Wang, Cong Wang, Wenjie Fu, Yue Min, Mingquan Feng, Isabel Guan, Xuming Hu, Conghui He, Cunxiang Wang, Kexin Yang, et al. Rethinking llm evaluation: Can we evaluate llms with 200× less data? In *International Conference on Learning Representations*, volume 2026, pp. 109852–109870, 2026.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024b.

Yuhao Wang and Xinran Li. Asymptotic theory of the best-choice rerandomization using the mahalanobis distance. *Journal of Econometrics*, 251:106049, 2025.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Xiaodong Wu, Minhao Wang, Yichen Liu, Xiaoming Shi, He Yan, Lu Xiangju, Junmin Zhu, and Wei Zhang. Lifbench: Evaluating the instruction following performance and stability of large language models in long-context scenarios. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16445–16468, 2025.

xAI. Grok-1.5 vision preview. `https://x.ai/news/grok-1.5v`, 2024.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Jihan Yao, Gantavya Bhatt, Arnav Das, Peter Jin, Ke Bao, Qiaolin Yu, Khushi Bhardwaj, Chang Su, Jialei Wang, Yikai Zhu, et al. Coresets before score sets: Evaluation-unsupervised prompt subset selection for llm benchmarks. *arXiv preprint arXiv:2607.09739*, 2026.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Gregory Yauney, Shahzaib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations*, volume 2026, pp. 27389–27412, 2026.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets. In *International Conference on Learning Representations*, volume 2024, pp. 55361–55414, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 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*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 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.

Guanhua Zhang, Florian E. Dorner, and Moritz Hardt. How benchmark prediction from fewer data misses the mark. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 135638–135678. Curran Associates, Inc., 2025a. doi: 10.52202/085713-4528. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c57bacb5ea621582235c54ce2a83136d-Paper-Conference.pdf.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Taolin Zhang, Hang Guo, Wang Lu, Tao Dai, Shu-Tao Xia, and Jindong Wang. Sparseeval: Efficient evaluation of large language models by sparse optimization. *arXiv preprint arXiv:2602.07909*, 2026.

Yuan Zhang, Jason Baldridge, and Luheng He. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1298–1308, 2019.

Zicheng Zhang, Xiangyu Zhao, Xinyu Fang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Haodong Duan, Kai Chen, and Guangtao Zhai. Redundancy principles for mllms benchmarks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12492–12504, 2025b.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020.

Lin Zhao, Tianchen Zhao, Zinan Lin, Xuefei Ning, Guohao Dai, Huazhong Yang, and Yu Wang. Flasheval: Towards fast and accurate evaluation of text-to-image diffusion generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16122–16131. IEEE, 2024.

Lucen Zhong, Zhengxiao Du, Xiaohan Zhang, Haiyi Hu, and Jie Tang. Complexfuncbench: exploring multi-step and constrained function calling under long-context scenario. *arXiv preprint arXiv:2501.10132*, 2025.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13246–13257. IEEE, 2024.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, et al. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. In *International Conference on Learning Representations*, volume 2025, pp. 66602–66656, 2025.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 48337–48383, 2025.

APPENDIX CONTENTS

A Additional Experiment Details **20**

A.1 Details of Data Splits 20

A.2 Details of Benchmarks Considered 20

A.3 Details of Baselines Considered 22

A.4 Implementation and Hyperparameters 22

B Additional Visualizations **24**

C Additional Analysis **28**

C.1 Effectiveness of BestRand on VLA Benchmarks 28

C.2 Effectiveness of BestRand on Multi-Metric Benchmarks 28

C.3 Can Data-Only Methods Achieve Effective Benchmark Compression? 29

D Additional Statements **30**

D.1 Limitations of BestRand 30

D.2 Broader Impact of BestRand 30

D.3 Computational Resources Statement 30

E Additional Experimental Results **31**

E.1 Benchmark-Wise Results for Spearman Rank Correlation 31

E.2 Benchmark-Wise Detailed Results for MAE 31

E.3 Comparison with a Broader Set of Baselines 38

F Full Proofs and Derivations **41**

F.1 Notation and basic conventions 41

F.2 Numerical validation of the assumptions 41

F.3 Basic lemmas: moments of sampling without replacement 44

F.4 Proof of Theorem 1 45

F.5 Closed forms under Assumption 1 50

F.6 Proof of Theorem 2 52

F.7 Proof of Theorem 3 56

G Related Work **63**

G.1 LLM, VLM, and VLA Benchmarks 63

G.2 Low-Cost Deep Learning 63

G.3 Benchmark Compression Methods 63

H The Novelty and Contribution of BestRand **64**

Table 3: Details of the 16 multimodal and 10 agent benchmarks used in the main experiments and the two VLA benchmarks used in the extension. Metric (Single?) reports the metric used in our experiments and whether it is the benchmark's only official metric.

Type	Benchmark	Public Link	# Items	# Test Models	Metric (Single?)
Multimodal Benchmarks	CountBenchQA (Beyer et al., 2024)	HF/CountBenchQA	491	11	Accuracy (✓)
	DynaMath (Zou et al., 2025)	HF/DynaMath_Sample	5,010	20	Accuracy (✗)
	LogicVista (Xiao et al., 2024)	GH/LogicVista	448	20	Accuracy (✓)
	MathVerse (Zhang et al., 2024)	HF/MathVerse	788	20	Accuracy (✓)
	MathVista (Lu et al., 2024)	HF/MathVista	1,000	10	Accuracy (✓)
	MME (Fu et al., 2025a)	GH/MME	2,374	16	MME Score (✓)
	MMMU (Yue et al., 2024)	HF/MMMU	1,050	8	Accuracy (✓)
	MMMU-Pro (Yue et al., 2025)	HF/MMMU_Pro	1,730	20	Accuracy (✓)
	MMStar (Chen et al., 2024)	HF/MMStar	1,500	7	Accuracy (✓)
	MMVP (Tong et al., 2024)	HF/MMVP	300	7	Accuracy (✓)
	OCRBench v2 (Fu et al., 2025b)	HF/OCRBenchv2	10,000	8	OCR Score (✓)
	OmniDocBench (Ouyang et al., 2025)	GH/OmniDocBench@v1.5	1,355	6	NED (✗)
	RealWorldQA (xAI, 2024)	HF/RealworldQA	765	7	Accuracy (✓)
	SpatialEval (Wang et al., 2024a)	HF/SpatialEval	4,635	9	Accuracy (✓)
	VisualPuzzles (Song et al., 2025b)	HF/VisualPuzzles	1,168	5	Exact Match (✓)
	VisualWebBench (Liu et al., 2024a)	GH/VisualWebBench	1,534	5	Accuracy (✗)
Agent Benchmarks	ACEBench (Chen et al., 2025)	GH/ACEBench	1,023	20	Score (✗)
	BFCL (Patil et al., 2025)	GH/BFCL	5,106	16	Accuracy (✗)
	ComplexFuncBench (Zhong et al., 2025)	HF/ComplexFuncBench	1,000	20	Success Rate (✗)
	DrafterBench (Li et al., 2025)	HF/DrafterBench	1,920	20	Score (✗)
	MultiChallenge (Deshpande et al., 2025)	HF/MultiChallenge	273	20	Accuracy (✓)
	NexusBench (Nexusflow.ai, 2024)	GH/NexusBench	1,118	14	Accuracy (✓)
	OSWorld (Xie et al., 2024)	GH/OSWorld	361	20	Success Rate (✓)
	tau-bench (Yao et al., 2024)	GH/tau-bench	165	20	Pass Rate (✗)
	Terminal-Bench (Merrill et al., 2026)	HF/terminal-bench-2.0	89	20	Reward (✓)
	Toolathlon (Li et al., 2026)	GH/Toolathlon	108	20	Success Rate (✓)
VLA	Bench2Drive (Jia et al., 2024)	GH/Bench2Drive	220	50	Success Rate (✗)
	Fail2Drive (Gerstenecker et al., 2026)	GH/Fail2Drive	200	37	Success Rate (✗)

A Additional Experiment Details

A.1 Details of Data Splits

We strictly separate construction and test models to evaluate whether a compressed benchmark transfers to unseen models. For each benchmark, we select five construction models and reserve a disjoint set of models for testing. At each retention ratio, all methods use the same construction and test split. Methods may use only responses from the construction models for item selection or estimator fitting, while the test models are used exclusively for evaluation. Response-free methods use no model responses during construction and are evaluated on the same test models.

A.2 Details of Benchmarks Considered

Section 3 evaluates BestRand on 70 benchmarks spanning text, multimodal, and agent tasks, while Appendix C.1 extends the evaluation to two VLA benchmarks. Tables 4 and 3 list all benchmarks and report their names, references, public links, numbers of items and test models, the metrics used in our experiments, and whether they report multiple metrics. For benchmarks with multiple metrics, we select one representative metric for the main experiments. Appendix C.2 further evaluates BestRand in multi-metric settings and discusses its extension to jointly consider multiple metrics.

Table 4: Details of the 44 text benchmarks used in the main experiments. Each benchmark uses five construction models and the remaining models for testing. Metric (Single?) reports the metric used in our experiments and whether it is the benchmark's only official metric.

Type	Benchmark	Public Link	# Items	# Test Models	Metric (Single?)
Text Benchmarks	ACPBench Hard (Kokel et al., 2026)	HF/acp_bench	1,040	20	Accuracy (✓)
	AGC-Bench (Beaty et al., 2026)	HF/agcbench-2026	100	20	Accuracy (✗)
	AIME (MAA, 2026)	↗ MAA/AIME	60	20	Accuracy (✓)
	ARC-Challenge (Clark et al., 2018)	HF/ai2_arc	1,172	20	Accuracy (✓)
	ARC-Easy (Clark et al., 2018)	HF/ai2_arc	2,376	20	Accuracy (✓)
	BBEH (Kazemi et al., 2025)	GH/bbeh	4,520	20	Accuracy (✓)
	BBH (Suzgun et al., 2023)	GH/BIG-Bench-Hard	5,761	20	Accuracy (✓)
	BEIR NFCorpus (Thakur et al., 2021)	HF/nfcorpus	323	17	nDCG@10 (✗)
	BigCodeBench (Zhuo et al., 2025)	HF/bigcodebench	1,140	20	Pass@1 (✓)
	BoolQ (Clark et al., 2019)	HF/boolq	3,270	20	Accuracy (✓)
	C3 (Sun et al., 2020)	GH/CLUE	1,825	20	Accuracy (✓)
	CommonsenseQA (Talmor et al., 2019)	HF/commonsense_qa	1,221	20	Accuracy (✓)
	CyberSecEval 2 (Bhatt et al., 2024)	GH/CybersecurityBenchmarks	1,336	5	Success Rate (✗)
	Finance Fund. (Krumdick et al., 2024)	HF/FinanceFundamentals	600	20	Accuracy (✓)
	FLASK (Ye et al., 2024)	GH/FLASK	1,740	5	Score (✗)
	GPQA (Rein et al., 2023)	HF/gpqa	546	20	Accuracy (✓)
	GSM-Symbolic (Mirzadeh et al., 2025)	GH/ml-gsm-symbolic	7,500	20	Accuracy (✓)
	GSM8K (Cobbe et al., 2021)	HF/gsm8k	1,319	20	Accuracy (✓)
	HellaSwag (Zellers et al., 2019)	HF/hellaswag	10,042	20	Accuracy (✓)
	HumanEval (Chen et al., 2021)	GH/human-eval	164	20	Pass@1 (✓)
	IFEval (Zhou et al., 2023)	HF/IFEval	541	20	Accuracy (✗)
	InFoBench (Qin et al., 2024)	HF/InFoBench	500	14	DRFR (✓)
	JudgeBench (Tan et al., 2025)	HF/JudgeBench	350	20	Judge Accuracy (✓)
	LIFBench (Wu et al., 2025)	GH/LIFBench-2024	2,802	15	Total Score (✗)
	LiveCodeBench (Jain et al., 2025)	HF/code_generation_lite	880	17	Pass@1 (✓)
	LogiQA (Liu et al., 2020)	GH/LogiQA-dataset	1,302	20	Exact Match (✓)
	LongBench v2 (Bai et al., 2025)	HF/LongBench-v2	503	12	Accuracy (✓)
	MATH lv5 (Hendrycks et al., 2021)	HF/hendrycks_math	1,324	20	Accuracy (✓)
	MedQA (Jin et al., 2021)	GH/MedQA	1,273	20	Accuracy (✓)
	MMLU (Hendrycks et al., 2020)	HF/mmlu	14,042	20	Accuracy (✓)
	MMLU-Pro (Wang et al., 2024b)	HF/MMLU-Pro	12,032	20	Accuracy (✓)
	MuSR (Sprague et al., 2024)	HF/MuSR	756	20	Accuracy (✓)
	OpenBookQA (Mihaylov et al., 2018)	HF/openbookqa	500	20	Accuracy (✓)
	PAWS (Zhang et al., 2019)	HF/paws	8,000	20	Accuracy (✓)
	PlanBench (Valmeekam et al., 2023)	GH/LLMs-Planning	500	17	Accuracy (✓)
	Pre-Flight (Brooker & Hughes, 2026)	GH/pre_flight	300	20	Accuracy (✓)
	RACE-H (Lai et al., 2017)	HF/race	3,498	20	Accuracy (✓)
	RepoBench (Liu et al., 2024b)	GH/repobench	937	7	CodeBLEU (✗)
	SciBench (Wang et al., 2023)	GH/scibench	583	19	Accuracy (✓)
	SciCode (Tian et al., 2024)	GH/SciCode	288	16	Accuracy (✓)
	SimpleQA (Wei et al., 2024)	GH/simple-evals	4,326	13	Correctness (✗)
	TheoremQA (Chen et al., 2023)	HF/TheoremQA	800	20	Accuracy (✓)
	TruthfulQA (Lin et al., 2022)	GH/TruthfulQA	817	20	Truthfulness (✗)
	WinoGrande (Sakaguchi et al., 2021)	HF/winogrande	1,267	20	Accuracy (✓)

In Section 4, we examine how performance changes as more construction models become available, using evaluation records from up to 1,000 models on ARC-Challenge, BBH, GPQA, GSM8K, TruthfulQA, HellaSwag, IFEval, MATH, MMLU-Pro, MMLU, MuSR, and WinoGrande provided by Huang et al. (2025a).

A.3 Details of Baselines Considered

We group the baselines by whether benchmark construction uses historical model responses. Response-based methods use these records to identify representative items and, when needed, learn how to estimate full-benchmark scores from selected-item responses. Random and the data-only methods use no historical performance records and are collectively termed response-free methods.

Response-based baselines differ in their item representations and score estimators. **SparseEval** (Zhang et al., 2026) represents each item by its response vector across construction models. Starting from a clustered or random subset, it iteratively replaces items to reduce full-score prediction error on the construction models. A linear model or MLP then maps a new model's selected-item responses to its estimated full score. **Anchor Points** (Vivek et al., 2024) derives item similarities from correlations between response patterns and applies k-medoids clustering. It selects each cluster medoid and estimates performance using a cluster-size-weighted average of the selected-item scores. **TinyBenchmarks** (Polo et al., 2024) fits an item response theory model to learn item difficulty and discrimination, clusters the resulting item representations, and selects one representative per cluster. For a new model, it infers ability from the selected responses and predicts performance on unobserved items. We use its gp-IRT estimator, which combines these predictions with the weighted observed-item score. **ZipBench** (Huang et al., 2026) is designed for settings with limited construction budgets. It constructs synthetic response records by majority voting within a group of construction models and retains records spanning the observed performance range. It then fits an item response theory model to the augmented records, clusters the learned item representations, and estimates the full score using a cluster-size-weighted average of the selected-item scores.

Among the response-free baselines, **Random** uniformly samples a subset of items, whereas the four data-only methods select subsets using prompt-derived representations. All five methods estimate each model's full-benchmark score by averaging its scores on the selected items. Facility Location, MIS, and Coverage use Sentence-BERT embeddings (Reimers & Gurevych, 2019), whereas Agreement uses multilingual BERT representations (Devlin et al., 2019). **Facility Location** (Yao et al., 2026) greedily selects items that maximize semantic coverage, ensuring that each benchmark item is close to a selected item. **MIS** (Kjorvezir et al., 2026) connects semantically similar prompts in a graph and greedily selects an independent set to reduce redundancy. **Agreement** (Purificato et al., 2026) applies an eight-head attention module to the BERT representations, ranks prompts by the variance of their attention weights, and retains those with the greatest disagreement. **Coverage** (Purificato et al., 2026) begins with a random prompt and repeatedly selects the prompt farthest from the most recently chosen one, encouraging coverage of distinct regions in the embedding space.

A.4 Implementation and Hyperparameters

Main experiment. BestRand uses five probe models, 128 candidate subsets, a pool multiplier of four, and L_2 matching. We evaluate Random and BestRand over 10,000 trials to obtain stable estimates. We evaluate other baselines with three seeds to obtain reliable results. Anchor Points uses FasterPAM for k-medoids clustering. TinyBenchmarks and ZipBench use multidimensional two-parameter logistic IRT models with 10 and 5 dimensions, respectively. Both IRT models are trained for 2,000 epochs with an initial learning rate of 0.1 and a decay factor of 0.9999. ZipBench adds six synthetic response profiles. SparseEval performs 10 subset-refinement rounds and trains each proxy for 2,000 epochs. Its MLP has hidden widths of 128, 64, and 32 with dropout 0.1. Both the linear and MLP variants use a learning rate of 5×10^{-4} and train the final predictor for up to 20,000 epochs, with early stopping checked every 500 epochs using a patience of 30.

Data-only experiment. We evaluate data-only methods in Section 4 and Appendix C.3 as response-free baselines. We evaluate the four methods at 20% retention on the 12 text benchmarks used in the cost analysis. The primary setting uses 384-dimensional Sentence-BERT representations, while RoBERTa, RoBERTa's last-layer representation, and Longformer are used only for embedding sensitivity. Agreement uses `bert-base-multilingual-uncased`, a maximum sequence length of 64, eight attention heads, and a batch size of 32. MIS estimates its similarity threshold from 10,000 randomly sampled prompt pairs and searches percentiles 80 through 100 until the item budget can be met. Facility Location chooses its neighborhood scale from $\{0.1, 0.5, 1, 5, 10, 100\}$ by three-fold validation based only on geometric coverage.

Table 5: Comparison of the construction information required by BestRand and 15 diverse baselines. Text and Embed. denote direct prompt features and pretrained text embeddings. Scores denotes historical scores for individual models and items, All items indicates whether these scores must cover the entire benchmark, and Full score denotes the corresponding full benchmark aggregate. Among response-based methods, BestRand has lower information requirements because it uses scores from only a subset of items, reducing construction cost while retaining strong empirical performance.

		Construction Information				
Type	**Method**	**Text**	**Embed.**	**Scores**	**All items**	**Full score**
Response-free	Random	—	—	—	—	—
	Facility Location (Yao et al., 2026)	—	✓	—	—	—
	MIS (Kjorvezir et al., 2026)	—	✓	—	—	—
	Agreement (Purificato et al., 2026)	—	✓	—	—	—
	Coverage (Purificato et al., 2026)	—	✓	—	—	—
Response-based	SparseEval (Linear) (Zhang et al., 2026)	—	—	✓	✓	✓
	SparseEval (MLP) (Zhang et al., 2026)	—	—	✓	✓	✓
	Anchor Points (Vivek et al., 2024)	—	—	✓	✓	—
	TinyBenchmarks (Polo et al., 2024)	—	—	✓	✓	—
	ZipBench (Huang et al., 2026)	—	—	✓	✓	✓
	SubLIME (Saranathan et al., 2025)	✓	✓	✓	✓	✓
	MINCE (Das et al., 2026)	—	—	✓	✓	✓
	EssenceBench (Wang et al., 2026)	—	✓	✓	✓	✓
	FlashEval (Zhao et al., 2024)	—	—	✓	✓	✓
	DatBench (Joshi et al., 2026)	—	—	✓	✓	✓
	BestRand (ours)	—	—	✓	—	—

OOD experiment. In Section 4, we evaluate OOD generalization at 20% retention under score-range, model-family, and temporal shifts. For score-range OOD, we rank models by their full-benchmark scores, randomly sample construction models from the lowest, highest, or middle 75%, and reserve the complementary 25% for testing. Full scores are used only to define the OOD splits, and no test-model responses are used for item selection or estimator fitting. Model-family OOD uses disjoint families for construction and testing, while temporal OOD uses only models released earlier than every test model for construction. We use five splits with different random seeds for each shift setting to reduce sensitivity to any single model split. For each split, we evaluate Random and BestRand with 100 seeds and all other baselines with three seeds. We average MAE over seeds and splits within each benchmark, then report the mean and standard deviation across benchmarks.

Additional baseline experiment. In Appendix E.3, we evaluate five additional baselines on 12 benchmarks using three random seeds for each stochastic method and a single run for the deterministic DatBench. SubLIME uses BERT representations and a rank correlation predictor to select from 1,000 candidate subsets after transport filtering. MINCE estimates score drift from 10,000 Monte Carlo subsets per retention ratio. EssenceBench uses BGE-M3 representations, redundancy thresholds of 0.9, three genetic search rounds, and a regularized predictor with 25 spline terms. FlashEval retains the top 100,000 of three million candidates according to Kendall's τ_b. DatBench ranks binary items by point-biserial correlation and uses Pearson correlation for nonbinary scores. All methods use identical data splits and item budgets, and test responses are used only for final evaluation.

Construction information requirements. Table 5 summarizes the construction information required by BestRand and 15 baselines. Among response-based methods, BestRand has relatively low information requirements because its probe models score only a candidate pool rather than the full benchmark. This partial coverage reduces the required construction data and construction cost.

B ADDITIONAL VISUALIZATIONS

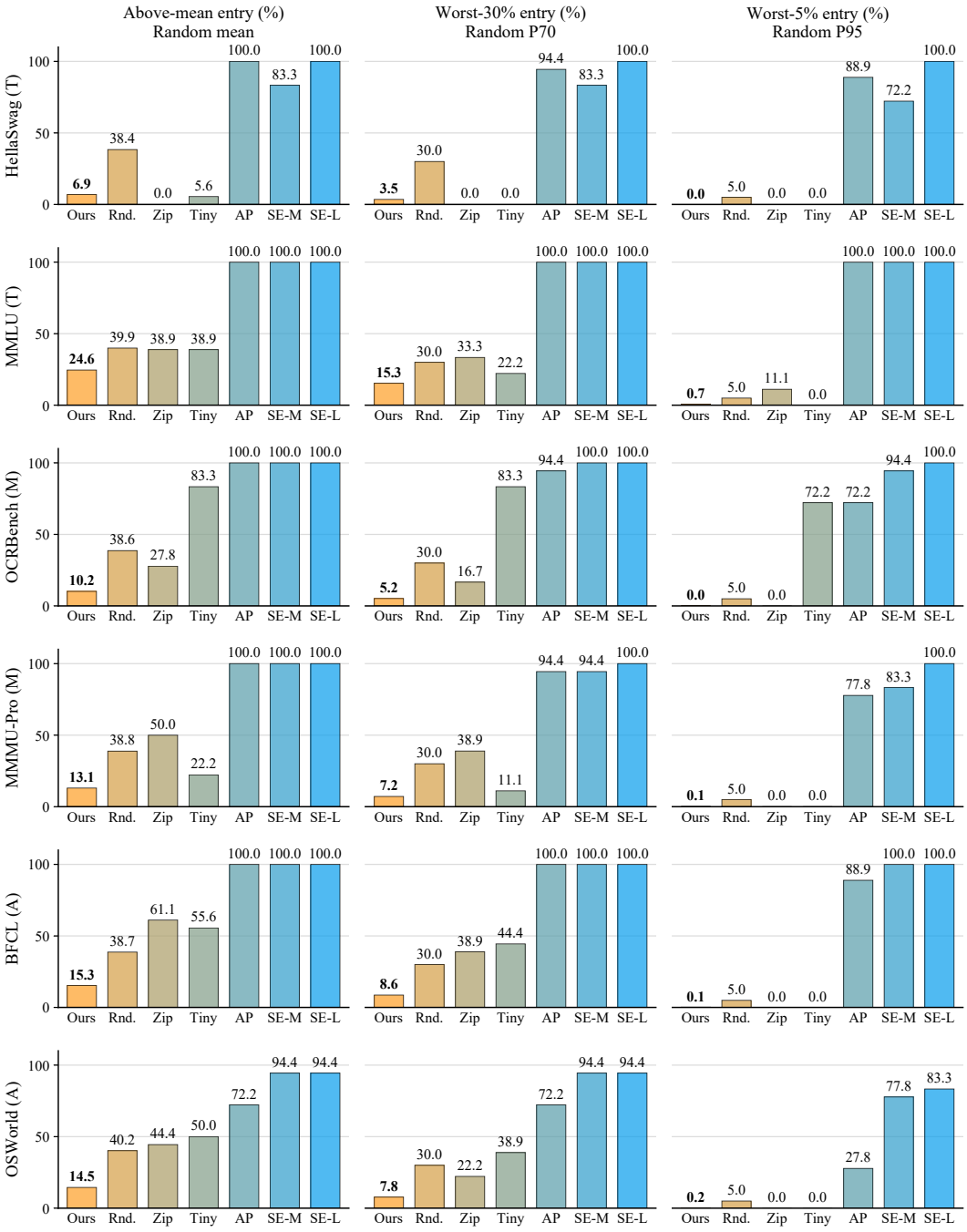

Figure 6: High-error risk across six representative text, multimodal, and agent benchmarks. Labels T, M, and A denote the text, multimodal, and agent benchmark categories, respectively. Columns use Random's empirical mean, 70th percentile, and 95th percentile MAE as thresholds for the worse-than-average, worst 30% tail, and worst 5% tail region. Results are averaged equally across six retention ratios. Random and BestRand use 10,000 trials per ratio, while the other baselines use three trials. Rand, Zip, Tiny, AP, SE-M, and SE-L denote Random, ZipBench, TinyBenchmarks, Anchor Points, SparseEval MLP, and SparseEval Linear, respectively.

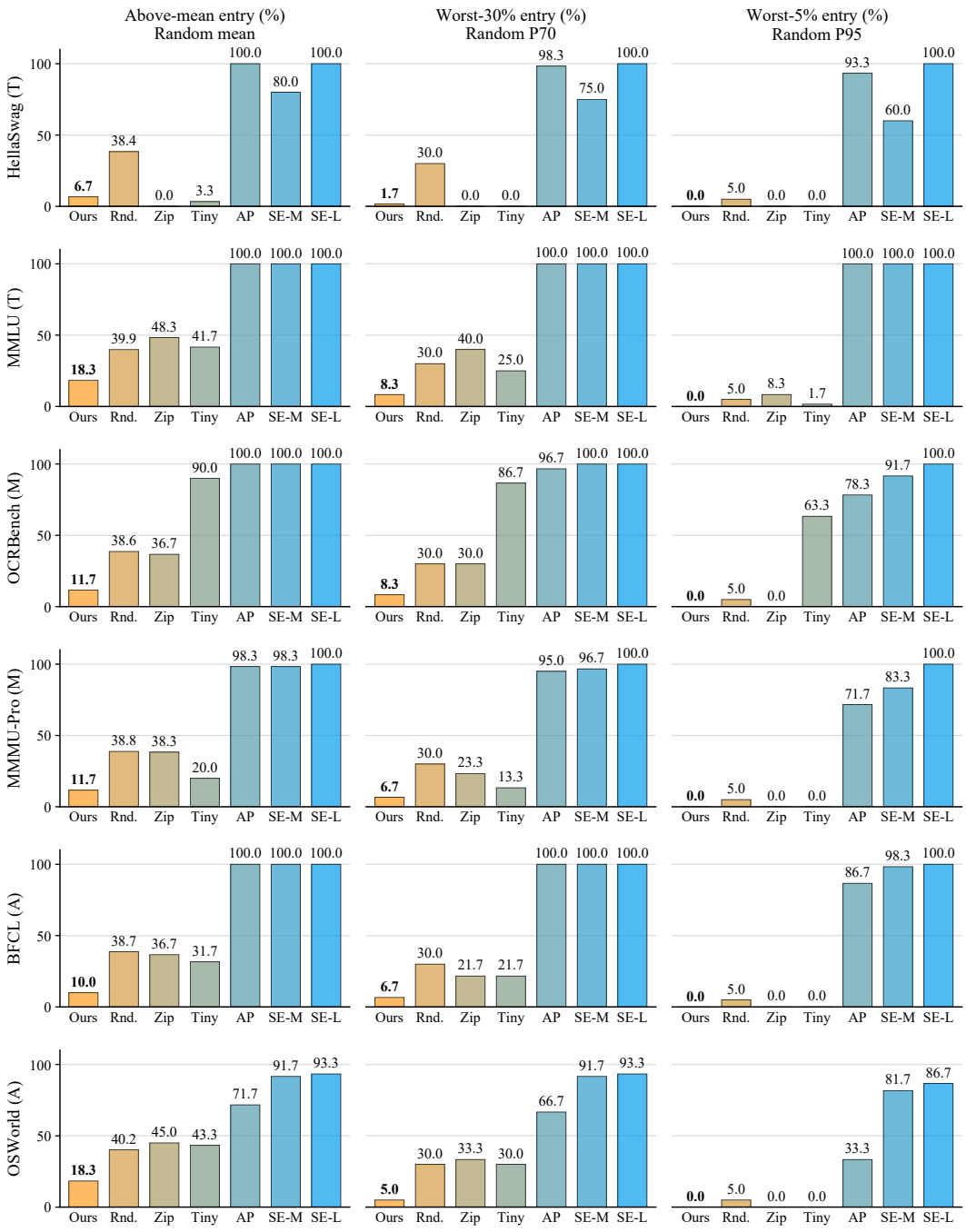

Figure 7: We conduct an additional experiment on the same six representative benchmarks as in Figure 6. Random uses the same 10,000 trials, while BestRand and all other methods are evaluated with ten seeds per retention ratio. The high error risk patterns remain similar across the two experiments. Averaged over the worse-than-average and worst-30% metrics, BestRand reduces high error risk relative to Random, ZipBench, and TinyBenchmarks by 68.06% on the six benchmarks in Figure 6 and by 71.06% in this new experiment. This close agreement confirms that the main conclusion is robust to the number of seeds and that BestRand consistently reduces high error risk.

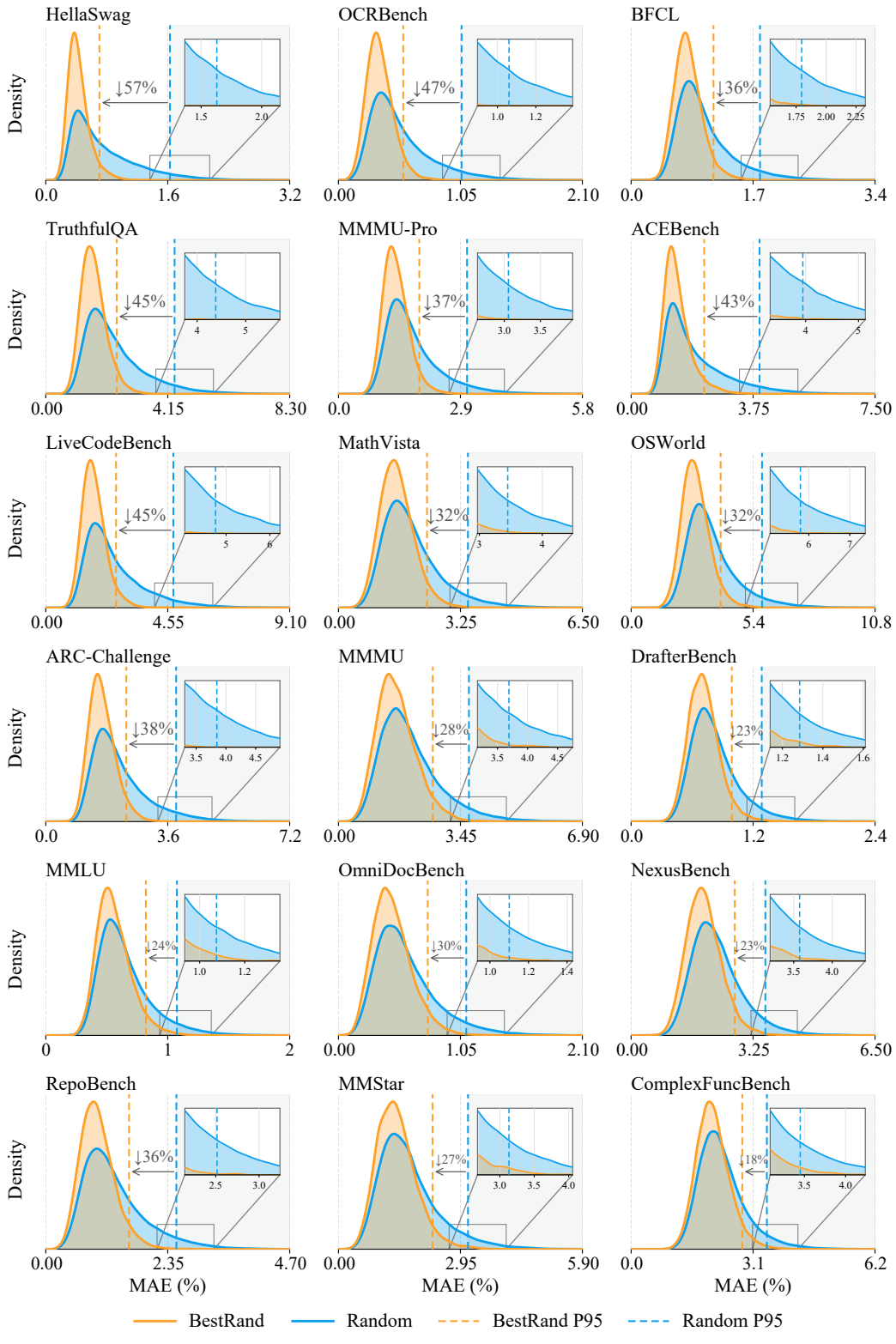

Figure 8: MAE distributions of Random and BestRand across 18 text, multimodal, and agent benchmarks. BestRand consistently concentrates errors near zero and suppresses high-error outcomes.

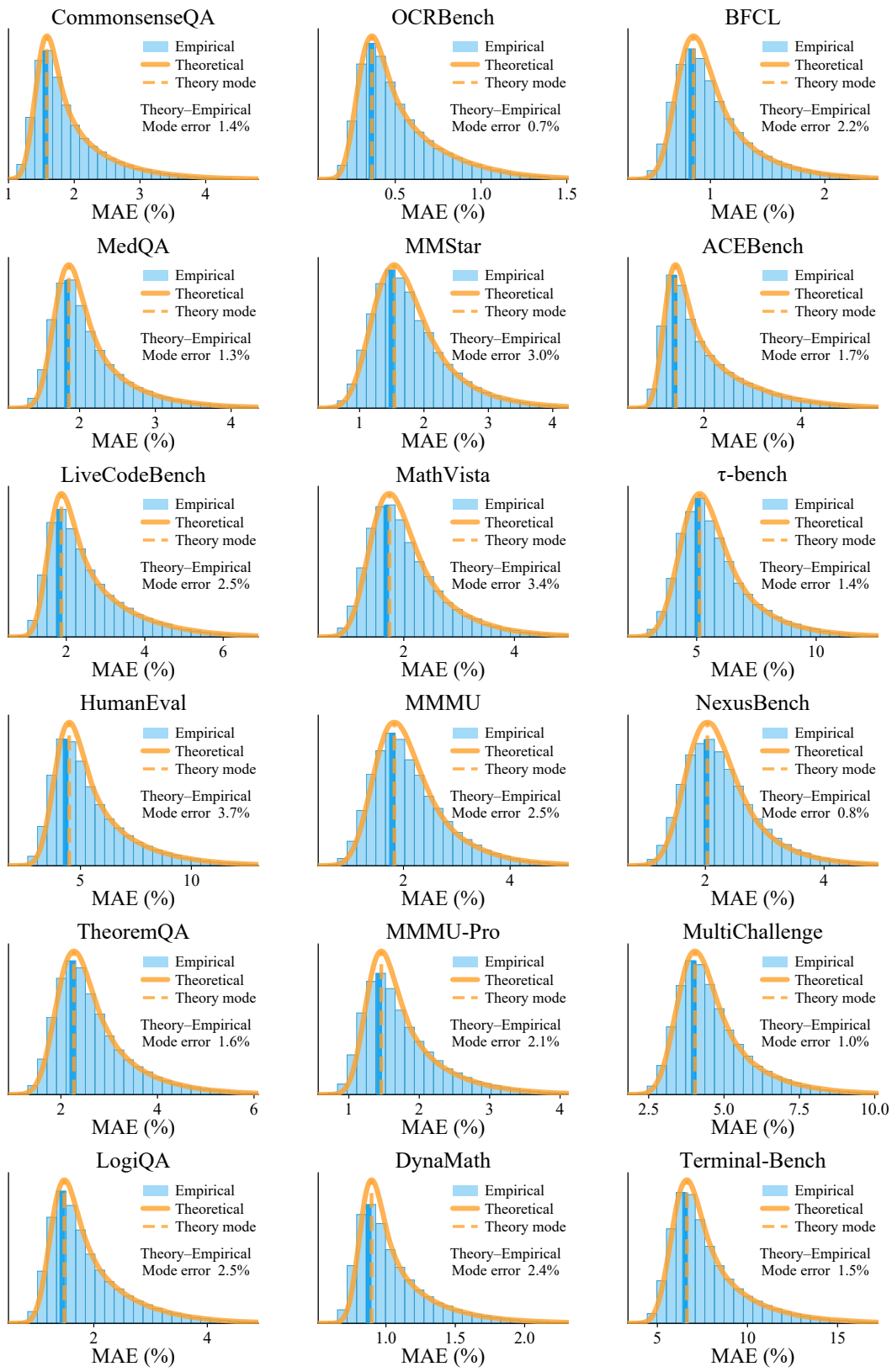

Figure 9: Empirical MAE distributions closely match theoretical predictions across representative text, multimodal, and agent benchmarks, including the predicted shapes and modes.

Table 6: Bench2Drive compression results using five training models. Within each run, all train-dependent methods use the same SR-stratified model sample. We report percentage-point MAE for SR and Smoothness, score MAE for DS, and mean relative error in percent for Efficiency. Avg. Error averages these four errors. Lower is better. Standard deviations appear in smaller type. The best and second-best means in each removal-ratio block are **bolded** and underlined, respectively.

	Prediction error ↓ mean ± std.				
Method	**SR**	**DS**	**Eff. MRE**	**Smooth.**	**Avg. Error**
Bench2Drive, removal ratio $= 50\%$					
SparseEval (Linear)	16.14 ±0.24	1.22 ±0.16	2.73 ±1.73	1.38 ±0.31	5.37 ±0.49
SparseEval (MLP)	3.50 ±0.15	1.78 ±0.19	2.31 ±2.13	1.85 ±0.73	2.36 ±0.37
Anchor Points	5.21 ±1.48	4.55 ±1.62	2.40 ±1.71	3.44 ±2.15	3.90 ±1.38
TinyBenchmarks	2.15 ±0.12	2.38 ±0.50	3.80 ±3.18	2.02 ±0.62	2.58 ±0.83
ZipBench	2.45 ±0.26	3.18 ±0.31	**1.64** ±0.33	1.92 ±0.47	2.30 ±0.08
Random	2.34 ±0.64	1.14 ±0.24	2.74 ±1.67	1.72 ±0.73	1.99 ±0.49
BestRand-SR (ours)	**1.94** ±0.29	1.03 ±0.15	3.05 ±1.96	1.55 ±0.54	1.89 ±0.51
BestRand-4Metric (ours)	1.95 ±0.32	**1.00** ±0.17	1.69 ±0.71	**1.26** ±0.25	**1.47** ±0.21
Bench2Drive, removal ratio $= 65\%$					
SparseEval (Linear)	20.49 ±1.87	2.18 ±0.29	3.19 ±1.36	1.74 ±0.53	6.90 ±0.37
SparseEval (MLP)	2.85 ±0.40	2.61 ±1.21	3.88 ±1.32	**1.40** ±0.10	2.68 ±0.54
Anchor Points	5.80 ±1.58	4.11 ±0.95	2.33 ±1.02	3.78 ±2.63	4.00 ±0.95
TinyBenchmarks	3.01 ±0.05	2.86 ±0.68	4.67 ±3.47	2.10 ±0.24	3.16 ±0.89
ZipBench	3.12 ±0.50	3.91 ±0.83	2.54 ±0.92	2.63 ±0.25	3.05 ±0.41
Random	3.38 ±1.20	1.66 ±0.54	3.99 ±2.74	2.33 ±0.98	2.84 ±0.92
BestRand-SR (ours)	2.63 ±0.44	1.42 ±0.26	3.78 ±2.28	2.03 ±0.74	2.46 ±0.63
BestRand-4Metric (ours)	**2.63** ±0.44	**1.36** ±0.25	**2.30** ±0.95	1.70 ±0.36	**2.00** ±0.31

C Additional Analysis

Section 4 analyzes the key design choices of BestRand and evaluates its effectiveness, efficiency, robustness, and OOD generalization. This section complements those results with additional experiments on VLA benchmarks, multi-metric evaluation, and data-only compression methods.

C.1 Effectiveness of BestRand on VLA Benchmarks

Section 3 reports results on text, multimodal, and agent benchmarks. Here, we provide additional results on VLA benchmarks, where evaluation is especially costly because each item requires an interactive task rollout rather than a single model inference. We evaluate BestRand on Bench2Drive (Jia et al., 2024) and Fail2Drive (Gerstenecker et al., 2026). Both experiments construct subsets from five training records and evaluate them on held-out records at removal ratios of 50% and 65%. To evaluate the direct transfer of the original method, BestRand-SR selects routes using success rate alone. The next subsection extends BestRand to jointly optimize multiple metrics.

Tables 6 and 7 show that BestRand-SR is the strongest single-metric method on both benchmarks. Compared with Random, it reduces Avg. Error from 1.99 to 1.89 at 50% removal and from 2.84 to 2.46 at 65% removal on Bench2Drive. On Fail2Drive, it reduces the error from 2.56 to 2.14 and from 3.38 to 2.88, respectively. It also improves success-rate estimation at both compression levels on both benchmarks. These results demonstrate that lightweight probe matching transfers effectively to closed-loop VLA evaluation across benchmarks and evaluation splits.

C.2 Effectiveness of BestRand on Multi-Metric Benchmarks

Appendix A.2 summarizes all benchmarks used in our study. The main experiments use one metric per benchmark, selecting a representative metric when multiple metrics are available. Here, we examine whether BestRand can extend to multi-metric settings and preserve several metrics simultaneously. We normalize each metric separately and jointly match their probe profiles. BestRand-

Table 7: Fail2Drive paired-route compression results using five uniformly sampled training records. The same selected route pairs are used for ID and generalization evaluation. Random and BestRand use 100 seeds, while fitted baselines use three. We report test-record MAE as mean ± standard deviation. Lower is better. Standard deviations appear in smaller type. The best and second-best means in each removal-ratio block are **bolded** and underlined, respectively.

	Prediction error ↓ mean ± std.						
Method	**ID SR**	**ID DS**	**ID HM**	**Gen SR**	**Gen DS**	**Gen HM**	**Avg. Error**
Fail2Drive, removal ratio $= 50\%$							
SparseEval (Linear)	17.40±8.19	4.63±0.62	10.44±7.40	18.48±2.77	5.56±0.78	13.17±4.26	11.62±3.38
SparseEval (MLP)	4.56±0.67	3.53±0.95	3.07±0.66	6.71±2.20	3.86±0.55	3.17±0.81	4.15±0.36
Anchor Points	2.79±0.09	2.51±0.63	2.09±0.43	5.33±1.30	3.99±0.42	3.88±1.18	3.43±0.39
TinyBenchmarks	2.52±0.23	1.81±0.49	1.89±0.28	3.49±0.93	**1.63**±0.58	2.57±0.56	2.32±0.23
ZipBench	2.43±0.35	3.06±1.17	2.17±0.60	**2.87**±0.37	2.51±0.89	**2.42**±0.29	2.58±0.39
Random	2.69±0.71	1.66±0.58	2.14±0.63	3.80±1.06	2.02±0.61	3.04±0.88	2.56±0.56
BestRand-SR (ours)	2.34±0.50	1.41±0.40	1.80±0.41	3.08±0.58	1.74±0.43	2.46±0.47	2.14±0.32
BestRand-2Metric (ours)	**2.30**±0.41	**1.27**±0.23	**1.72**±0.29	3.14±0.64	1.65±0.38	2.48±0.51	**2.09**±0.29
Fail2Drive, removal ratio $= 65\%$							
SparseEval (Linear)	19.99±9.27	5.24±2.08	13.30±8.77	18.93±2.93	4.59±0.77	13.01±5.05	12.51±3.89
SparseEval (MLP)	5.62±1.43	4.58±1.24	4.19±0.75	5.48±1.22	3.66±0.51	**3.24**±0.23	4.46±0.11
Anchor Points	3.60±0.15	4.27±0.95	3.37±0.27	5.00±1.75	4.34±1.08	3.32±0.47	3.98±0.43
TinyBenchmarks	3.39±0.45	2.79±1.17	2.81±0.86	5.50±0.90	2.62±0.65	4.24±0.83	3.56±0.71
ZipBench	3.20±0.75	3.32±1.64	2.70±1.12	4.54±0.59	3.25±0.71	3.41±0.39	3.40±0.68
Random	3.76±1.04	2.22±0.80	2.94±0.89	4.88±1.19	2.60±0.71	3.89±1.01	3.38±0.71
BestRand-SR (ours)	3.20±0.55	1.82±0.42	2.42±0.45	4.23±0.89	2.26±0.48	3.33±0.71	2.88±0.37
BestRand-2Metric (ours)	**3.18**±0.51	**1.75**±0.34	**2.39**±0.39	**4.20**±0.82	**2.19**±0.46	3.31±0.68	**2.83**±0.36

4Metric uses success rate, driving score, efficiency, and smoothness on Bench2Drive. BestRand-2Metric uses success rate and driving score on Fail2Drive, with the harmonic mean used only for evaluation. This extension changes only the selection objective and retains the same subset size.

Joint metric matching achieves the lowest overall error on both benchmarks. Compared with BestRand-SR, BestRand-4Metric reduces Avg. Error from 1.89 to 1.47 at 50% removal and from 2.46 to 2.00 at 65% removal on Bench2Drive. The largest gains are in efficiency and smoothness, while success-rate accuracy is maintained. On Fail2Drive, BestRand-2Metric reduces Avg. Error from 2.14 to 2.09 and from 2.88 to 2.83. Although smaller, these gains are consistent across compression levels and maintain balanced in-distribution and generalization performance. These results show that one compressed subset can accurately preserve multiple benchmark metrics.

C.3 Can Data-Only Methods Achieve Effective Benchmark Compression?

Some benchmark compression methods require no historical model responses and select items solely from representations of the benchmark data. This makes subset construction inexpensive and practical when performance records are unavailable. We study four such methods. Facility Location (Yao et al., 2026) selects items that cover the prompt embedding space, while MIS (Kjorvezir et al., 2026) constructs a similarity graph and chooses a nonredundant independent set. Agreement selects prompts with the greatest variation in attention weights, while Coverage iteratively selects distant prompts to span diverse regions of the representation space (Purificato et al., 2026). Appendix A.3 describes these methods, and Appendix A.4 provides their implementation settings.

Figure 10 compares the four data-only methods with Random and BestRand at 20% retention. Although inexpensive to construct, the data-only subsets often yield substantially higher MAE than Random and BestRand, with performance varying markedly across benchmarks. BestRand achieves the lowest MAE on five of the six benchmarks and is nearly tied for the lowest on GPQA, yielding the strongest overall accuracy. These results show that a small initial investment in probe-model inference clearly improves accuracy over methods that rely only on benchmark data.

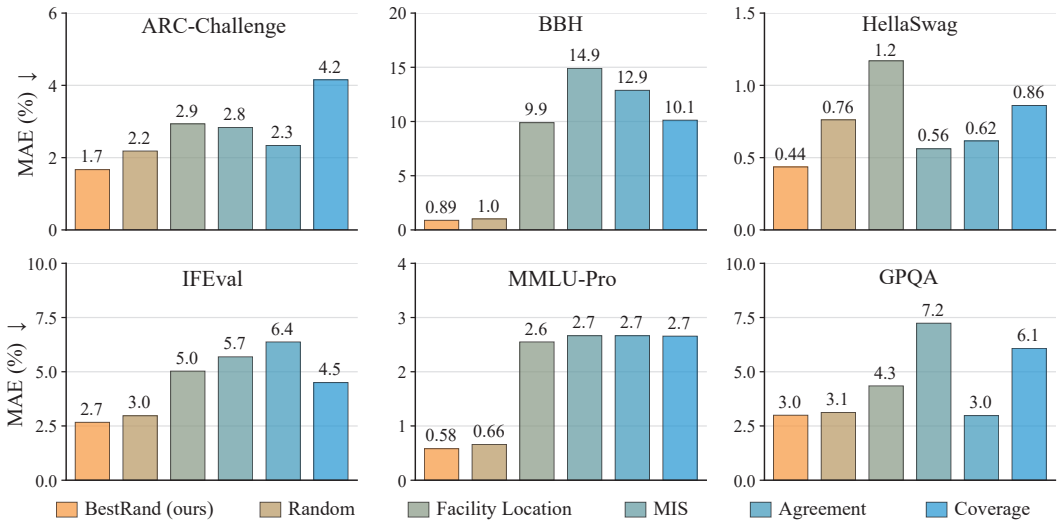

Figure 10: Performance of data-only compression methods at 20% retention on six representative benchmarks. Facility Location, MIS, Agreement, and Coverage select items from prompt-derived representations without historical model responses. Random also requires no historical responses, while BestRand evaluates a few probe models only on its candidate pool. Lower MAE is better.

D Additional Statements

D.1 Limitations of BestRand

BestRand provides an efficient approximation to full evaluation, but several limitations remain.

Nonzero estimation error. Although BestRand outperforms many existing benchmark compression methods at lower construction cost, it cannot guarantee zero estimation error. This is a common limitation of benchmark compression, but the compression ratio can be adjusted to balance evaluation efficiency and reliability. BestRand is most useful for rapid model iteration and comparison, but compressed benchmarks cannot fully replace comprehensive evaluation. We therefore recommend reporting full-benchmark results for final evaluations when feasible.

Probe inference cost. BestRand still requires probe models to run full inference on some benchmark items, so its construction cost is not zero. However, this cost is small in practice and brings clear accuracy gains. BestRand substantially outperforms data-only methods that use no model responses. Moreover, only a few probe models are needed, and they score only the candidate pool, which can be much smaller than the full benchmark at low retention ratios. BestRand therefore achieves better accuracy at a lower construction cost than existing response-based methods.

D.2 Broader Impact of BestRand

BestRand provides stable, accurate, and low-cost benchmark compression, reducing evaluation costs while preserving reliable results. Its efficient construction process also keeps the cost of creating compressed benchmarks low. Together, these benefits can reduce compute and energy use, lower associated carbon emissions, and support faster experimental iteration. By making reliable evaluation more affordable, BestRand can also lower barriers for smaller laboratories and researchers with limited computing resources, enabling broader participation in model development and evaluation.

D.3 Computational Resources Statement

Our experiments use historical model performance records collected through prior research, so their original computational cost cannot be reliably estimated. All benchmark compression experiments can run on a single GPU. BestRand's subset selection requires only a CPU and completes in under 0.1 seconds with parallel execution. We will release the code and data to support reproducibility.

Table 8: Spearman rank correlation across 70 benchmarks at 5–50% retention. Values are mean ± sample standard deviation across benchmarks, and the Mean column pools all displayed settings. Higher is better. Best means are **bold** and second-best means are underlined.

	Average Spearman correlation ↑ mean ± std.					
Method	**5%**	**10%**	**15%**	**20%**	**50%**	**Mean**
Text 44 benchmarks						
SparseEval (Linear)	0.381 ±0.238	0.364 ±0.252	0.384 ±0.242	0.347 ±0.271	0.407 ±0.265	0.377 ±0.252
SparseEval (MLP)	0.466 ±0.288	0.514 ±0.263	0.586 ±0.263	0.588 ±0.303	0.638 ±0.308	0.558 ±0.290
Anchor Points	0.756 ±0.190	0.791 ±0.165	0.801 ±0.165	0.800 ±0.172	0.862 ±0.146	0.802 ±0.170
TinyBenchmarks	0.841 ±0.181	0.883 ±0.152	0.900 ±0.164	0.932 ±0.100	0.966 ±0.053	0.904 ±0.143
ZipBench	0.837 ±0.193	0.890 ±0.151	0.901 ±0.163	0.927 ±0.118	0.967 ±0.059	0.904 ±0.149
Random	0.855 ±0.170	0.903 ±0.129	0.926 ±0.103	0.941 ±0.084	0.978 ±0.031	0.921 ±0.120
BestRand (ours)	**0.857** ±0.167	**0.904** ±0.129	**0.927** ±0.101	**0.942** ±0.083	**0.978** ±0.031	**0.922** ±0.118
Multimodal 16 benchmarks						
SparseEval (Linear)	0.327 ±0.319	0.252 ±0.329	0.256 ±0.228	0.254 ±0.319	0.274 ±0.237	0.272 ±0.284
SparseEval (MLP)	0.363 ±0.377	0.576 ±0.225	0.506 ±0.267	0.588 ±0.280	0.607 ±0.324	0.528 ±0.305
Anchor Points	0.772 ±0.193	0.749 ±0.186	0.757 ±0.206	0.774 ±0.183	0.842 ±0.160	0.779 ±0.185
TinyBenchmarks	0.818 ±0.202	0.885 ±0.142	0.922 ±0.081	0.930 ±0.089	0.961 ±0.044	0.903 ±0.131
ZipBench	0.795 ±0.311	0.868 ±0.181	0.917 ±0.114	0.940 ±0.055	0.963 ±0.051	0.897 ±0.178
Random	**0.850** ±0.145	**0.903** ±0.106	0.927 ±0.084	0.941 ±0.070	0.976 ±0.032	**0.919** ±0.102
BestRand (ours)	0.840 ±0.168	0.902 ±0.106	**0.927** ±0.084	**0.941** ±0.069	**0.976** ±0.033	0.917 ±0.110
Agent 10 benchmarks						
SparseEval (Linear)	0.275 ±0.157	0.243 ±0.226	0.326 ±0.221	0.339 ±0.212	0.254 ±0.260	0.288 ±0.212
SparseEval (MLP)	0.230 ±0.292	0.330 ±0.331	0.316 ±0.274	0.454 ±0.243	0.554 ±0.292	0.377 ±0.299
Anchor Points	0.678 ±0.176	0.726 ±0.121	0.750 ±0.127	0.724 ±0.160	0.860 ±0.101	0.748 ±0.147
TinyBenchmarks	0.727 ±0.179	0.815 ±0.134	0.851 ±0.116	0.878 ±0.101	0.943 ±0.049	0.843 ±0.138
ZipBench	0.752 ±0.161	0.829 ±0.124	0.880 ±0.094	0.891 ±0.088	0.935 ±0.054	0.858 ±0.123
Random	**0.759** ±0.166	0.842 ±0.118	0.880 ±0.093	0.905 ±0.075	0.965 ±0.028	0.870 ±0.123
BestRand (ours)	0.755 ±0.173	**0.843** ±0.118	**0.881** ±0.093	**0.905** ±0.075	**0.965** ±0.028	**0.870** ±0.126

E Additional Experimental Results

The main experiments in Section 3 evaluate BestRand on 70 text, multimodal, and agent benchmarks and report results averaged across benchmarks. This section complements these aggregate findings with detailed results on representative benchmarks. We first report benchmark-level results for Spearman's rank correlation coefficient (SP) and mean absolute error (MAE). Finally, we provide additional visualizations that complement and extend the figures in the main text.

E.1 Benchmark-Wise Results for Spearman Rank Correlation

Table 8 summarizes Spearman rank correlation (SP) across all 70 benchmarks, while Tables 9, 10, and 11 report results for six representative benchmarks in each category. SP quantifies agreement between model rankings derived from compressed and full benchmarks. Higher values indicate better rank preservation, with 1 denoting identical rankings, but do not guarantee small score errors. Leading methods achieve high SP at moderate and high retention ratios. Averaged across TinyBenchmarks, ZipBench, Random, and BestRand over all benchmarks, SP rises from 0.930 at 20% retention to 0.969 at 50%. This strong rank preservation, also observed in prior work (Yauney et al., 2026), motivates our use of MAE as the primary metric. Full SP results are reported here.

E.2 Benchmark-Wise Detailed Results for MAE

Tables 12, 13, and 14 provide detailed MAE results for six representative text, multimodal, and agent benchmarks, respectively, using the same 18 benchmarks as the SP tables.

Table 9: Text benchmark Spearman correlations at 5–50% retention. Values are mean ± sample standard deviation; higher is better. Best means are **bold** and second-best means are underlined.

	Average Spearman correlation ↑ mean ± std.					
Method	**5%**	**10%**	**15%**	**20%**	**50%**	**Mean**
TruthfulQA						
SparseEval (Linear)	-0.002 ±0.420	0.627 ±0.227	0.672 ±0.174	0.697 ±0.109	0.729 ±0.071	0.545 ±0.347
SparseEval (MLP)	-0.145 ±0.776	0.255 ±0.951	0.415 ±0.554	0.522 ±0.364	0.684 ±0.303	0.346 ±0.614
Anchor Points	0.884 ±0.010	**0.974** ±0.010	0.905 ±0.054	0.974 ±0.006	0.974 ±0.006	0.942 ±0.046
TinyBenchmarks	0.896 ±0.075	0.925 ±0.042	0.959 ±0.014	0.970 ±0.009	0.981 ±0.013	0.946 ±0.047
ZipBench	**0.946** ±0.045	0.958 ±0.034	0.970 ±0.008	0.958 ±0.012	0.990 ±0.005	0.965 ±0.027
Random	0.930 ±0.036	0.959 ±0.021	0.970 ±0.015	0.977 ±0.012	0.992 ±0.004	0.966 ±0.029
BestRand (ours)	0.933 ±0.033	0.961 ±0.019	**0.972** ±0.014	**0.979** ±0.011	**0.992** ±0.004	**0.967** ±0.027
BoolQ						
SparseEval (Linear)	0.263 ±0.145	0.519 ±0.197	0.497 ±0.220	0.509 ±0.325	0.278 ±0.321	0.413 ±0.245
SparseEval (MLP)	0.197 ±0.203	0.294 ±0.195	0.333 ±0.512	0.489 ±0.371	0.196 ±0.633	0.302 ±0.372
Anchor Points	0.407 ±0.178	0.573 ±0.218	0.784 ±0.043	0.612 ±0.356	0.790 ±0.021	0.633 ±0.228
TinyBenchmarks	0.903 ±0.023	0.959 ±0.014	0.966 ±0.022	0.976 ±0.012	0.993 ±0.006	0.959 ±0.034
ZipBench	0.855 ±0.005	0.933 ±0.032	0.976 ±0.003	0.978 ±0.007	0.986 ±0.003	0.946 ±0.052
Random	0.934 ±0.032	**0.965** ±0.017	0.976 ±0.011	0.982 ±0.009	0.994 ±0.003	0.970 ±0.027
BestRand (ours)	**0.934** ±0.032	0.965 ±0.016	**0.977** ±0.011	**0.983** ±0.009	**0.994** ±0.003	**0.970** ±0.027
IFEval						
SparseEval (Linear)	0.564 ±0.140	0.670 ±0.125	0.601 ±0.188	0.667 ±0.094	0.232 ±0.559	0.547 ±0.290
SparseEval (MLP)	0.387 ±0.251	0.793 ±0.116	0.812 ±0.118	0.871 ±0.089	0.889 ±0.105	0.751 ±0.229
Anchor Points	0.891 ±0.018	0.912 ±0.035	0.900 ±0.050	0.888 ±0.081	0.954 ±0.011	0.909 ±0.046
TinyBenchmarks	0.923 ±0.020	0.949 ±0.033	0.978 ±0.003	0.980 ±0.002	0.992 ±0.003	0.965 ±0.030
ZipBench	0.861 ±0.029	0.949 ±0.035	0.971 ±0.015	0.982 ±0.003	0.987 ±0.007	0.950 ±0.051
Random	0.942 ±0.031	0.970 ±0.016	0.980 ±0.010	**0.985** ±0.008	**0.993** ±0.004	0.974 ±0.024
BestRand (ours)	**0.944** ±0.028	**0.970** ±0.015	**0.980** ±0.010	0.985 ±0.008	0.993 ±0.004	**0.974** ±0.023
Finance-Fundamentals						
SparseEval (Linear)	0.409 ±0.320	0.021 ±0.510	0.303 ±0.187	0.264 ±0.154	0.040 ±0.649	0.208 ±0.381
SparseEval (MLP)	0.267 ±0.242	0.369 ±0.259	0.564 ±0.136	0.645 ±0.040	0.753 ±0.091	0.520 ±0.237
Anchor Points	0.757 ±0.067	0.830 ±0.050	0.822 ±0.032	0.834 ±0.072	0.940 ±0.018	0.836 ±0.075
TinyBenchmarks	0.796 ±0.028	0.838 ±0.028	0.923 ±0.037	0.946 ±0.019	0.973 ±0.005	0.895 ±0.073
ZipBench	0.819 ±0.038	0.807 ±0.087	0.900 ±0.041	0.923 ±0.014	0.972 ±0.016	0.884 ±0.076
Random	0.822 ±0.084	0.901 ±0.043	**0.933** ±0.029	**0.949** ±0.022	**0.984** ±0.008	0.918 ±0.071
BestRand (ours)	**0.825** ±0.080	**0.903** ±0.041	0.932 ±0.028	0.949 ±0.022	0.984 ±0.008	**0.919** ±0.069
LIFBench						
SparseEval (Linear)	0.657 ±0.186	0.261 ±0.635	0.210 ±0.374	0.377 ±0.375	0.288 ±0.398	0.359 ±0.390
SparseEval (MLP)	0.792 ±0.051	0.827 ±0.034	0.852 ±0.030	0.807 ±0.055	0.848 ±0.021	0.825 ±0.042
Anchor Points	0.926 ±0.039	0.965 ±0.039	0.962 ±0.014	0.962 ±0.018	0.986 ±0.009	0.960 ±0.030
TinyBenchmarks	0.932 ±0.040	0.952 ±0.020	0.952 ±0.009	0.950 ±0.028	0.958 ±0.002	0.949 ±0.022
ZipBench	0.964 ±0.009	**0.986** ±0.007	**0.987** ±0.002	**0.992** ±0.008	0.992 ±0.004	**0.984** ±0.012
Random	0.961 ±0.024	0.976 ±0.016	0.982 ±0.012	0.985 ±0.010	0.992 ±0.006	0.979 ±0.018
BestRand (ours)	**0.966** ±0.020	0.979 ±0.014	0.984 ±0.010	0.987 ±0.009	**0.993** ±0.006	0.981 ±0.016
GPQA						
SparseEval (Linear)	0.391 ±0.138	0.187 ±0.427	0.297 ±0.364	0.113 ±0.143	-0.096 ±0.366	0.178 ±0.315
SparseEval (MLP)	-0.103 ±0.248	-0.112 ±0.361	-0.078 ±0.457	0.151 ±0.025	0.608 ±0.141	0.093 ±0.376
Anchor Points	0.396 ±0.351	0.268 ±0.343	0.367 ±0.109	0.397 ±0.070	0.441 ±0.099	0.374 ±0.204
TinyBenchmarks	0.231 ±0.121	0.474 ±0.104	**0.668** ±0.083	0.722 ±0.070	0.819 ±0.132	0.583 ±0.233
ZipBench	0.092 ±0.316	**0.598** ±0.183	0.537 ±0.048	0.531 ±0.143	0.766 ±0.112	0.505 ±0.278
Random	0.430 ±0.212	0.573 ±0.171	0.658 ±0.142	0.717 ±0.120	0.884 ±0.055	0.652 ±0.212
BestRand (ours)	**0.445** ±0.196	0.576 ±0.159	0.662 ±0.131	**0.725** ±0.111	**0.886** ±0.052	**0.659** ±0.202

Table 10: Multimodal benchmark Spearman correlations at 5–50% retention. Values are mean ± sample standard deviation; higher is better. Best means are **bold** and second-best are underlined.

	Average Spearman correlation ↑ mean ± std.					
Method	**5%**	**10%**	**15%**	**20%**	**50%**	**Mean**
MathVerse						
SparseEval (Linear)	0.199 ±0.372	-0.318 ±0.355	-0.306 ±0.128	0.043 ±0.435	0.107 ±0.411	-0.055 ±0.375
SparseEval (MLP)	-0.177 ±0.440	0.450 ±0.423	0.618 ±0.160	0.747 ±0.048	0.445 ±0.065	0.417 ±0.407
Anchor Points	0.771 ±0.142	0.783 ±0.071	0.802 ±0.114	0.846 ±0.053	0.958 ±0.010	0.832 ±0.104
TinyBenchmarks	0.853 ±0.054	0.936 ±0.026	0.946 ±0.016	0.962 ±0.020	**0.989** ±0.004	0.937 ±0.053
ZipBench	0.897 ±0.055	0.893 ±0.027	0.942 ±0.029	0.957 ±0.009	0.987 ±0.007	0.935 ±0.045
Random	0.907 ±0.043	**0.948** ±0.022	**0.963** ±0.016	0.971 ±0.013	0.988 ±0.006	0.955 ±0.036
BestRand (ours)	**0.907** ±0.041	0.948 ±0.022	0.963 ±0.016	**0.971** ±0.013	0.988 ±0.006	**0.955** ±0.036
OmniDocBench						
SparseEval (Linear)	-0.448 ±0.315	0.105 ±0.662	0.162 ±0.493	-0.181 ±0.616	-0.029 ±0.674	-0.078 ±0.531
SparseEval (MLP)	0.543 ±0.605	0.467 ±0.436	0.619 ±0.293	0.848 ±0.066	0.581 ±0.175	0.611 ±0.338
Anchor Points	**0.886** ±0.099	0.829 ±0.000	0.829 ±0.151	0.905 ±0.066	0.905 ±0.119	0.870 ±0.093
TinyBenchmarks	0.638 ±0.175	0.810 ±0.144	0.733 ±0.175	0.771 ±0.099	0.886 ±0.099	0.768 ±0.147
ZipBench	0.829 ±0.099	**0.943** ±0.057	0.886 ±0.057	0.924 ±0.033	0.924 ±0.087	0.901 ±0.073
Random	0.831 ±0.152	0.895 ±0.091	0.920 ±0.074	0.936 ±0.064	0.976 ±0.033	0.912 ±0.103
BestRand (ours)	0.836 ±0.143	0.898 ±0.089	**0.922** ±0.073	**0.937** ±0.064	**0.976** ±0.032	**0.914** ±0.099
SpatialEval						
SparseEval (Linear)	0.776 ±0.160	0.479 ±0.464	0.462 ±0.655	0.490 ±0.665	0.445 ±0.714	0.531 ±0.498
SparseEval (MLP)	0.681 ±0.218	0.843 ±0.229	0.899 ±0.044	0.888 ±0.035	0.759 ±0.245	0.814 ±0.175
Anchor Points	0.978 ±0.026	0.978 ±0.026	0.989 ±0.010	0.978 ±0.026	0.989 ±0.019	0.982 ±0.020
TinyBenchmarks	**0.989** ±0.010	0.961 ±0.019	0.983 ±0.029	**1.000** ±0.000	**1.000** ±0.000	0.987 ±0.020
ZipBench	0.983 ±0.017	0.961 ±0.035	0.978 ±0.026	0.978 ±0.026	**1.000** ±0.000	0.980 ±0.024
Random	0.979 ±0.023	0.987 ±0.017	0.991 ±0.014	0.993 ±0.012	0.998 ±0.005	**0.990** ±0.017
BestRand (ours)	0.974 ±0.026	**0.988** ±0.017	**0.991** ±0.014	0.993 ±0.012	0.998 ±0.005	0.989 ±0.018
LogicVista						
SparseEval (Linear)	-0.038 ±0.282	-0.008 ±0.448	0.084 ±0.312	0.168 ±0.158	0.367 ±0.200	0.115 ±0.293
SparseEval (MLP)	0.175 ±0.533	0.387 ±0.289	0.412 ±0.491	0.382 ±0.514	0.869 ±0.028	0.445 ±0.425
Anchor Points	**0.795** ±0.070	0.664 ±0.042	0.829 ±0.085	0.691 ±0.247	0.866 ±0.035	0.769 ±0.132
TinyBenchmarks	0.729 ±0.046	**0.868** ±0.020	0.884 ±0.034	0.902 ±0.039	0.939 ±0.021	0.865 ±0.080
ZipBench	0.739 ±0.154	0.777 ±0.087	0.862 ±0.028	0.882 ±0.041	0.968 ±0.020	0.846 ±0.109
Random	0.763 ±0.122	0.866 ±0.065	**0.905** ±0.043	0.927 ±0.033	**0.974** ±0.013	0.887 ±0.098
BestRand (ours)	0.766 ±0.111	0.867 ±0.060	0.905 ±0.041	**0.927** ±0.031	0.973 ±0.013	**0.888** ±0.093
RealWorldQA						
SparseEval (Linear)	0.115 ±0.580	0.200 ±0.429	0.309 ±0.073	0.333 ±0.200	0.079 ±0.431	0.207 ±0.344
SparseEval (MLP)	-0.358 ±0.138	0.115 ±0.606	0.285 ±0.738	0.188 ±0.207	-0.248 ±0.421	-0.004 ±0.483
Anchor Points	0.382 ±0.601	0.721 ±0.255	0.236 ±0.419	0.398 ±0.803	0.572 ±0.418	0.462 ±0.483
TinyBenchmarks	0.248 ±0.704	0.673 ±0.109	**0.842** ±0.164	**0.891** ±0.109	0.927 ±0.063	0.716 ±0.381
ZipBench	-0.297 ±0.480	0.624 ±0.309	0.818 ±0.063	0.855 ±0.221	0.891 ±0.096	0.578 ±0.519
Random	**0.598** ±0.304	0.732 ±0.222	0.801 ±0.177	0.846 ±0.147	**0.955** ±0.061	**0.786** ±0.232
BestRand (ours)	0.543 ±0.328	**0.735** ±0.219	0.783 ±0.186	0.831 ±0.159	0.954 ±0.061	0.769 ±0.249
OCRBench						
SparseEval (Linear)	0.556 ±0.179	0.778 ±0.099	0.381 ±0.304	0.357 ±0.290	0.071 ±0.234	0.429 ±0.312
SparseEval (MLP)	0.929 ±0.041	0.905 ±0.024	0.921 ±0.036	0.929 ±0.024	0.952 ±0.063	0.927 ±0.038
Anchor Points	0.984 ±0.014	0.976 ±0.000	0.841 ±0.172	0.841 ±0.234	0.881 ±0.126	0.905 ±0.136
TinyBenchmarks	**0.992** ±0.014	**1.000** ±0.000	0.992 ±0.014	**1.000** ±0.000	**1.000** ±0.000	**0.997** ±0.008
ZipBench	**0.992** ±0.014	0.992 ±0.014	**1.000** ±0.000	0.992 ±0.014	**1.000** ±0.000	0.995 ±0.010
Random	0.989 ±0.016	0.995 ±0.010	0.997 ±0.008	0.998 ±0.006	1.000 ±0.001	0.996 ±0.010
BestRand (ours)	0.990 ±0.015	0.995 ±0.010	0.997 ±0.008	0.998 ±0.006	1.000 ±0.001	0.996 ±0.010

Table 11: Agent benchmark Spearman correlations at 5–50% retention. Values are mean ± sample standard deviation; higher is better. Best means are **bold** and second-best means are underlined.

	Average Spearman correlation ↑ mean ± std.					
Method	**5%**	**10%**	**15%**	**20%**	**50%**	**Mean**
Drafter-Bench						
SparseEval (Linear)	0.287 ±0.353	-0.065 ±0.383	-0.055 ±0.471	0.011 ±0.322	0.240 ±0.153	0.084 ±0.336
SparseEval (MLP)	0.667 ±0.170	0.685 ±0.101	0.227 ±0.627	0.556 ±0.080	0.683 ±0.097	0.564 ±0.311
Anchor Points	0.797 ±0.067	0.868 ±0.056	0.853 ±0.111	0.951 ±0.038	0.964 ±0.024	0.887 ±0.086
TinyBenchmarks	0.842 ±0.029	0.897 ±0.009	0.919 ±0.013	0.927 ±0.029	0.938 ±0.006	0.905 ±0.039
ZipBench	0.896 ±0.053	0.946 ±0.035	0.959 ±0.016	**0.980** ±0.013	0.984 ±0.009	0.953 ±0.042
Random	0.916 ±0.041	0.955 ±0.021	0.968 ±0.014	0.976 ±0.011	0.991 ±0.005	0.961 ±0.034
BestRand (ours)	**0.920** ±0.036	**0.957** ±0.019	**0.969** ±0.014	0.977 ±0.011	**0.991** ±0.005	**0.963** ±0.031
OSWorld						
SparseEval (Linear)	0.556 ±0.241	0.422 ±0.197	0.675 ±0.101	0.533 ±0.024	0.435 ±0.148	0.524 ±0.166
SparseEval (MLP)	0.253 ±0.235	-0.223 ±0.394	0.063 ±0.376	0.409 ±0.186	0.743 ±0.168	0.249 ±0.415
Anchor Points	0.811 ±0.090	0.847 ±0.041	0.861 ±0.028	0.909 ±0.017	0.953 ±0.011	0.876 ±0.065
TinyBenchmarks	0.779 ±0.082	0.904 ±0.031	0.925 ±0.030	0.947 ±0.014	0.985 ±0.005	0.908 ±0.080
ZipBench	**0.919** ±0.043	**0.949** ±0.012	**0.962** ±0.012	0.952 ±0.025	**0.988** ±0.005	**0.954** ±0.030
Random	0.896 ±0.051	0.939 ±0.030	0.955 ±0.023	0.965 ±0.018	0.987 ±0.008	0.948 ±0.043
BestRand (ours)	0.902 ±0.045	0.943 ±0.027	0.958 ±0.020	**0.967** ±0.017	0.987 ±0.008	0.951 ±0.039
BFCL						
SparseEval (Linear)	0.229 ±0.369	0.288 ±0.431	0.261 ±0.313	0.340 ±0.320	0.220 ±0.528	0.268 ±0.341
SparseEval (MLP)	0.450 ±0.371	0.790 ±0.144	0.804 ±0.058	0.791 ±0.041	0.695 ±0.332	0.706 ±0.241
Anchor Points	0.825 ±0.131	0.706 ±0.202	0.786 ±0.041	0.483 ±0.391	0.865 ±0.111	0.733 ±0.227
TinyBenchmarks	**0.973** ±0.008	**0.989** ±0.007	0.982 ±0.006	0.988 ±0.000	0.994 ±0.003	0.985 ±0.009
ZipBench	0.964 ±0.004	**0.989** ±0.004	**0.991** ±0.003	0.990 ±0.009	**1.000** ±0.000	**0.987** ±0.013
Random	0.969 ±0.017	0.983 ±0.009	0.988 ±0.007	0.990 ±0.006	0.996 ±0.004	0.985 ±0.013
BestRand (ours)	0.970 ±0.016	0.984 ±0.009	0.988 ±0.007	**0.991** ±0.006	0.996 ±0.004	0.986 ±0.013
TauBench						
SparseEval (Linear)	0.276 ±0.468	0.046 ±0.404	0.441 ±0.269	0.479 ±0.198	0.354 ±0.232	0.319 ±0.322
SparseEval (MLP)	0.419 ±0.151	0.345 ±0.164	0.589 ±0.046	0.474 ±0.206	0.739 ±0.013	0.513 ±0.184
Anchor Points	0.354 ±0.317	0.658 ±0.210	0.634 ±0.167	0.668 ±0.147	0.723 ±0.111	0.607 ±0.218
TinyBenchmarks	0.633 ±0.057	0.632 ±0.070	0.739 ±0.092	0.723 ±0.114	0.921 ±0.023	0.730 ±0.127
ZipBench	0.552 ±0.238	0.686 ±0.125	0.783 ±0.105	0.730 ±0.147	0.853 ±0.055	0.721 ±0.162
Random	**0.638** ±0.154	**0.747** ±0.111	**0.800** ±0.090	**0.834** ±0.076	**0.939** ±0.029	**0.792** ±0.141
BestRand (ours)	0.629 ±0.158	0.744 ±0.113	0.798 ±0.092	0.832 ±0.077	0.938 ±0.029	0.788 ±0.145
NexusBench						
SparseEval (Linear)	0.453 ±0.225	0.140 ±0.390	0.225 ±0.523	0.242 ±0.454	0.549 ±0.306	0.322 ±0.369
SparseEval (MLP)	-0.061 ±0.058	-0.015 ±0.153	0.027 ±0.214	0.201 ±0.104	0.209 ±0.164	0.072 ±0.171
Anchor Points	0.760 ±0.036	0.671 ±0.062	0.698 ±0.035	0.690 ±0.122	0.861 ±0.053	0.736 ±0.093
TinyBenchmarks	0.811 ±0.031	**0.916** ±0.065	**0.953** ±0.036	0.936 ±0.030	**0.977** ±0.007	0.919 ±0.068
ZipBench	0.853 ±0.071	0.897 ±0.044	0.933 ±0.040	**0.949** ±0.017	0.936 ±0.044	0.914 ±0.053
Random	0.851 ±0.080	0.909 ±0.047	0.930 ±0.037	0.942 ±0.031	0.971 ±0.017	0.921 ±0.062
BestRand (ours)	**0.859** ±0.076	0.908 ±0.047	0.933 ±0.035	0.945 ±0.030	0.972 ±0.017	**0.923** ±0.059
ComplexFuncBench						
SparseEval (Linear)	-0.012 ±0.255	0.640 ±0.095	0.483 ±0.254	0.580 ±0.190	0.269 ±0.354	0.392 ±0.322
SparseEval (MLP)	0.274 ±0.424	0.444 ±0.472	0.466 ±0.489	0.794 ±0.107	0.585 ±0.378	0.513 ±0.381
Anchor Points	0.801 ±0.087	0.785 ±0.027	0.869 ±0.054	0.828 ±0.052	0.876 ±0.039	0.832 ±0.060
TinyBenchmarks	**0.958** ±0.005	0.948 ±0.007	**0.972** ±0.015	**0.975** ±0.004	0.983 ±0.003	**0.967** ±0.015
ZipBench	0.834 ±0.084	0.903 ±0.051	0.945 ±0.026	0.949 ±0.011	0.977 ±0.003	0.922 ±0.064
Random	0.922 ±0.034	**0.955** ±0.019	0.968 ±0.014	0.974 ±0.011	0.989 ±0.005	0.962 ±0.030
BestRand (ours)	0.922 ±0.034	0.955 ±0.019	0.968 ±0.014	0.974 ±0.012	**0.989** ±0.005	0.962 ±0.030

Table 12: Text benchmark MAE at 2–50% retention. Values are mean ± sample standard deviation in percent; lower is better. Best means are **bold** and second-best means are underlined.

	Average MAE (%) ↓ mean ± std.						
Method	**2%**	**5%**	**10%**	**15%**	**20%**	**50%**	**Mean**
TruthfulQA							
SparseEval (Linear)	15.04 ±5.08	28.42 ±3.90	18.03 ±5.94	16.60 ±3.22	15.08 ±5.39	19.15 ±5.62	18.72 ±6.29
SparseEval (MLP)	9.21 ±3.15	10.38 ±1.97	9.66 ±3.54	9.83 ±3.86	9.72 ±3.59	9.52 ±2.94	9.72 ±2.74
Anchor Points	7.15 ±1.00	7.81 ±0.87	4.81 ±0.64	5.24 ±0.08	2.70 ±0.33	2.35 ±0.17	5.01 ±2.16
TinyBenchmarks	6.21 ±1.15	4.50 ±1.04	3.03 ±0.24	2.25 ±0.28	2.27 ±0.37	1.26 ±0.26	3.25 ±1.79
ZipBench	**5.36** ±1.05	4.53 ±0.85	2.62 ±0.75	3.65 ±1.65	2.73 ±0.87	1.13 ±0.36	3.34 ±1.65
Random	8.33 ±3.52	5.12 ±2.20	3.52 ±1.52	2.79 ±1.20	2.35 ±1.01	1.17 ±0.51	3.88 ±3.02
BestRand (ours)	6.34 ±1.96	**3.85** ±1.13	**2.60** ±0.74	**2.00** ±0.53	**1.64** ±0.42	**0.79** ±0.18	**2.87** ±2.08
BoolQ							
SparseEval (Linear)	15.53 ±4.48	21.38 ±4.02	15.34 ±2.54	16.23 ±3.09	16.56 ±3.33	29.22 ±19.94	19.04 ±8.97
SparseEval (MLP)	20.51 ±14.05	24.82 ±8.90	26.04 ±3.12	26.04 ±2.83	22.15 ±13.50	23.02 ±8.99	23.76 ±8.37
Anchor Points	8.86 ±1.22	10.64 ±1.24	15.29 ±6.83	5.98 ±0.57	8.65 ±4.85	5.80 ±2.80	9.20 ±4.52
TinyBenchmarks	4.44 ±0.55	2.63 ±0.57	**1.29** ±0.23	**1.00** ±0.45	**0.78** ±0.26	0.47 ±0.02	1.77 ±1.46
ZipBench	3.66 ±0.42	2.61 ±0.05	1.39 ±0.14	1.56 ±0.19	0.97 ±0.22	0.60 ±0.07	1.80 ±1.09
Random	3.51 ±1.11	2.17 ±0.69	1.50 ±0.48	1.19 ±0.38	1.00 ±0.31	0.50 ±0.16	1.64 ±1.15
BestRand (ours)	**3.21** ±0.86	**1.99** ±0.53	1.37 ±0.36	1.07 ±0.28	0.90 ±0.23	**0.45** ±0.11	**1.50** ±1.01
IFEval							
SparseEval (Linear)	32.75 ±32.08	21.84 ±6.06	21.50 ±2.41	21.18 ±4.79	17.15 ±3.31	28.37 ±15.80	23.80 ±13.71
SparseEval (MLP)	14.45 ±2.31	18.21 ±5.69	11.18 ±4.12	12.93 ±2.27	11.23 ±5.81	11.82 ±6.12	13.30 ±4.67
Anchor Points	**8.86** ±3.45	8.36 ±1.57	7.27 ±1.82	7.57 ±1.37	7.99 ±1.27	4.60 ±0.61	7.44 ±2.13
TinyBenchmarks	10.56 ±0.92	6.08 ±0.33	4.69 ±1.54	3.31 ±0.22	**2.49** ±0.18	1.63 ±0.45	4.79 ±3.11
ZipBench	10.38 ±1.44	8.12 ±1.74	5.12 ±0.20	3.61 ±0.20	3.60 ±0.39	1.85 ±0.14	5.45 ±3.11
Random	10.41 ±2.53	6.49 ±1.62	4.47 ±1.11	3.54 ±0.89	2.97 ±0.75	1.48 ±0.37	4.89 ±3.22
BestRand (ours)	9.55 ±1.96	**5.92** ±1.21	**4.06** ±0.82	**3.20** ±0.64	2.67 ±0.53	**1.32** ±0.26	**4.45** ±2.88
Finance-Fundamentals							
SparseEval (Linear)	25.66 ±6.07	19.15 ±5.22	22.12 ±1.99	16.94 ±2.13	19.21 ±1.26	36.02 ±26.96	23.18 ±11.71
SparseEval (MLP)	**8.90** ±1.29	8.72 ±0.82	8.37 ±1.09	7.09 ±1.14	6.34 ±0.35	8.74 ±1.79	8.03 ±1.40
Anchor Points	9.90 ±5.05	9.57 ±3.61	8.09 ±2.43	8.49 ±3.44	8.16 ±4.27	6.25 ±2.63	8.41 ±3.33
TinyBenchmarks	11.84 ±1.85	6.58 ±0.65	5.53 ±1.71	4.45 ±0.68	2.81 ±0.63	1.71 ±0.33	5.49 ±3.49
ZipBench	9.52 ±0.39	7.11 ±1.43	4.90 ±0.67	4.13 ±0.19	3.36 ±0.50	1.93 ±0.27	5.16 ±2.65
Random	11.25 ±3.35	7.00 ±2.11	4.81 ±1.45	3.81 ±1.15	3.20 ±0.97	1.60 ±0.48	5.28 ±3.63
BestRand (ours)	9.99 ±2.31	**6.17** ±1.39	**4.23** ±0.93	**3.30** ±0.69	**2.75** ±0.57	**1.37** ±0.28	**4.64** ±3.06
LIFBench							
SparseEval (Linear)	22.59 ±8.91	17.68 ±6.25	39.51 ±36.09	32.91 ±23.26	27.01 ±26.44	21.75 ±17.14	26.91 ±20.12
SparseEval (MLP)	13.20 ±7.24	8.33 ±1.14	8.89 ±1.49	10.34 ±1.55	11.39 ±1.15	12.29 ±1.14	10.74 ±3.22
Anchor Points	2.84 ±0.26	1.94 ±0.56	1.65 ±0.76	1.34 ±0.16	1.70 ±0.62	1.04 ±0.35	1.75 ±0.72
TinyBenchmarks	7.22 ±1.70	6.24 ±0.24	5.67 ±0.43	6.46 ±0.47	6.67 ±0.17	6.55 ±0.39	6.47 ±0.81
ZipBench	3.06 ±1.44	1.58 ±0.46	1.11 ±0.06	0.88 ±0.19	0.94 ±0.07	0.48 ±0.21	1.34 ±1.01
Random	2.80 ±0.91	1.74 ±0.56	1.20 ±0.39	0.95 ±0.31	0.80 ±0.26	0.40 ±0.13	1.32 ±0.92
BestRand (ours)	**2.36** ±0.58	**1.45** ±0.36	**1.00** ±0.24	**0.77** ±0.18	**0.64** ±0.15	**0.31** ±0.07	**1.09** ±0.74
GPQA							
SparseEval (Linear)	19.39 ±2.81	23.64 ±6.78	20.13 ±1.11	20.74 ±4.52	21.69 ±1.96	34.14 ±13.55	23.29 ±7.60
SparseEval (MLP)	**6.60** ±1.15	6.85 ±0.66	7.57 ±1.40	6.67 ±0.61	5.91 ±0.82	5.42 ±0.52	6.50 ±1.04
Anchor Points	14.17 ±2.34	8.78 ±1.89	8.47 ±2.81	8.87 ±1.13	10.12 ±1.92	7.41 ±1.31	9.64 ±2.79
TinyBenchmarks	13.69 ±3.37	9.34 ±2.44	**3.71** ±0.43	**3.40** ±1.10	3.25 ±1.33	2.32 ±0.47	5.95 ±4.55
ZipBench	12.59 ±4.88	8.60 ±0.96	5.41 ±0.65	4.88 ±1.12	3.80 ±0.49	2.32 ±0.55	6.27 ±3.94
Random	10.85 ±2.39	6.87 ±1.49	4.68 ±1.01	3.71 ±0.81	3.12 ±0.68	1.56 ±0.34	5.13 ±3.29
BestRand (ours)	10.56 ±2.25	**6.64** ±1.36	4.53 ±0.93	3.58 ±0.74	**3.00** ±0.60	**1.49** ±0.30	**4.97** ±3.19

Table 13: Multimodal benchmark MAE at 2–50% retention. Values are mean ± sample standard deviation in percent; lower is better. Best means are **bold** and second-best means are underlined.

	Average MAE (%) ↓ mean ± std.						
Method	**2%**	**5%**	**10%**	**15%**	**20%**	**50%**	**Mean**
MathVerse							
SparseEval (Linear)	42.87 ±30.75	33.96 ±26.14	42.56 ±27.64	44.15 ±17.58	30.57 ±23.97	16.27 ±4.08	35.06 ±22.11
SparseEval (MLP)	18.20 ±4.98	15.64 ±3.33	17.82 ±1.94	13.49 ±2.57	10.64 ±2.59	13.95 ±1.79	14.96 ±3.72
Anchor Points	9.14 ±1.59	7.81 ±1.58	7.46 ±1.23	7.81 ±1.84	7.63 ±1.61	4.41 ±0.87	7.38 ±1.93
TinyBenchmarks	9.54 ±0.11	5.57 ±0.42	3.79 ±0.98	3.40 ±1.05	2.54 ±0.16	1.26 ±0.47	4.35 ±2.79
ZipBench	**7.13** ±1.16	5.30 ±1.12	4.66 ±0.80	3.45 ±0.90	3.05 ±0.28	1.41 ±0.37	4.17 ±1.99
Random	8.90 ±2.19	5.57 ±1.40	3.81 ±0.96	3.02 ±0.76	2.54 ±0.64	1.27 ±0.32	4.19 ±2.76
BestRand (ours)	8.27 ±1.70	**5.12** ±1.04	**3.48** ±0.71	**2.74** ±0.55	**2.28** ±0.45	**1.14** ±0.22	**3.84** ±2.49
OmniDocBench							
SparseEval (Linear)	20.11 ±6.80	20.37 ±12.29	13.22 ±2.91	12.83 ±8.10	9.04 ±5.33	9.32 ±3.95	14.15 ±7.71
SparseEval (MLP)	2.59 ±1.30	2.03 ±0.91	2.75 ±1.92	2.14 ±0.80	2.53 ±1.54	2.24 ±1.50	2.38 ±1.19
Anchor Points	1.66 ±0.20	**0.88** ±0.38	0.90 ±0.25	1.37 ±0.79	1.03 ±0.20	0.60 ±0.05	1.07 ±0.48
TinyBenchmarks	2.85 ±0.09	2.34 ±1.27	2.74 ±0.50	2.82 ±0.44	2.73 ±0.25	2.52 ±0.11	2.67 ±0.54
ZipBench	**1.37** ±0.69	1.31 ±0.46	**0.63** ±0.10	0.61 ±0.09	0.57 ±0.24	0.37 ±0.17	0.81 ±0.50
Random	2.03 ±0.92	1.27 ±0.57	0.87 ±0.40	0.69 ±0.31	0.58 ±0.26	0.29 ±0.13	0.96 ±0.76
BestRand (ours)	1.70 ±0.65	1.06 ±0.40	0.72 ±0.27	**0.57** ±0.21	**0.47** ±0.17	**0.23** ±0.08	**0.79** ±0.59
SpatialEval							
SparseEval (Linear)	18.82 ±5.34	12.37 ±4.15	19.93 ±3.49	17.77 ±4.96	17.27 ±6.06	15.49 ±9.10	16.94 ±5.49
SparseEval (MLP)	25.25 ±10.16	33.02 ±6.53	23.33 ±12.23	24.39 ±12.19	30.81 ±4.04	21.17 ±10.90	26.33 ±9.31
Anchor Points	4.97 ±0.33	5.93 ±2.25	6.09 ±2.49	4.92 ±2.53	5.80 ±0.64	3.60 ±1.64	5.22 ±1.80
TinyBenchmarks	3.99 ±0.53	1.54 ±0.50	1.11 ±0.20	0.84 ±0.13	**0.61** ±0.28	0.57 ±0.08	1.44 ±1.25
ZipBench	2.33 ±1.00	1.75 ±0.50	**0.96** ±0.65	0.85 ±0.24	0.81 ±0.26	0.64 ±0.17	1.22 ±0.78
Random	2.62 ±1.06	1.63 ±0.66	1.12 ±0.45	0.89 ±0.36	0.75 ±0.30	0.37 ±0.15	1.23 ±0.93
BestRand (ours)	**2.30** ±0.83	**1.42** ±0.52	0.98 ±0.35	**0.77** ±0.27	0.64 ±0.23	**0.32** ±0.11	**1.07** ±0.79
LogicVista							
SparseEval (Linear)	28.92 ±3.49	21.78 ±3.79	25.59 ±5.46	35.28 ±5.20	29.37 ±7.28	26.44 ±5.22	27.90 ±6.11
SparseEval (MLP)	11.20 ±1.89	10.27 ±1.57	9.12 ±0.96	9.24 ±2.57	9.72 ±1.42	8.34 ±0.56	9.65 ±1.66
Anchor Points	11.41 ±1.66	7.81 ±1.91	8.10 ±0.64	6.20 ±0.79	7.52 ±2.76	4.49 ±1.59	7.59 ±2.60
TinyBenchmarks	11.62 ±1.77	8.34 ±0.79	**5.00** ±0.66	**3.82** ±0.32	3.45 ±0.13	2.54 ±0.07	5.79 ±3.36
ZipBench	**9.72** ±0.39	8.02 ±0.68	6.37 ±0.35	5.26 ±0.46	4.08 ±1.03	1.68 ±0.61	5.85 ±2.74
Random	12.81 ±2.98	7.96 ±1.90	5.44 ±1.29	4.33 ±1.02	3.65 ±0.86	1.81 ±0.43	6.00 ±3.93
BestRand (ours)	11.86 ±2.34	**7.40** ±1.47	5.06 ±0.98	4.01 ±0.77	**3.34** ±0.63	**1.64** ±0.30	**5.55** ±3.55
RealWorldQA							
SparseEval (Linear)	17.28 ±2.79	24.08 ±14.00	19.67 ±4.21	17.69 ±4.59	20.10 ±7.48	14.89 ±9.57	18.95 ±7.39
SparseEval (MLP)	19.57 ±14.34	26.67 ±8.86	27.42 ±11.33	26.42 ±11.46	17.26 ±9.83	21.77 ±9.79	23.18 ±10.13
Anchor Points	8.40 ±2.87	10.46 ±4.50	6.68 ±2.63	5.14 ±1.46	9.16 ±3.79	8.60 ±3.43	8.07 ±3.26
TinyBenchmarks	9.62 ±6.00	5.09 ±1.68	3.37 ±1.12	**2.03** ±0.25	3.30 ±2.57	1.91 ±1.00	4.22 ±3.60
ZipBench	8.82 ±3.05	6.37 ±0.83	3.66 ±1.25	2.99 ±1.79	**1.78** ±0.91	2.14 ±1.26	4.30 ±2.95
Random	7.90 ±3.31	4.91 ±2.00	3.35 ±1.38	2.65 ±1.11	2.24 ±0.92	1.12 ±0.46	3.70 ±2.83
BestRand (ours)	**7.24** ±2.79	**4.54** ±1.66	**3.06** ±1.12	2.42 ±0.93	2.02 ±0.75	**1.00** ±0.37	**3.38** ±2.52
OCRBench							
SparseEval (Linear)	16.95 ±5.37	15.14 ±1.55	12.74 ±3.40	33.01 ±22.07	22.23 ±10.23	20.70 ±1.75	20.13 ±11.00
SparseEval (MLP)	6.84 ±5.75	9.07 ±8.89	7.67 ±8.39	6.47 ±6.18	5.38 ±5.46	3.88 ±1.00	6.55 ±5.69
Anchor Points	3.13 ±1.35	2.07 ±0.65	2.67 ±0.59	2.55 ±1.74	3.87 ±2.26	2.82 ±1.06	2.85 ±1.31
TinyBenchmarks	1.72 ±0.09	2.32 ±0.30	1.81 ±0.19	2.20 ±0.12	2.21 ±0.22	1.81 ±0.13	2.01 ±0.29
ZipBench	1.83 ±0.44	1.07 ±0.09	0.59 ±0.16	**0.39** ±0.14	0.39 ±0.21	0.29 ±0.09	0.76 ±0.59
Random	1.86 ±0.92	1.15 ±0.57	0.80 ±0.39	0.63 ±0.31	0.53 ±0.26	0.27 ±0.13	0.87 ±0.72
BestRand (ours)	**1.38** ±0.50	**0.85** ±0.29	**0.57** ±0.19	0.44 ±0.14	**0.36** ±0.11	**0.17** ±0.05	**0.63** ±0.47

Table 14: Agent benchmark MAE at 2–50% retention. Values are mean ± sample standard deviation in percent; lower is better. Best means are **bold** and second-best means are underlined.

	Average MAE (%) ↓ mean ± std.						
Method	**2%**	**5%**	**10%**	**15%**	**20%**	**50%**	**Mean**
Drafter-Bench							
SparseEval (Linear)	23.58 ±11.47	27.37 ±13.35	41.79 ±42.16	39.31 ±40.39	17.49 ±5.47	14.58 ±4.61	27.35 ±23.54
SparseEval (MLP)	11.92 ±13.27	11.08 ±12.11	12.94 ±15.06	13.69 ±14.96	13.94 ±16.42	13.01 ±15.91	12.77 ±12.39
Anchor Points	3.06 ±1.67	2.44 ±0.40	2.12 ±0.66	1.88 ±0.30	1.48 ±0.12	1.13 ±0.07	2.02 ±0.91
TinyBenchmarks	6.84 ±2.08	3.66 ±0.62	3.69 ±0.55	3.79 ±0.70	3.88 ±0.67	3.26 ±0.32	4.19 ±1.50
ZipBench	3.96 ±0.75	1.61 ±0.16	**1.09** ±0.16	**0.84** ±0.21	**0.53** ±0.12	0.45 ±0.07	1.41 ±1.27
Random	2.91 ±0.85	1.80 ±0.53	1.24 ±0.36	0.98 ±0.29	0.83 ±0.24	0.41 ±0.12	1.36 ±0.93
BestRand (ours)	**2.60** ±0.60	**1.60** ±0.36	1.10 ±0.25	0.87 ±0.19	0.72 ±0.15	**0.36** ±0.07	**1.21** ±0.80
OSWorld							
SparseEval (Linear)	14.41 ±3.28	17.37 ±2.78	18.95 ±5.56	19.52 ±5.86	16.75 ±7.35	16.70 ±1.96	17.28 ±4.43
SparseEval (MLP)	15.82 ±3.77	14.88 ±2.14	16.93 ±3.68	14.58 ±3.25	12.86 ±0.73	9.89 ±1.44	14.16 ±3.29
Anchor Points	13.59 ±3.67	7.78 ±1.33	6.16 ±2.24	6.26 ±1.78	6.38 ±1.76	3.21 ±0.72	7.23 ±3.71
TinyBenchmarks	12.84 ±1.28	8.77 ±1.02	5.56 ±1.23	4.02 ±1.47	4.48 ±0.95	1.83 ±0.39	6.25 ±3.82
ZipBench	13.74 ±2.73	**5.50** ±2.27	5.18 ±1.22	3.59 ±1.19	3.84 ±0.93	1.84 ±0.68	5.62 ±4.18
Random	12.78 ±3.97	7.85 ±2.44	5.38 ±1.68	4.26 ±1.34	3.58 ±1.12	1.78 ±0.56	5.94 ±4.17
BestRand (ours)	**10.71** ±2.56	6.57 ±1.51	**4.44** ±0.98	**3.49** ±0.76	**2.87** ±0.61	**1.40** ±0.29	**4.91** ±3.31
BFCL							
SparseEval (Linear)	24.27 ±4.48	18.69 ±1.99	25.90 ±9.78	17.41 ±1.88	19.55 ±1.36	20.48 ±4.04	21.05 ±5.14
SparseEval (MLP)	14.63 ±7.26	13.73 ±8.13	13.29 ±7.43	11.69 ±6.03	12.56 ±8.35	12.68 ±5.74	13.10 ±6.14
Anchor Points	7.39 ±1.82	4.66 ±1.04	5.60 ±1.08	5.17 ±0.26	5.98 ±1.15	4.09 ±0.70	5.48 ±1.43
TinyBenchmarks	3.25 ±1.09	**1.77** ±0.14	1.73 ±0.73	1.30 ±0.56	1.05 ±0.26	0.66 ±0.04	1.62 ±0.98
ZipBench	3.58 ±0.52	2.27 ±0.55	1.57 ±0.46	1.20 ±0.34	**0.80** ±0.25	0.61 ±0.05	1.67 ±1.09
Random	3.58 ±1.34	2.22 ±0.83	1.53 ±0.58	1.21 ±0.46	1.02 ±0.39	0.51 ±0.19	1.68 ±1.24
BestRand (ours)	**2.98** ±0.81	1.84 ±0.49	**1.25** ±0.32	**0.97** ±0.24	0.80 ±0.19	**0.39** ±0.09	**1.37** ±0.95
TauBench							
SparseEval (Linear)	17.81 ±3.91	36.39 ±33.85	24.90 ±5.52	19.37 ±4.48	20.51 ±1.95	20.72 ±3.45	23.28 ±13.63
SparseEval (MLP)	**14.85** ±1.05	15.01 ±2.02	12.98 ±1.97	13.04 ±3.90	10.59 ±2.17	10.16 ±1.69	12.77 ±2.73
Anchor Points	22.44 ±2.50	14.86 ±2.42	10.95 ±4.18	7.95 ±1.23	7.32 ±1.55	6.93 ±0.95	11.74 ±6.01
TinyBenchmarks	22.46 ±5.51	13.57 ±3.74	9.68 ±3.10	7.51 ±2.17	6.63 ±1.51	2.98 ±0.53	10.47 ±6.96
ZipBench	18.39 ±5.60	15.18 ±4.23	11.08 ±3.24	6.96 ±2.07	7.77 ±2.09	3.56 ±0.96	10.49 ±5.93
Random	21.55 ±6.52	13.02 ±3.80	8.62 ±2.57	6.91 ±2.05	5.84 ±1.74	2.90 ±0.86	9.81 ±6.99
BestRand (ours)	19.16 ±5.39	**11.72** ±2.74	**7.64** ±1.71	**6.11** ±1.33	**5.10** ±1.07	**2.52** ±0.51	**8.71** ±6.06
NexusBench							
SparseEval (Linear)	28.04 ±23.24	22.53 ±4.64	23.56 ±7.45	22.86 ±5.36	20.68 ±2.42	21.42 ±3.79	23.18 ±9.18
SparseEval (MLP)	11.39 ±1.05	9.70 ±0.65	8.51 ±0.97	8.16 ±0.18	7.57 ±0.48	7.43 ±0.79	8.79 ±1.55
Anchor Points	10.94 ±2.70	8.46 ±2.62	8.96 ±4.66	10.36 ±2.03	9.41 ±2.13	6.35 ±0.86	9.08 ±2.76
TinyBenchmarks	**5.51** ±0.78	**3.82** ±0.36	**2.90** ±0.83	2.56 ±0.45	2.32 ±0.67	1.45 ±0.16	**3.09** ±1.42
ZipBench	8.96 ±2.41	4.64 ±1.11	3.75 ±0.51	2.64 ±0.29	2.03 ±0.14	1.39 ±0.57	3.90 ±2.75
Random	8.20 ±2.38	5.02 ±1.48	3.45 ±1.02	2.74 ±0.81	2.30 ±0.68	1.15 ±0.34	3.81 ±2.63
BestRand (ours)	7.21 ±1.73	4.40 ±1.07	2.99 ±0.72	**2.34** ±0.56	**1.95** ±0.46	**0.97** ±0.22	3.31 ±2.24
ComplexFuncBench							
SparseEval (Linear)	31.30 ±14.50	37.30 ±12.37	16.94 ±2.07	19.40 ±3.77	18.43 ±3.17	35.76 ±28.61	26.52 ±14.79
SparseEval (MLP)	13.50 ±0.98	12.03 ±3.50	10.72 ±3.50	10.82 ±2.75	12.35 ±3.56	11.30 ±1.94	11.79 ±2.61
Anchor Points	10.36 ±3.87	10.21 ±3.02	10.88 ±3.14	10.20 ±1.22	9.08 ±1.73	6.88 ±1.02	9.60 ±2.55
TinyBenchmarks	7.59 ±1.17	**3.45** ±0.41	3.56 ±0.56	**2.42** ±0.64	2.27 ±0.66	1.39 ±0.30	**3.45** ±2.13
ZipBench	8.33 ±1.87	7.10 ±0.80	3.69 ±0.65	3.30 ±0.86	2.57 ±0.28	1.54 ±0.32	4.42 ±2.64
Random	8.15 ±2.07	5.05 ±1.30	3.48 ±0.90	2.76 ±0.71	2.31 ±0.60	1.16 ±0.30	3.82 ±2.54
BestRand (ours)	**7.50** ±1.57	4.63 ±0.98	**3.17** ±0.65	2.50 ±0.50	**2.09** ±0.42	**1.04** ±0.20	3.49 ±2.26

Table 15: Compared with 15 baselines, BestRand delivers the strongest overall performance, achieving the lowest average MAE. Each entry reports the average MAE (%) and sample standard deviation across benchmarks; the final column aggregates all benchmark–budget combinations. Lower is better; the best and second-best results are shown in **bold** and underlined, respectively.

	Average MAE (%) ↓ mean ± std.						
Method	**2%**	**5%**	**10%**	**15%**	**20%**	**50%**	**Mean**
SparseEval (Linear)	19.73 ±6.78	21.72 ±5.86	24.98 ±8.09	18.65 ±5.32	19.43 ±5.88	21.39 ±7.08	20.98 ±6.67
SparseEval (MLP)	11.94 ±6.77	11.72 ±5.06	11.30 ±4.28	10.32 ±4.44	10.81 ±4.47	10.90 ±5.64	11.17 ±5.03
DatBench	14.70 ±5.29	12.12 ±5.22	11.11 ±5.12	9.57 ±3.75	8.44 ±2.78	6.54 ±2.02	10.41 ±4.86
Anchor Points	7.85 ±2.60	8.04 ±1.51	6.99 ±1.27	7.32 ±2.30	7.71 ±2.73	5.25 ±2.15	7.19 ±2.29
EssenceBench	10.00 ±6.47	8.14 ±6.33	8.06 ±5.31	6.75 ±4.12	6.22 ±4.67	2.97 ±2.30	7.02 ±5.35
MIS	9.71 ±5.68	7.27 ±4.53	5.70 ±4.09	5.30 ±4.99	4.30 ±4.47	2.70 ±2.33	5.83 ±4.85
Agreement	9.28 ±5.43	7.64 ±6.03	5.69 ±4.73	4.75 ±4.24	4.38 ±3.86	2.00 ±1.93	5.62 ±4.99
Coverage	7.82 ±3.49	6.16 ±3.70	6.18 ±4.47	5.90 ±3.85	5.30 ±3.29	2.09 ±1.25	5.57 ±3.79
Facility Location	8.41 ±3.83	5.68 ±3.24	5.58 ±3.67	5.00 ±3.24	4.15 ±3.09	1.78 ±2.18	5.10 ±3.71
MINCE	6.91 ±4.14	3.60 ±2.10	2.76 ±1.62	2.25 ±1.21	1.98 ±1.20	0.88 ±0.49	3.06 ±2.81
TinyBenchmarks	6.07 ±3.88	4.19 ±2.56	2.76 ±1.63	2.32 ±1.41	1.93 ±1.12	1.01 ±0.64	3.04 ±2.67
SubLIME	6.86 ±3.55	3.97 ±2.28	2.64 ±1.47	2.11 ±1.18	1.77 ±1.10	0.82 ±0.43	3.03 ±2.72
FlashEval	6.76 ±3.71	3.84 ±2.39	2.56 ±1.57	2.15 ±1.30	1.89 ±1.11	0.92 ±0.49	3.02 ±2.74
Random	6.18 ±3.26	3.83 ±2.03	2.63 ±1.39	2.09 ±1.10	1.76 ±0.93	0.88 ±0.46	2.89 ±2.44
ZipBench	5.84 ±3.85	3.66 ±2.37	**2.23** ±1.30	**1.77** ±0.97	1.60 ±0.93	1.01 ±0.65	2.68 ±2.54
BestRand (ours)	**5.45** ±3.06	**3.36** ±1.90	2.29 ±1.30	1.81 ±1.03	**1.50** ±0.86	**0.74** ±0.43	**2.52** ±2.23

E.3 Comparison with a Broader Set of Baselines

Table 1 in Section 3 compares BestRand with Random, two SparseEval variants, Anchor Points, TinyBenchmarks, and ZipBench across 70 benchmarks. Appendix C.3 further evaluates Facility Location, MIS, Agreement, and Coverage. Here, we combine these methods with five additional baselines. They are SubLIME (Saranathan et al., 2025), MINCE (Das et al., 2026), EssenceBench (Wang et al., 2026), FlashEval (Zhao et al., 2024), and DatBench (Joshi et al., 2026). We evaluate all 15 baselines and BestRand on 12 benchmarks. All methods use the same data splits and retention ratios.

The five additional baselines cover a diverse set of compression strategies. SubLIME predicts how well a subset preserves model rankings using anchor model responses and properties of benchmark items. MINCE uses Monte Carlo calibration to estimate the score drift of random subsets. EssenceBench combines semantic and response filtering with score prediction and subset search. FlashEval searches for subsets that preserve model rankings. DatBench selects items according to the point-biserial correlation between item responses and full benchmark scores.

Table 15 summarizes the results across 12 benchmarks and six retention ratios. Against 15 baselines, BestRand achieves the lowest overall MAE and ranks first or second at every retention ratio. Its advantage therefore holds from aggressive to mild compression. Figure 11 further breaks down the comparison by benchmark. BestRand remains among the strongest methods across individual benchmarks, with its error decreasing steadily as more items are retained. These consistent trends show that the aggregate advantage holds broadly rather than arising from a few favorable benchmarks. This comparison complements our main study on 70 text, multimodal, and agent benchmarks, as well as our extension to VLA evaluation. Across the full study, we consider a wide range of retention ratios, evaluation metrics, model splits, construction budgets, and distribution shifts. We also compare methods from many baseline families and examine average accuracy, ranking fidelity, variation across seeds, high-error risk, construction cost, and OOD generalization. Additional ablations, data-only experiments, and multi-metric evaluations test the method under further settings and design choices. Together, these experiments provide evidence at the aggregate, category, and individual-benchmark levels. BestRand consistently delivers accurate and stable estimates at low construction cost across this extensive evaluation. The scale and diversity of the results show that its gains persist across benchmarks, compression settings, evaluation criteria, and competing methods.

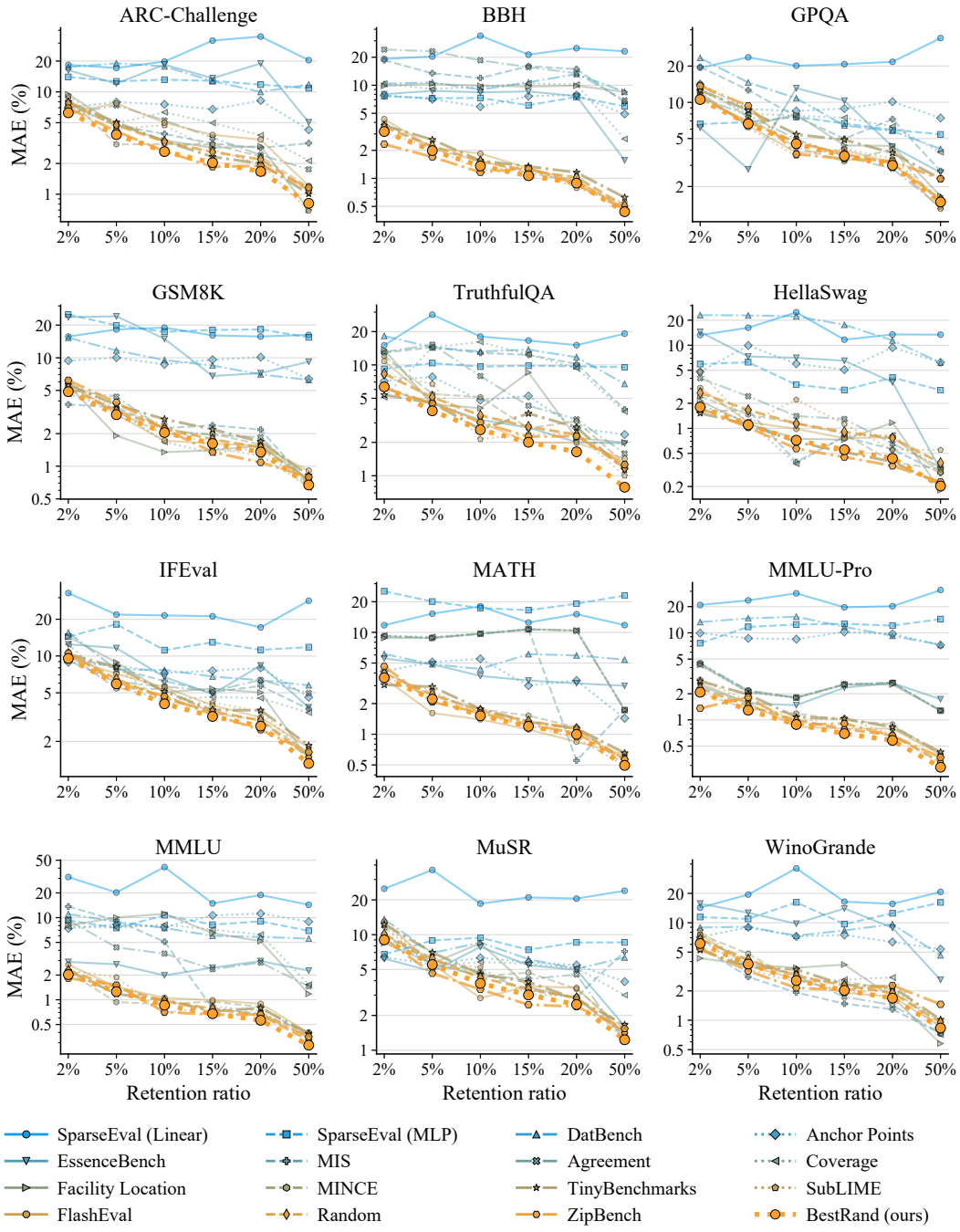

Figure 11: MAE of BestRand and 15 baselines across six retention ratios on 12 benchmarks. Each panel shows one benchmark, with MAE plotted on a logarithmic scale. BestRand consistently achieves the lowest or near-lowest error across benchmarks and retention ratios.

Table 16: Stability of Random and BestRand on the 12 benchmarks. Entries report mean MAE and sample standard deviation across seeds. Reduction rows show BestRand's relative reductions in mean MAE and standard deviation compared with Random, formatted as mean/std. Random and BestRand use 10,000 seeds. Across all settings, BestRand achieves both a lower mean MAE and a lower standard deviation than Random, demonstrating consistently better accuracy and stability.

		Mean Absolute Error (%) ↓ mean ± std.					
Benchmark	**Method**	**2%**	**5%**	**10%**	**15%**	**20%**	**50%**
ARC-Challenge	Random	7.70 ±2.89	4.73 ±1.78	3.27 ±1.23	2.59 ±0.98	2.18 ±0.82	1.09 ±0.41
	BestRand	**6.25** ±1.66	**3.82** ±1.00	**2.61** ±0.64	**2.03** ±0.49	**1.67** ±0.38	**0.81** ±0.18
	Reduction	**18.9%/42.5%**	**19.1%/43.6%**	**20.2%/47.8%**	**21.5%/49.9%**	**23.6%/53.5%**	**25.2%/56.9%**
BBH	Random	3.56 ±0.92	2.22 ±0.57	1.53 ±0.39	1.21 ±0.31	1.02 ±0.26	0.51 ±0.13
	BestRand	**3.19** ±0.68	**1.99** ±0.42	**1.36** ±0.28	**1.07** ±0.22	**0.89** ±0.18	**0.44** ±0.09
	Reduction	**10.5%/26.4%**	**10.4%/26.6%**	**11.0%/29.1%**	**11.8%/30.0%**	**12.5%/32.9%**	**13.8%/34.2%**
GPQA	Random	10.85 ±2.39	6.87 ±1.49	4.68 ±1.01	3.71 ±0.81	3.12 ±0.68	1.56 ±0.34
	BestRand	**10.56** ±2.25	**6.64** ±1.36	**4.53** ±0.93	**3.58** ±0.74	**3.00** ±0.60	**1.49** ±0.30
	Reduction	**2.7%/5.8%**	**3.4%/8.8%**	**3.1%/8.6%**	**3.4%/8.3%**	**4.1%/11.5%**	**4.4%/11.1%**
GSM8K	Random	5.27 ±1.56	3.25 ±0.97	2.24 ±0.67	1.78 ±0.53	1.49 ±0.45	0.74 ±0.22
	BestRand	**4.87** ±1.23	**2.99** ±0.75	**2.04** ±0.50	**1.62** ±0.39	**1.35** ±0.32	**0.67** ±0.16
	Reduction	**7.6%/21.6%**	**8.0%/22.7%**	**8.7%/25.3%**	**8.7%/26.4%**	**9.3%/27.7%**	**9.8%/28.9%**
HellaSwag	Random	2.67 ±1.50	1.66 ±0.93	1.14 ±0.64	0.91 ±0.51	0.76 ±0.43	0.38 ±0.21
	BestRand	**1.81** ±0.74	**1.10** ±0.45	**0.72** ±0.27	**0.55** ±0.19	**0.44** ±0.14	**0.20** ±0.06
	Reduction	**32.2%/50.6%**	**33.4%/51.8%**	**37.0%/57.7%**	**39.1%/61.9%**	**42.8%/67.7%**	**46.8%/74.0%**
IFEval	Random	10.41 ±2.53	6.49 ±1.62	4.47 ±1.11	3.54 ±0.89	2.97 ±0.75	1.48 ±0.37
	BestRand	**9.55** ±1.96	**5.92** ±1.21	**4.06** ±0.82	**3.20** ±0.64	**2.67** ±0.53	**1.32** ±0.26
	Reduction	**8.2%/22.5%**	**8.8%/25.0%**	**9.1%/26.7%**	**9.6%/28.3%**	**10.2%/29.5%**	**10.8%/30.3%**
MATH	Random	3.97 ±1.18	2.45 ±0.73	1.70 ±0.50	1.34 ±0.39	1.13 ±0.33	0.56 ±0.17
	BestRand	**3.58** ±0.85	**2.21** ±0.53	**1.52** ±0.35	**1.20** ±0.27	**1.00** ±0.22	**0.49** ±0.11
	Reduction	**9.8%/28.2%**	**9.8%/26.8%**	**10.6%/30.2%**	**10.6%/30.9%**	**11.6%/33.1%**	**12.4%/33.5%**
MMLU	Random	2.27 ±0.75	1.41 ±0.47	0.97 ±0.32	0.77 ±0.26	0.65 ±0.21	0.32 ±0.11
	BestRand	**2.02** ±0.53	**1.26** ±0.33	**0.86** ±0.22	**0.68** ±0.17	**0.56** ±0.14	**0.28** ±0.07
	Reduction	**10.8%/28.8%**	**11.1%/29.6%**	**11.7%/32.1%**	**12.3%/33.5%**	**13.2%/34.8%**	**13.4%/36.2%**
MMLU-Pro	Random	2.31 ±0.75	1.44 ±0.47	0.99 ±0.32	0.78 ±0.25	0.66 ±0.21	0.33 ±0.11
	BestRand	**2.08** ±0.55	**1.29** ±0.34	**0.89** ±0.23	**0.70** ±0.18	**0.58** ±0.15	**0.29** ±0.07
	Reduction	**10.2%/27.2%**	**10.1%/26.9%**	**10.5%/28.5%**	**11.3%/29.8%**	**11.9%/31.7%**	**13.3%/35.3%**
MuSR	Random	10.03 ±2.88	6.16 ±1.79	4.24 ±1.24	3.38 ±0.99	2.84 ±0.83	1.42 ±0.42
	BestRand	**9.01** ±2.05	**5.51** ±1.25	**3.79** ±0.84	**3.01** ±0.66	**2.49** ±0.53	**1.23** ±0.26
	Reduction	**10.1%/28.9%**	**10.6%/30.2%**	**10.7%/32.5%**	**10.9%/33.4%**	**12.3%/36.2%**	**13.1%/37.9%**
TruthfulQA	Random	8.33 ±3.52	5.12 ±2.20	3.52 ±1.52	2.79 ±1.20	2.35 ±1.01	1.17 ±0.51
	BestRand	**6.34** ±1.96	**3.85** ±1.13	**2.60** ±0.74	**2.00** ±0.53	**1.64** ±0.42	**0.79** ±0.18
	Reduction	**23.9%/44.3%**	**24.8%/48.5%**	**26.3%/51.3%**	**28.2%/55.8%**	**30.1%/58.6%**	**32.9%/63.6%**
WinoGrande	Random	6.77 ±1.98	4.20 ±1.23	2.87 ±0.84	2.28 ±0.67	1.91 ±0.56	0.96 ±0.28
	BestRand	**6.12** ±1.44	**3.77** ±0.89	**2.55** ±0.58	**2.03** ±0.47	**1.69** ±0.38	**0.83** ±0.18
	Reduction	**9.6%/27.1%**	**10.1%/28.2%**	**10.9%/31.2%**	**11.0%/30.5%**	**11.9%/32.8%**	**13.0%/36.0%**

Table 16 presents the accuracy and stability of Random and BestRand. Each entry reports the mean and standard deviation of MAE across independent seeds. BestRand achieves both a lower mean error and a lower standard deviation than Random in every setting, demonstrating consistently higher accuracy and greater stability across benchmarks and compression levels. In particular, the lower standard deviation demonstrates that BestRand is consistently more stable across random seeds. These results are consistent with Theorem 3, which shows that probe-guided matching reduces the probe-explained component of the common-shift variance. The consistent improvements across all settings provide strong empirical support for the proposed variance-reduction mechanism.

F Full Proofs and Derivations

F.1 Notation and basic conventions

Fix the score matrix $X \in [0,1]^{N\times M}$, where $X_{s,m}$ is the score of model m on item s over N items and M models. In the finite-population analysis, X is non-random and randomness comes only from subset sampling. A single subset $\mathcal{S}$ is drawn uniformly without replacement from $\{1,\dots,N\}$, with $|\mathcal{S}| = k$ and $1 \le k \le N$, where $N > 1$. The non-degenerate distributional analysis below concerns $k < N$; at $k = N$, all errors vanish. The selection analysis additionally considers multiple candidate draws. Following the main text,

$$\mu_m = \frac{1}{N}\sum_{s=1}^{N} X_{s,m}, \quad \hat{\mu}_m = \frac{1}{k}\sum_{s\in\mathcal{S}} X_{s,m}, \quad E_m = \hat{\mu}_m - \mu_m, \quad \mathrm{MAE} = \frac{1}{M}\sum_{m=1}^{M}|E_m|, \tag{12}$$

and the errors are collected into $E = (E_1,\dots,E_M)^\top$. Item difficulty and its empirical mean and variance are

$$g_s = \frac{1}{M}\sum_{m=1}^{M} X_{s,m}, \qquad \bar{g}_N = \frac{1}{N}\sum_{s=1}^{N} g_s, \qquad \mathrm{Var}_N(g) = \frac{1}{N}\sum_{s=1}^{N}\big(g_s - \bar{g}_N\big)^2. \tag{13}$$

The common shift is $\bar{E} = \frac{1}{M}\sum_m E_m$ and the finite-population factor is $c(k) = \sqrt{(N-k)/[k(N-1)]} > 0$ with $c(N) = 0$.

Two variance symbols must be kept apart. A subscript N denotes an empirical average over the fixed population of N items and carries no randomness, whereas $\mathrm{Var}(\cdot)$ and $\mathrm{Cov}(\cdot)$ without a subscript denote sampling variance and covariance taken as the subset $\mathcal{S}$ is redrawn. We write Φ for the standard normal cumulative distribution function and $\varphi(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$ for its density.

Let $z_s = X_{s,\cdot} - \mu \in \mathbb{R}^M$ be the s-th row of X with the column means $\mu = (\mu_1,\dots,\mu_M)^\top$ removed. Because each column is centered,

$$\sum_{s=1}^{N} z_s = \sum_{s=1}^{N} X_{s,\cdot} - N\mu = 0. \tag{A.1}$$

In line with Assumption 1, the population score covariance matrix is

$$C_{ml} = \frac{1}{N}\sum_{s=1}^{N}\big(X_{s,m}-\mu_m\big)\big(X_{s,l}-\mu_l\big), \qquad C = \frac{1}{N}\sum_{s=1}^{N} z_s z_s^\top, \tag{A.2}$$

which is positive semidefinite since $a^\top C a = \frac{1}{N}\sum_s (a^\top z_s)^2 \ge 0$ for every $a \in \mathbb{R}^M$. Finally, let $I_s = \mathbf{1}\{s \in \mathcal{S}\}$ be the sampling indicators, so that $\sum_{s=1}^{N} I_s = k$.

The matrix C is the population score covariance, whereas the sampling error covariance is $\mathrm{Cov}(E) = c(k)^2 C$. Theorem 2 uses C_0 for the scalar constant $C_0 = \sqrt{\pi/2-1}\,u^* \approx 0.578$. The subscript distinguishes it from the covariance matrix C. To keep the limiting-density prefactor separate from both, Appendix F.6 denotes it by A_0. Exact finite-population identities are distinguished below from distributional conclusions obtained within the Gaussian approximation of Theorem 1.

F.2 Numerical validation of the assumptions

The empirical checks concern how well Assumptions 1 and 2 approximate the benchmark data. Assumption 1 posits equal positive diagonal entries σ^2 and equal off-diagonal entries $\rho\sigma^2$ in C, with $\rho \in [0,1)$. The empirical summaries used in the main text are $\sigma^2 = M^{-1}\sum_m C_{mm}$ and, for $M > 1$, $\rho = [M(M-1)]^{-1}\sum_{m\neq l} C_{ml}/\sigma^2$. Under equal variances, the latter is the average pairwise correlation; with unequal variances it is a normalized average covariance. Assumption 2 requires $\mathrm{Var}_N(g) > 0$ and positive explained variance $R^2 > 0$ in a least-squares regression with an intercept. The normal equations ensure that the residual has zero empirical mean and is orthogonal to the probes; they do not require a perfect linear fit or independence of residuals and probe scores.

Table 17: Empirical results strongly support Assumption 1. Across diverse benchmarks, model-score variances are highly similar, while pairwise model-score covariances are overwhelmingly positive and close to a common value. A_{diag} measures variance similarity, P_+ is the proportion of positive pairwise covariances, and A_{off} measures covariance similarity. All are percentages; higher values indicate closer agreement with the assumption, and 100% indicates perfect agreement.

Benchmark	$\mathbf{A}_{\text{diag}}\uparrow$	$\mathbf{P}_+\uparrow$	$\mathbf{A}_{\text{off}}\uparrow$	**Benchmark**	$\mathbf{A}_{\text{diag}}\uparrow$	$\mathbf{P}_+\uparrow$	$\mathbf{A}_{\text{off}}\uparrow$
ACEBench	98.8	100.0	99.2	MathVista	94.9	100.0	95.3
AGC-Bench	98.7	100.0	94.9	MedQA	99.3	100.0	92.8
ARC-Challenge	99.5	100.0	94.1	MMMU	97.4	100.0	95.3
BBEH	92.5	100.0	87.7	MMMU-Pro	98.9	100.0	92.0
BBH	99.8	96.4	85.1	MMStar	98.2	100.0	95.9
Bench2Drive	99.2	100.0	95.6	MMVP	96.9	100.0	95.1
BFCL	99.1	97.1	93.2	MultiChallenge	99.4	100.0	94.5
BigCodeBench	95.2	100.0	87.9	MuSR	99.9	99.8	93.2
BoolQ	95.6	97.9	91.8	NexusBench	99.5	100.0	90.6
CommonsenseQA	98.8	97.9	86.8	OCRBench v2	99.7	100.0	99.5
ComplexFuncBench	96.5	99.5	88.6	OpenBookQA	96.4	97.2	72.3
CyberSecEval 2	100.0	100.0	98.5	OSWorld	95.6	99.4	86.9
DrafterBench	94.1	99.5	94.2	PAWS	97.3	97.9	92.4
DynaMath	98.8	100.0	92.7	Pre-Flight	99.6	100.0	97.7
Finance Fund.	99.7	100.0	95.9	RACE-H	95.0	100.0	97.1
GSM-Symbolic	94.1	100.0	90.4	RealWorldQA	97.9	100.0	92.0
HellaSwag	99.7	100.0	97.9	SciBench	98.8	100.0	92.8
HumanEval	97.1	100.0	92.0	SciCode	96.7	100.0	95.6
IFEval	96.8	97.5	86.0	tau-bench	99.0	99.7	91.1
InFoBench	99.3	100.0	98.8	Terminal-Bench	98.0	99.3	90.9
LiveCodeBench	99.0	100.0	96.3	TheoremQA	98.8	100.0	94.2
LogicVista	99.7	99.6	89.6	Toolathlon	94.2	98.1	87.4
LogiQA	98.5	100.0	97.7	TruthfulQA	99.6	100.0	95.7
LongBench v2	99.3	95.6	86.9	VisualPuzzles	99.8	100.0	94.9
MathVerse	96.8	98.0	84.3	WinoGrande	98.0	97.7	90.6
Mean:	$A_{\text{diag}} = 97.9\%$		$P_+ = 99.4\%$		$A_{\text{off}} = 92.4\%$		

Validation of Assumption 1. We validate Assumption 1 by testing the equal-variance and equal-covariance conditions directly from the observed model scores. For each benchmark, we remove only models whose scores are constant across all items, as required by Assumption 1. We then center each remaining model over the N items and form the empirical covariance matrix

$$C = \frac{1}{N} Z^\top Z, \qquad Z_{sm} = X_{s,m} - \mu_m.$$

Let $d = (C_{11}, \ldots, C_{MM})$ contain the score variances of the M models. Let o contain the $K = M(M-1)/2$ covariances between unordered model pairs. We compute

$$A_{\text{diag}} = 100 \frac{\langle d, \mathbf{1} \rangle}{\|d\|_2 \sqrt{M}}, \qquad P_+ = \frac{100}{K} \sum_{m<l} \mathbf{1}\{C_{ml} > 0\}, \qquad A_{\text{off}} = 100 \frac{\langle o, \mathbf{1} \rangle}{\|o\|_2 \sqrt{K}}.$$

A_{diag} measures variance similarity. P_+ is the proportion of model pairs with positive covariance. A_{off} measures covariance similarity. The two alignment scores are cosine similarities to a positive constant vector. They reach 100% only when the corresponding entries are equal and positive. Higher values therefore indicate closer agreement with Assumption 1.

Table 17 shows that score variances are highly similar across models and that almost all model pairs have positive covariance. The pairwise covariance values also remain close to a common level for most benchmarks. This agreement in both variance and covariance strongly supports Assumption 1 as a compact approximation of the observed covariance structure.

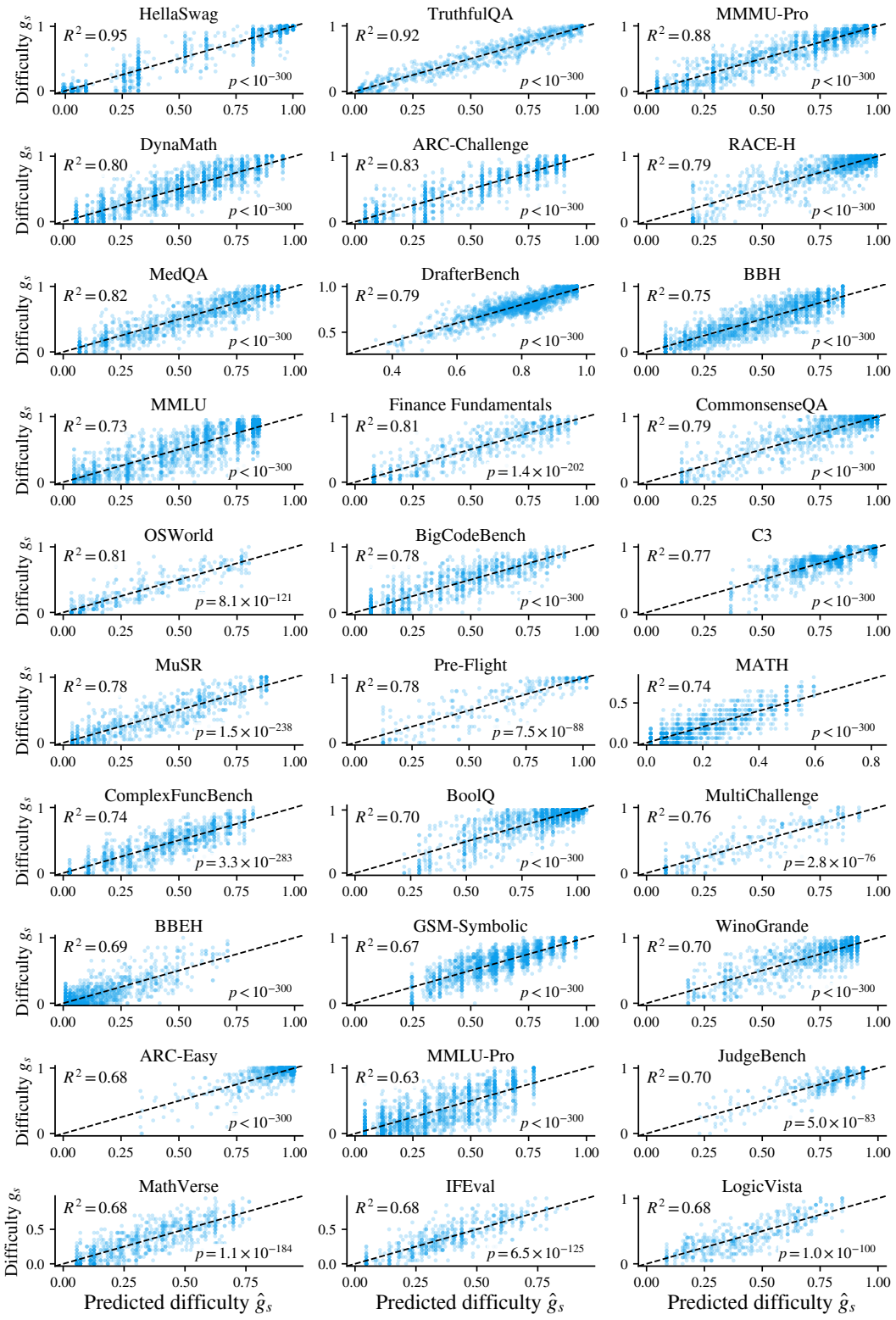

Figure 12: Item-level numerical validation on 30 benchmarks using ten probe models. The panels show a subset of the evaluated benchmarks, while aggregate results across all benchmarks are reported in Figure 13. Across the displayed benchmarks, R^2 is consistently well above zero, all p-values are small, and every regression is statistically significant. The dashed line denotes equality.

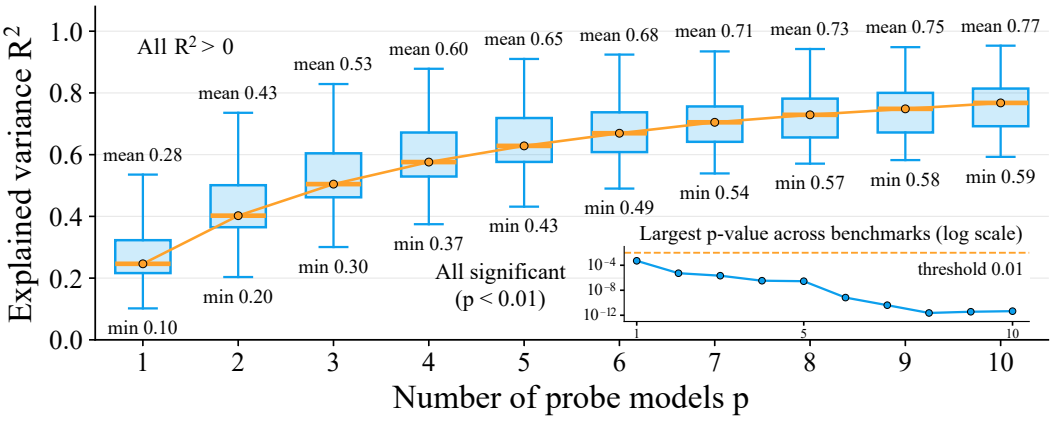

Figure 13: Numerical experiments consistently validate Assumption 2. Each box summarizes the distribution of explained variance R^2 across benchmarks for a given number of probe models, with whiskers spanning the full observed range. The small plot in the lower-right corner reports the largest benchmark-level p-value at each probe count. All R^2 values are positive and all benchmark-level results are statistically significant, providing strong empirical support for the assumption.

Validation of Assumption 2. We test whether probe scores explain item difficulty measured across the test models, which are disjoint from the probe models. Let $\mathcal{T}$ be the set of test models and $\mathcal{P}_p$ be a set of p probe models, with $\mathcal{T} \cap \mathcal{P}_p = \varnothing$. We define item difficulty as $g_s = |\mathcal{T}|^{-1} \sum_{m \in \mathcal{T}} X_{s,m}$ and vary the number of probes from one to ten. Since no probe score enters g_s, the measured association cannot result from overlap between the predictors and the response. A positive R^2 therefore shows that the probes capture difficulty shared with unseen models.

For each probe set, we regress g_s on the probe scores by ordinary least squares with an intercept. Repeated probe choices are summarized within each benchmark, and all benchmarks receive equal weight. We report R^2 and use the largest p-value across benchmark summaries as a conservative significance check. Figure 12 shows that predictions from ten probes closely track the difficulty defined by the disjoint test models at the item level. The consistent alignment with the equality line shows that the explained variance reflects a broad item-level relation rather than a small number of items. Figure 13 further shows that R^2 increases steadily with the number of probes, with smaller gains at larger probe counts. All R^2 values are positive, all benchmark summaries are significant, and most R^2 values are well above zero. Together, these results strongly validate Assumption 2 across the evaluated benchmarks. By Theorem 3, the positive explained variance allows probe matching to reduce common-shift variance, predicting greater stability for BestRand than for Random.

F.3 Basic lemmas: moments of sampling without replacement

The two lemmas below are the only source of every variance and covariance computed later.

Lemma B.1. Under uniform sampling without replacement, for any $s \neq t$,

$$\mathbb{E}[I_s] = \frac{k}{N}, \qquad \mathrm{Var}(I_s) = \frac{k(N-k)}{N^2}, \qquad \mathrm{Cov}(I_s, I_t) = -\frac{k(N-k)}{N^2(N-1)}. \tag{B.1}$$

Proof. Every item is equally likely to be chosen and every pair is equally likely to be chosen together,

$$\Pr(s \in \mathcal{S}) = \frac{k}{N}, \qquad \Pr(s, t \in \mathcal{S}) = \frac{k(k-1)}{N(N-1)} \quad (s \neq t).$$

The first probability holds because the k chosen positions are spread uniformly over N items, and the second because fixing two given items to be chosen amounts to choosing the remaining $k-2$ from the other $N-2$, with probability $\binom{N-2}{k-2}/\binom{N}{k} = \frac{k(k-1)}{N(N-1)}$. Hence $\mathbb{E}[I_s] = k/N$, and since $I_s \in \{0,1\}$ gives $I_s^2 = I_s$,

$$\mathrm{Var}(I_s) = \mathbb{E}[I_s] - \mathbb{E}[I_s]^2 = \frac{k}{N} - \left(\frac{k}{N}\right)^2 = \frac{k(N-k)}{N^2}.$$

For $s \neq t$,

$$\mathrm{Cov}(I_s, I_t) = \Pr(s, t \in \mathcal{S}) - \Pr(s \in \mathcal{S}) \Pr(t \in \mathcal{S}) = \frac{k(k-1)}{N(N-1)} - \frac{k^2}{N^2}.$$

Over a common denominator the numerator is $k[(k-1)N - k(N-1)] = k[kN - N - kN + k] = k(k-N) = -k(N-k)$, which gives $\mathrm{Cov}(I_s, I_t) = -k(N-k)/[N^2(N-1)]$. ■

A compact form that we use repeatedly aligns the variance and the covariance,

$$\mathrm{Cov}(I_s, I_t) = \frac{k(N-k)}{N^2}\Big(\delta_{st} - \frac{1-\delta_{st}}{N-1}\Big), \qquad \delta_{st} = \mathbf{1}\{s = t\}. \tag{B.2}$$

Lemma B.2. Let $h, h' : \{1, \dots, N\} \to \mathbb{R}$ be any two per-item scalars, write the subset-minus-population mean deviation as $\Delta\bar{h} := \bar{h}_{\mathcal{S}} - \bar{h}_N$ with $\bar{h}_{\mathcal{S}} = \frac{1}{k}\sum_{s\in\mathcal{S}} h_s$ and $\bar{h}_N = \frac{1}{N}\sum_s h_s$, and let $\mathrm{Cov}_N(h, h') = \frac{1}{N}\sum_s (h_s - \bar{h}_N)(h'_s - \bar{h}'_N)$. Then

$$\mathbb{E}[\Delta\bar{h}] = 0, \qquad \mathrm{Cov}(\Delta\bar{h}, \Delta\bar{h}') = c(k)^2 \,\mathrm{Cov}_N(h, h'). \tag{B.3}$$

In particular $\mathrm{Var}(\Delta\bar{h}) = c(k)^2 \,\mathrm{Var}_N(h)$.

Proof. Without loss of generality take h, h' centered, so that $\bar{h}_N = \bar{h}'_N = 0$, since replacing h_s by $h_s - \bar{h}_N$ leaves $\Delta\bar{h}$ unchanged. Then $\Delta\bar{h} = \bar{h}_{\mathcal{S}} = \frac{1}{k}\sum_s I_s h_s$. From $\mathbb{E}[I_s] = k/N$,

$$\mathbb{E}[\Delta\bar{h}] = \frac{1}{k}\sum_{s=1}^{N} \mathbb{E}[I_s]\, h_s = \frac{1}{k}\cdot\frac{k}{N}\sum_{s=1}^{N} h_s = \bar{h}_N = 0.$$

Because the mean is zero,

$$\mathrm{Cov}(\Delta\bar{h}, \Delta\bar{h}') = \mathbb{E}\Big[\Big(\tfrac{1}{k}\sum_s I_s h_s\Big)\Big(\tfrac{1}{k}\sum_t I_t h'_t\Big)\Big] = \frac{1}{k^2}\sum_{s,t}\mathbb{E}[I_s I_t]\, h_s h'_t.$$

Writing $\mathbb{E}[I_s I_t] = \mathrm{Cov}(I_s, I_t) + \mathbb{E}[I_s]\mathbb{E}[I_t]$ and using $\sum_s h_s = \sum_t h'_t = 0$, the product-of-means term $\frac{k^2}{N^2}(\sum_s h_s)(\sum_t h'_t)$ vanishes and only the covariance term survives. Substituting (B.2),

$$\mathrm{Cov}(\Delta\bar{h}, \Delta\bar{h}') = \frac{1}{k^2}\sum_{s,t}\frac{k(N-k)}{N^2}\Big(\delta_{st} - \frac{1-\delta_{st}}{N-1}\Big) h_s h'_t.$$

The diagonal contributes $\sum_s h_s h'_s$, and the off-diagonal contributes

$$\sum_{s\neq t} h_s h'_t = \Big(\sum_s h_s\Big)\Big(\sum_t h'_t\Big) - \sum_s h_s h'_s = -\sum_s h_s h'_s.$$

The bracket therefore equals $\big(1 + \frac{1}{N-1}\big)\sum_s h_s h'_s = \frac{N}{N-1}\sum_s h_s h'_s$, and

$$\mathrm{Cov}(\Delta\bar{h}, \Delta\bar{h}') = \frac{1}{k^2}\cdot\frac{k(N-k)}{N^2}\cdot\frac{N}{N-1}\sum_s h_s h'_s = \frac{N-k}{k(N-1)}\cdot\frac{1}{N}\sum_s h_s h'_s.$$

Here $\frac{1}{N}\sum_s h_s h'_s = \mathrm{Cov}_N(h, h')$ for centered h, h' and $\frac{N-k}{k(N-1)} = c(k)^2$, which gives (B.3). ■

The factor $c(k)^2 = \frac{N-k}{k(N-1)}$ has two parts, the usual $\frac{1}{k}$ dilution from averaging over k draws and the finite-population correction $\frac{N-k}{N-1}$ from sampling without replacement, which vanishes at $k = N$ where no fluctuation remains (Cochran, 1977).

F.4 Proof of Theorem 1

Theorem 1. Under Assumption 1, when the effective sample size $\min\{k, N-k\}$ is large,

$$E \;\dot{\sim}\; \mathcal{N}\big(0, c(k)^2 C\big), \qquad \mathrm{MAE} = \frac{1}{M}\sum_{m=1}^{M} |E_m|.$$

Under this Gaussian approximation, the MAE distribution is that of the average absolute value of correlated zero-mean Gaussian variables. The pointwise inequality $\mathrm{MAE} \geq |\bar{E}|$ and the approximation $\bar{E} \;\dot{\sim}\; \mathcal{N}(0, s^2)$ give, for every threshold $t > 0$,

$$\Pr(\mathrm{MAE} > t) \geq \Pr(|\bar{E}| > t) \approx 2\Phi(-t/s),$$

where $s = c(k)\sigma\sqrt{\rho + (1-\rho)/M}$ is the standard deviation of the common shift $\bar{E} = M^{-1}\sum_m E_m$.

The proof has four parts, the exact first and second moments of E, its multivariate normal approximation and the induced MAE distribution, the identity for the common shift together with a two-sided sandwich for the MAE, and the right-tail lower bound. The tail bound follows because $\mathrm{MAE} \geq |\bar{E}|$ holds pointwise and $\bar{E}$ is approximately $\mathcal{N}(0, s^2)$ under Assumption 1, so $\Pr(\mathrm{MAE} > t) \geq \Pr(|\bar{E}| > t) \approx 2\Phi(-t/s)$.

F.4.1 Exact first and second moments of the error vector

Proposition C.1. Under uniform sampling without replacement, for any finite N and $1 \leq k < N$,

$$\mathbb{E}[E] = 0, \qquad \mathrm{Cov}(E) = c(k)^2\, C. \tag{C.1}$$

This is exact and requires no distributional assumption.

Proof. Write E as a linear combination of the indicators. From $\hat{\mu}_m = \frac{1}{k}\sum_{s\in\mathcal{S}} X_{s,m} = \frac{1}{k}\sum_{s=1}^{N} I_s X_{s,m}$ and $X_{s,\cdot} = z_s + \mu$,

$$E = \frac{1}{k}\sum_{s=1}^{N} I_s X_{s,\cdot} - \mu = \frac{1}{k}\sum_{s=1}^{N} I_s z_s + \frac{1}{k}\Big(\sum_{s=1}^{N} I_s\Big)\mu - \mu.$$

Since $\sum_s I_s = k$, the middle term equals μ and cancels $-\mu$, which leaves the clean linear form

$$E = \frac{1}{k}\sum_{s=1}^{N} I_s z_s. \tag{C.2}$$

For the first moment, $\mathbb{E}[I_s] = k/N$ and $\sum_s z_s = 0$ from (A.1) give

$$\mathbb{E}[E] = \frac{1}{k}\sum_{s=1}^{N} \mathbb{E}[I_s] z_s = \frac{1}{k}\cdot\frac{k}{N}\sum_{s=1}^{N} z_s = 0.$$

For the second moment, since the mean is zero,

$$\mathrm{Cov}(E) = \mathbb{E}[EE^\top] = \frac{1}{k^2}\sum_{s,t}\mathbb{E}[I_s I_t]\, z_s z_t^\top = \frac{1}{k^2}\sum_{s,t}\mathrm{Cov}(I_s, I_t)\, z_s z_t^\top,$$

where the product-of-means term again vanishes because $\sum_s z_s = 0$. Splitting into $s = t$ and $s \neq t$ and using Lemma B.1,

$$\mathrm{Cov}(E) = \frac{1}{k^2}\left[\frac{k(N-k)}{N^2}\sum_{s} z_s z_s^\top - \frac{k(N-k)}{N^2(N-1)}\sum_{s\neq t} z_s z_t^\top\right].$$

With $\sum_{s\neq t} z_s z_t^\top = (\sum_s z_s)(\sum_t z_t)^\top - \sum_s z_s z_s^\top = -\sum_s z_s z_s^\top$, the bracket becomes $\frac{k(N-k)}{N^2}\big(1 + \frac{1}{N-1}\big)\sum_s z_s z_s^\top = \frac{k(N-k)}{N(N-1)}\sum_s z_s z_s^\top$, hence

$$\mathrm{Cov}(E) = \frac{N-k}{k(N-1)}\cdot\frac{1}{N}\sum_{s} z_s z_s^\top = c(k)^2\, C.$$

■

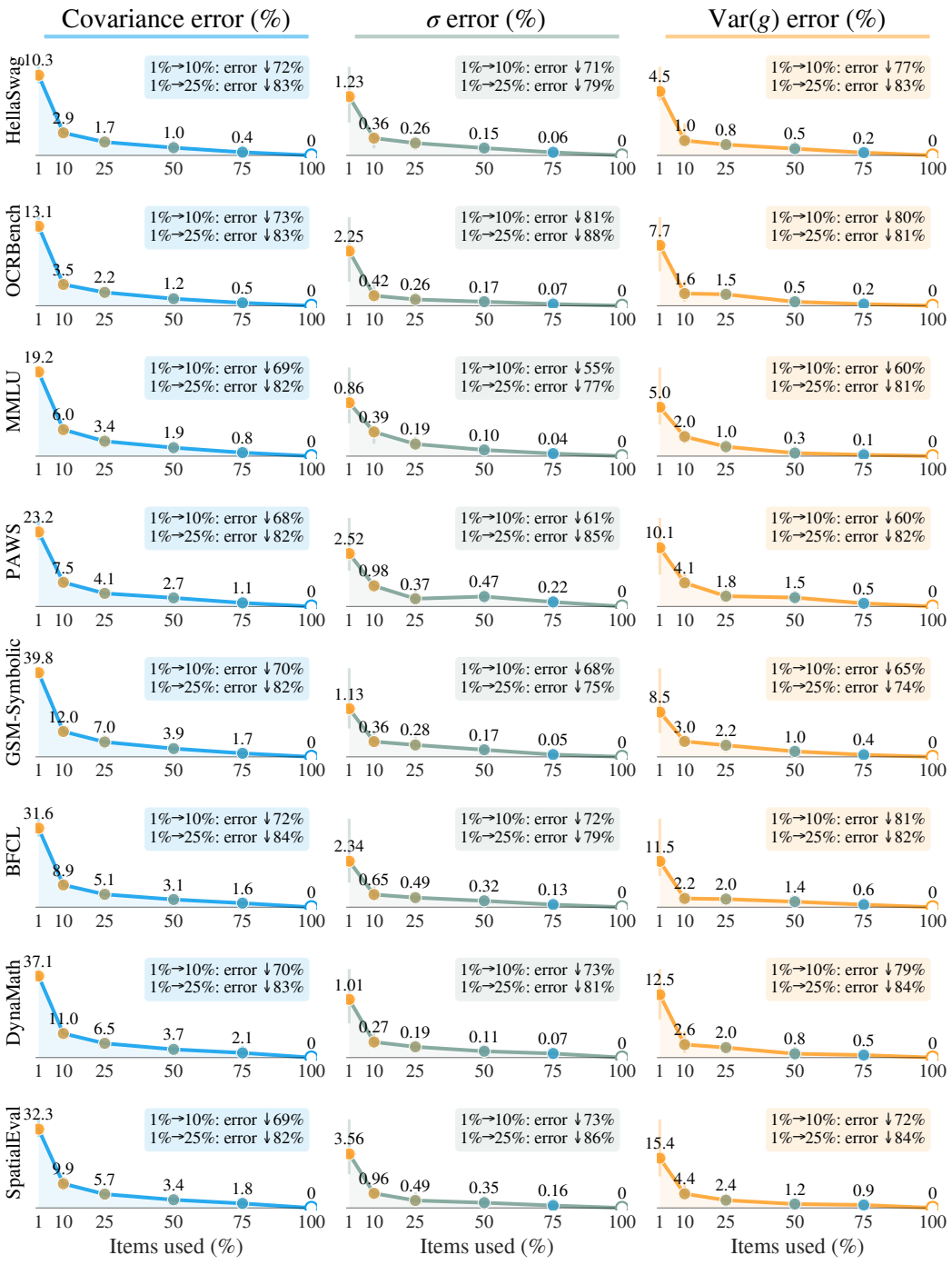

Figure 14: Finite sample stability of the empirical covariance and its derived summaries. Each benchmark uses the same model set across all item subsets. The quantities C_n, σ_n, and $\mathrm{Var}_n(g)$ are recomputed from nested random item subsets. Each point is the median over 30 random item orderings, and the vertical bars cover the middle half of the results. The text boxes report the error reduction from 1% to 10% and 25%. Across all eight benchmarks, the errors decrease sharply after only 10% of the items and become small by 25%, showing that partial item sets provide accurate estimates of the full benchmark covariance structure and its derived summaries.

F.4.2 MULTIVARIATE NORMAL APPROXIMATION

Proposition C.1 gives exact moments, but a statement about near normality needs convergence, which requires embedding the fixed population into a sequence indexed by a growing N.

For each N, fix a population $X_{N,1}, \ldots, X_{N,N} \in [0,1]^M$ and a subset size k_N, with the associated quantities carrying the subscript N, namely μ_N, $z_{N,s} = X_{N,s} - \mu_N$, $C_N = \frac{1}{N}\sum_s z_{N,s} z_{N,s}^\top$, $c_N = \sqrt{(N-k_N)/[k_N(N-1)]}$, and $E_N = \frac{1}{k_N}\sum_s I_s z_{N,s}$. Define the effective sample size $h_N = \min\{k_N, N-k_N\}$. We impose two conditions. First, $h_N \to \infty$, so both the retained and omitted item counts diverge. Second, $C_N \to \Sigma$ for a positive semidefinite Σ, possibly singular, which fixes the limiting covariance after scaling. These conditions specify the asymptotic regime underlying the normal approximation; a large effective sample size alone does not guarantee its accuracy for an arbitrary finite population (Bowyer et al., 2025).

The second condition above requires the score covariance to stabilize as N grows. We examine its finite sample behavior by fixing all models with nonconstant scores and recomputing the covariance from nested random subsets of n items. Figure 14 reports the relative Frobenius error of C_n and the relative errors of σ_n and $\mathrm{Var}_n(g)$, each measured against its full benchmark estimate. Here $\sigma_n^2 = M^{-1}\sum_m (C_n)_{mm}$. Across all eight benchmarks, the errors fall sharply when item coverage increases from 1% to 10%. By 25% of the items, all three estimates are close to their full benchmark values and improve further as more items are used. This consistent behavior provides empirical support for the covariance stability required by the asymptotic regime.

Proposition C.2. Under the framework above, if M is fixed and the two conditions hold, then

$$c_N^{-1} E_N \xrightarrow{d} W \sim \mathcal{N}_M(0, \Sigma). \tag{C.3}$$

Moreover, writing $\mathrm{MAE}_N = M^{-1}\sum_{m=1}^M |E_{N,m}|$, we have

$$c_N^{-1}\mathrm{MAE}_N \xrightarrow{d} \frac{1}{M}\sum_{m=1}^M |W_m|.$$

Hence the working Gaussian approximation for E also determines the approximate MAE distribution through the average of its absolute coordinates. Here W denotes only the limiting vector of the scaled errors in (C.3).

Proof. We use the Cramér–Wold device, so that convergence of every scalar projection $a^\top(c_N^{-1}E_N)$ to the matching one-dimensional normal implies convergence of the vector. Fix a direction $a \in \mathbb{R}^M$ and study the projection.

First, express the projection as a sample-mean deviation. Define the scalar population $y_{N,s} = a^\top z_{N,s}$, whose population mean is $\bar{y}_N = \frac{1}{N}\sum_s y_{N,s} = 0$ by $\sum_s z_{N,s} = 0$, and whose population variance normalized by $N-1$ is

$$v_{N,a} = \frac{1}{N-1}\sum_{s=1}^N (a^\top z_{N,s})^2 = \frac{N}{N-1}\, a^\top C_N a.$$

With $\bar{y}_{\mathcal{S}} = \frac{1}{k_N}\sum_{s\in\mathcal{S}} y_{N,s}$ and $E_N = \frac{1}{k_N}\sum_s I_s z_{N,s}$ from (C.2), the sample mean minus the population mean is the projection itself,

$$\bar{y}_{\mathcal{S}} - \bar{y}_N = a^\top E_N. \tag{C.4}$$

Second, treat a nondegenerate direction with $a^\top\Sigma a > 0$. Because $X_{N,s} \in [0,1]^M$ gives $z_{N,s} \in [-1,1]^M$, each term is bounded by $|y_{N,s}| = |a^\top z_{N,s}| \le \|a\|_1 < \infty$. The second condition gives $v_{N,a} \to a^\top\Sigma a > 0$ and the first gives $h_N \to \infty$, so

$$\frac{\max_s y_{N,s}^2}{h_N\, v_{N,a}} \le \frac{\|a\|_1^2}{h_N\, v_{N,a}} \longrightarrow 0,$$

which is the Lindeberg-type condition of the Hájek finite-population central limit theorem (Hájek, 1960; Li & Ding, 2017). That theorem gives, for simple random sampling without replacement,

$$\frac{\bar{y}_{\mathcal{S}} - \bar{y}_N}{\sqrt{(\frac{1}{k_N} - \frac{1}{N})v_{N,a}}} \xrightarrow{d} \mathcal{N}(0,1).$$

The variance in the denominator simplifies through $v_{N,a} = \frac{N}{N-1} a^\top C_N a$,

$$\Big(\frac{1}{k_N} - \frac{1}{N}\Big) v_{N,a} = \frac{N - k_N}{k_N N} \cdot \frac{N}{N-1}\, a^\top C_N a = c_N^2\, a^\top C_N a,$$

so with (C.4) it reads $a^\top E_N / (c_N \sqrt{a^\top C_N a}) \xrightarrow{d} \mathcal{N}(0,1)$. Since $a^\top C_N a \to a^\top \Sigma a$, Slutsky's theorem yields

$$c_N^{-1} a^\top E_N \xrightarrow{d} \mathcal{N}(0,\, a^\top \Sigma a). \tag{C.5}$$

Third, treat a degenerate direction with $a^\top \Sigma a = 0$, so that $a^\top C_N a \to 0$. By Proposition C.1, $\mathbb{E}[c_N^{-1} a^\top E_N] = 0$ and

$$\mathrm{Var}(c_N^{-1} a^\top E_N) = c_N^{-2}\, a^\top \mathrm{Cov}(E_N) a = a^\top C_N a \to 0.$$

Chebyshev's inequality then gives $c_N^{-1} a^\top E_N \xrightarrow{p} 0$, which is the normal limit $\mathcal{N}(0,0)$ and agrees with (C.5) at $a^\top \Sigma a = 0$.

Fourth, reassemble the vector. For every a we have $c_N^{-1} a^\top E_N \xrightarrow{d} \mathcal{N}(0, a^\top \Sigma a)$. Taking $W = \Sigma^{1/2} G$ with $G \sim \mathcal{N}_M(0, I_M)$ and $\Sigma^{1/2}$ the positive semidefinite square root, every projection satisfies $a^\top W \sim \mathcal{N}(0, a^\top \Sigma a)$, so all one-dimensional limits come from the single $W \sim \mathcal{N}_M(0, \Sigma)$. The Cramér–Wold device gives $c_N^{-1} E_N \xrightarrow{d} W \sim \mathcal{N}_M(0, \Sigma)$.

Finally, the map $v \mapsto M^{-1} \sum_{m=1}^M |v_m|$ is continuous on $\mathbb{R}^M$. Applying the continuous mapping theorem to (C.3) yields

$$\frac{\mathrm{MAE}_N}{c_N} = \frac{1}{M} \sum_{m=1}^M \left| \frac{E_{N,m}}{c_N} \right| \xrightarrow{d} \frac{1}{M} \sum_{m=1}^M |W_m|.$$

This establishes the induced MAE limit without assuming independence of the error coordinates. ■

Using the finite-N covariance C and scale $c(k)$ gives the working formulation of Theorem 1, in the asymptotic regime specified above:

$$E \;\dot\sim\; \mathcal{N}_M\big(0, c(k)^2 C\big), \qquad \mathrm{MAE} = \frac{1}{M} \sum_{m=1}^M |E_m|.$$

The MAE identity is exact; its distribution is approximated by evaluating this function under the joint Gaussian law for E. Each absolute coordinate then has a half-normal marginal, but the coordinates are generally dependent, so MAE is not itself asserted to be Gaussian.

F.4.3 The common-shift identity and the two-sided sandwich

Proposition C.3. The common shift satisfies the identity

$$\bar E = \bar g_{\mathcal{S}} - \bar g_N, \qquad \bar g_{\mathcal{S}} = \frac{1}{k} \sum_{s \in \mathcal{S}} g_s, \quad \bar g_N = \frac{1}{N} \sum_{s=1}^N g_s, \tag{C.6}$$

and the MAE is sandwiched by

$$|\bar E| \;\le\; \mathrm{MAE} \;\le\; |\bar E| + \mathrm{floor}, \qquad \mathrm{floor} := \frac{1}{M} \sum_{m=1}^M |R_m|, \quad R_m := E_m - \bar E. \tag{C.7}$$

Both are identities or exact inequalities and need no assumption.

Proof. For (C.6), substitute the definitions and exchange the order of summation,

$$\bar E = \frac{1}{M} \sum_{m=1}^M E_m = \frac{1}{k} \sum_{s \in \mathcal{S}} \frac{1}{M} \sum_{m=1}^M X_{s,m} - \frac{1}{N} \sum_{s=1}^N \frac{1}{M} \sum_{m=1}^M X_{s,m} = \bar g_{\mathcal{S}} - \bar g_N,$$

since the inner average over m is exactly g_s. The common shift is therefore the mean difficulty of the drawn subset minus that of the full set, which depends only on which items are drawn and acts identically on every model. For the lower bound, the triangle inequality gives $\mathrm{MAE} = \frac{1}{M}\sum_m |E_m| \geq |\frac{1}{M}\sum_m E_m| = |\bar{E}|$. For the upper bound, write $E_m = \bar{E} + R_m$ and use $|\bar{E} + R_m| \leq |\bar{E}| + |R_m|$, then average over m to obtain $\mathrm{MAE} \leq |\bar{E}| + \mathrm{floor}$. ■

The sandwich (C.7) shows that a large common shift necessarily produces a large MAE. The quantity floor is a random residual term, not a deterministic lower endpoint or a quantity known to remain fixed under selection. Thus the sandwich identifies the common shift as one source of high-error risk, but does not imply that it is the only source or that the entire MAE distribution is determined by s.

F.4.4 The scale of the common shift and the right-tail lower bound

Proposition C.4. The sampling variance of the common shift is

$$s^2 := \mathrm{Var}(\bar{E}) = c(k)^2\,\mathrm{Var}_N(g) = \frac{c(k)^2}{M^2}\,\mathbf{1}^\top C\,\mathbf{1}, \tag{C.8}$$

with $\mathbf{1} = (1,\ldots,1)^\top \in \mathbb{R}^M$. When the effective sample size is large, $\bar{E} \;\dot\sim\; \mathcal{N}(0, s^2)$ and therefore

$$\Pr(\mathrm{MAE} > t) \;\geq\; \Pr(|\bar{E}| > t) \;\approx\; 2\,\Phi\Big(-\frac{t}{s}\Big). \tag{C.9}$$

Proof. By (C.6), $\bar{E} = \bar{g}_{\mathcal{S}} - \bar{g}_N = \Delta\bar{g}$ is the sample-mean deviation of the per-item difficulty g_s, so Lemma B.2 with $h = h' = g$ gives $\mathrm{Var}(\bar{E}) = c(k)^2\mathrm{Var}_N(g)$. To express $\mathrm{Var}_N(g)$ through C, note $g_s - \bar{g}_N = \frac{1}{M}\sum_m (X_{s,m} - \mu_m) = \frac{1}{M}\mathbf{1}^\top z_s$, hence

$$\mathrm{Var}_N(g) = \frac{1}{N}\sum_{s=1}^{N}\Big(\frac{1}{M}\mathbf{1}^\top z_s\Big)^2 = \frac{1}{M^2}\,\mathbf{1}^\top\Big(\frac{1}{N}\sum_{s=1}^{N} z_s z_s^\top\Big)\mathbf{1} = \frac{1}{M^2}\,\mathbf{1}^\top C\,\mathbf{1}, \tag{C.10}$$

which is the mean of every entry of C, and (C.8) follows. For the tail, the lower bound $\mathrm{MAE} \geq |\bar{E}|$ of Proposition C.3 gives $\{|\bar{E}| > t\} \subseteq \{\mathrm{MAE} > t\}$, so $\Pr(\mathrm{MAE} > t) \geq \Pr(|\bar{E}| > t)$. The one-dimensional case of Proposition C.2 with $a = \frac{1}{M}\mathbf{1}$ makes $\bar{E}$ approximately $\mathcal{N}(0, s^2)$, so for $Z \sim \mathcal{N}(0, s^2)$,

$$\Pr(|\bar{E}| > t) \approx \Pr(|Z| > t) = 2\big(1 - \Phi(t/s)\big) = 2\,\Phi(-t/s).$$

■

Since $2\Phi(-t/s)$ increases with s, a larger scale raises the Gaussian approximation to the common-shift tail and hence the lower bound on high-error risk at any fixed threshold. This proves Theorem 1; it does not identify the lower bound with the actual MAE tail probability. Within the Gaussian model, $2\Phi(-t/s) \sim \sqrt{\frac{2}{\pi}}\,\frac{s}{t}\,e^{-t^2/(2s^2)}$ as $t/s \to \infty$. This asymptotic describes the Gaussian common-shift tail, not a relative-error guarantee for extreme tails of the finite-population distribution. In particular, "long-tailed" here describes the observed right-skewed error distributions, not a heavy-tailed asymptotic law. The empirical comparison discussed in the main text is Figure 1.

F.5 Closed forms under Assumption 1

This part inserts Assumption 1, the homogeneous variance and homogeneous pairwise correlation, and derives closed forms for the common-shift scale and the mean MAE, together with the single-common-factor representation used by Theorems 2 and 3.

F.5.1 The equicorrelation structure

Assumption 1 sets $C_{mm} = \sigma^2 > 0$ and $C_{ml} = \rho\sigma^2$ for $m \neq l$ with $\rho \in [0, 1)$, which is exactly

$$C = \sigma^2\big[(1-\rho)\,I + \rho\,\mathbf{1}\mathbf{1}^\top\big]. \tag{D.1}$$

The diagonal is $\sigma^2[(1-\rho)+\rho] = \sigma^2$ and each off-diagonal is $\sigma^2[\,0 + \rho\,] = \rho\sigma^2$, as required.

F.5.2 CLOSED FORM OF THE COMMON-SHIFT SCALE

Proposition D.2. Under Assumption 1,

$$s = c(k)\,\sigma\sqrt{\rho + \frac{1-\rho}{M}}. \tag{D.2}$$

Proof. From (C.8), $s^2 = \frac{c(k)^2}{M^2}\mathbf{1}^\top C\mathbf{1}$. With (D.1) and $\mathbf{1}^\top I\mathbf{1} = M$, $\mathbf{1}^\top\mathbf{1}\mathbf{1}^\top\mathbf{1} = M^2$,

$$\mathbf{1}^\top C\mathbf{1} = \sigma^2\big[(1-\rho)M + \rho M^2\big],$$

so $s^2 = \frac{c(k)^2}{M^2}\sigma^2[(1-\rho)M + \rho M^2] = c(k)^2\sigma^2\big(\rho + \frac{1-\rho}{M}\big)$, and taking the square root gives (D.2). ■

For $M > 1$, the squared scale increases with ρ because $\partial[\rho + (1-\rho)/M]/\partial\rho = 1 - 1/M > 0$. For fixed $k < N$, $\sigma > 0$, and $\rho > 0$, the expression tends to $c(k)\sigma\sqrt{\rho} > 0$ as M grows. Thus increasing the number of similarly behaving models does not remove the correlation-dependent component. At $\rho = 0$, by contrast, $s = c(k)\sigma/\sqrt{M}$ vanishes in this limit.

F.5.3 THE MEAN OF THE MAE

Proposition D.3. Under Assumption 1 and the normal approximation of Theorem 1,

$$\mathbb{E}[\mathrm{MAE}] \approx \sqrt{\tfrac{2}{\pi}}\,c(k)\,\sigma. \tag{D.3}$$

Proof. By Theorem 1 the marginal of each error is approximately $E_m \sim \mathcal{N}(0, c(k)^2\sigma^2)$, since the diagonal of C is σ^2. A zero-mean normal $Z \sim \mathcal{N}(0, \tau^2)$ has $\mathbb{E}|Z| = \sqrt{2/\pi}\,\tau$, which is the folded-normal mean of Lemma E.2 at zero mean, so with $\tau = c(k)\sigma$ each $\mathbb{E}|E_m| = \sqrt{2/\pi}\,c(k)\sigma$. Averaging the equal marginals over m leaves the value unchanged, giving (D.3). ■

Within the Gaussian approximation, the mean does not depend on ρ, since it is fixed by the single-model marginal scale $c(k)\sigma$. Dividing (D.2) by the Gaussian-model mean in (D.3) gives the common-shift scale relative to the mean error,

$$\frac{s}{\mathbb{E}[\mathrm{MAE}]} = \sqrt{\tfrac{\pi}{2}}\sqrt{\rho + \frac{1-\rho}{M}} \xrightarrow{M\to\infty} \sqrt{\tfrac{\pi}{2}}\,\sqrt{\rho}, \tag{D.4}$$

where σ and $c(k)$ cancel and only ρ and M remain.

F.5.4 THE SINGLE-COMMON-FACTOR REPRESENTATION

Proposition D.4. Within the Gaussian approximation of Theorem 1, Assumption 1 gives the representation

$$E_m \overset{d}{=} c(k)\,\sigma\big(\sqrt{\rho}\,G + \sqrt{1-\rho}\,Z_m\big), \qquad G, Z_1, \ldots, Z_M \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{D.5}$$

Proof. The right side is a linear combination of independent normals, so $(E_1, \ldots, E_M)$ is jointly normal and it suffices to match the mean and covariance with $\mathcal{N}(0, c(k)^2 C)$. The mean is zero, and the variance is

$$\mathrm{Var}(E_m) = c(k)^2\sigma^2\big(\rho + (1-\rho)\big) = c(k)^2\sigma^2 = c(k)^2 C_{mm}.$$

For $m \neq l$, only the shared G contributes, so

$$\mathrm{Cov}(E_m, E_l) = c(k)^2\sigma^2\,\rho = c(k)^2 C_{ml}.$$

Two zero-mean Gaussian vectors with equal covariance are identical in distribution. The equality in distribution in (D.5) is exact for this Gaussian model, not for the original finite-population sampling distribution. ■

The representation (D.5) separates the randomness into a common factor G, which reflects whether the drawn items are collectively harder or easier and is shared by all models, and an individual noise Z_m, which is the sampling luck of each model and is independent across models. This is the starting point of the asymptotic analysis of Theorem 2.

F.6 Proof of Theorem 2

Theorem 2. Under Assumption 1 with $0 < \rho < 1$ and the Gaussian approximation of Theorem 1, as $M \to \infty$ the MAE converges to a limiting distribution with lower endpoint $y_0 = \sqrt{2/\pi}\, c(k)\sigma\sqrt{1-\rho}$. The mode $\hat{y}(M)$ of the sampling distribution of the MAE approaches y_0 at the universal rate $\hat{y}(M)/y_0 - 1 = C_0/\sqrt{M} + O(1/M)$, with $C_0 = \sqrt{\pi/2-1}\,\arg\max_{u>0}\int_0^\infty v^{-1/2}\varphi(u-v)\,\mathrm{d}v \approx 0.578$, and the constant C_0 does not depend on $c(k), \sigma, \rho$. For brevity the proof writes $u^* = \arg\max_{u>0}\int_0^\infty v^{-1/2}\varphi(u-v)\,\mathrm{d}v \approx 0.765$ for the maximizer.

Throughout this part write $c := c(k) > 0$ and work within the Gaussian representation (D.5), holding c, $\sigma > 0$, and $0 < \rho < 1$ fixed as $M \to \infty$. This is a large-M analysis of the Gaussian model, separate from the fixed-M, large-N justification in Proposition C.2; it is not a joint-limit theorem for arbitrary growing score matrices. The remainder constants may depend on the fixed parameters, so the expansion is not uniform as $\rho \to 0$ or $\rho \to 1$. For $Z \sim \mathcal{N}(0,1)$ and a real argument g, define

$$r_\rho(g) := \mathbb{E}\big|\sqrt{\rho}\, g + \sqrt{1-\rho}\, Z\big|, \qquad v_\rho(g) := \mathrm{Var}\big(\big|\sqrt{\rho}\, g + \sqrt{1-\rho}\, Z\big|\big). \tag{E.0}$$

By (D.5), given the common factor $G = g$ the single-model absolute error is $|E_m| = c\sigma|\sqrt{\rho}\, g + \sqrt{1-\rho}\, Z_m|$, with conditional mean $c\sigma\, r_\rho(g)$ and conditional variance $c^2\sigma^2 v_\rho(g)$. Write $\mathrm{MAE}_M := \frac{1}{M}\sum_{m=1}^M |E_m|$.

F.6.1 Conditioning on the common factor

Lemma E.1. Given $G = g$,

$$\mathbb{E}[\mathrm{MAE}_M \mid G = g] = c\sigma\, r_\rho(g), \qquad \mathrm{Var}[\mathrm{MAE}_M \mid G = g] = \frac{c^2\sigma^2}{M}\, v_\rho(g). \tag{E.1}$$

Proof. Fix $G = g$. By (D.5) the $Z_1, \dots, Z_M$ remain independent and identically distributed given G, so the $|E_1|, \dots, |E_M|$ are independent and identically distributed with mean $c\sigma r_\rho(g)$ and variance $c^2\sigma^2 v_\rho(g)$. Averaging M such variables keeps the mean and divides the variance by M. ■

The randomness has two layers, an outer $G \sim \mathcal{N}(0,1)$ that decides how hard the draw is, and an inner average of M individual noises whose variance carries the factor $1/M$ and concentrates as M grows.

F.6.2 Closed form and local expansion of r_ρ

Lemma E.2. The function r_ρ is even and

$$r_\rho(g) = \sqrt{1-\rho}\,\sqrt{\tfrac{2}{\pi}}\, e^{-\frac{\rho g^2}{2(1-\rho)}} + \sqrt{\rho}\, g\,\mathrm{erf}\Big(\frac{\sqrt{\rho}\, g}{\sqrt{2(1-\rho)}}\Big). \tag{E.2}$$

In particular

$$r_\rho(0) = \sqrt{\tfrac{2}{\pi}}\sqrt{1-\rho}, \qquad r_\rho'(0) = 0, \qquad r_\rho''(0) = \sqrt{\tfrac{2}{\pi}}\,\frac{\rho}{\sqrt{1-\rho}} > 0, \tag{E.3}$$

so near $g = 0$,

$$r_\rho(g) = r_\rho(0) + \tfrac{1}{2} r_\rho''(0)\, g^2 + O(g^4), \tag{E.4}$$

and $g = 0$ is the unique strict minimum of r_ρ with a parabolic bottom.

Proof. Let $W = \sqrt{\rho}\, g + \sqrt{1-\rho}\, Z \sim \mathcal{N}(m, \varsigma^2)$ with $m = \sqrt{\rho}\, g$ and $\varsigma = \sqrt{1-\rho}$.

For the closed form, use the Gaussian identity $\mathbb{E}[\max(W,0)] = m\,\Phi(m/\varsigma) + \varsigma\,\varphi(m/\varsigma)$, obtained by substituting $u = (w-m)/\varsigma$ in $\int_0^\infty w\,\frac{1}{\varsigma}\varphi(\frac{w-m}{\varsigma})dw$ and integrating the resulting $u\varphi$ and φ parts. Since $|W| = 2\max(W,0) - W$ and $\mathbb{E}W = m$,

$$\mathbb{E}|W| = 2\big[m\Phi(m/\varsigma) + \varsigma\varphi(m/\varsigma)\big] - m = m\big[2\Phi(m/\varsigma) - 1\big] + 2\varsigma\varphi(m/\varsigma).$$

Using $2\Phi(z) - 1 = \operatorname{erf}(z/\sqrt{2})$ and $2\varsigma\varphi(m/\varsigma) = \varsigma\sqrt{2/\pi}\, e^{-m^2/2\varsigma^2}$,

$$\mathbb{E}|W| = \varsigma\sqrt{\tfrac{2}{\pi}}\, e^{-m^2/2\varsigma^2} + m \operatorname{erf}\Big(\frac{m}{\varsigma\sqrt{2}}\Big),$$

and substituting $m = \sqrt{\rho}\, g$ and $\varsigma = \sqrt{1-\rho}$ gives (E.2). Evenness follows from $r_\rho(-g) = \mathbb{E}|\sqrt{\rho}\, g - \sqrt{1-\rho}\, Z| = r_\rho(g)$ by the symmetry of Z.

For the first derivative, $\partial_g|\sqrt{\rho}\, g + \sqrt{1-\rho}\, Z| = \sqrt{\rho}\operatorname{sign}(W)$ away from a null set, so

$$r'_\rho(g) = \sqrt{\rho}\, \mathbb{E}\operatorname{sign}(W) = \sqrt{\rho}\big(1 - 2\Pr(W < 0)\big) = \sqrt{\rho}\operatorname{erf}\Big(\frac{\sqrt{\rho}\, g}{\sqrt{2(1-\rho)}}\Big), \tag{E.5}$$

and $r'_\rho(0) = \sqrt{\rho}\operatorname{erf}(0) = 0$. For the second derivative, differentiate (E.5) using $\frac{d}{du}\operatorname{erf}(u) = \frac{2}{\sqrt{\pi}} e^{-u^2}$ and the chain rule with inner derivative $\frac{\sqrt{\rho}}{\sqrt{2(1-\rho)}}$,

$$r''_\rho(g) = \sqrt{\rho} \cdot \frac{2}{\sqrt{\pi}}\, e^{-\frac{\rho g^2}{2(1-\rho)}} \cdot \frac{\sqrt{\rho}}{\sqrt{2(1-\rho)}} = \frac{2\rho}{\sqrt{2\pi}\,\sqrt{1-\rho}}\, e^{-\frac{\rho g^2}{2(1-\rho)}}.$$

At $g = 0$ this is $r''_\rho(0) = \sqrt{2/\pi}\,\rho/\sqrt{1-\rho} > 0$. Substituting $r'_\rho(0) = 0$ and $r''_\rho(0) > 0$ into the expansion gives (E.4), and since $r''_\rho(g) > 0$ everywhere, $g = 0$ is the unique global minimum. ■

F.6.3 The inner variance at the boundary

Lemma E.3.

$$v_\rho(0) = (1-\rho)\Big(1 - \tfrac{2}{\pi}\Big). \tag{E.6}$$

Proof. At $g = 0$, $W = \sqrt{1-\rho}\, Z$, so $|W| = \sqrt{1-\rho}\,|Z|$ and

$$v_\rho(0) = (1-\rho)\operatorname{Var}(|Z|) = (1-\rho)\big(\mathbb{E}Z^2 - (\mathbb{E}|Z|)^2\big) = (1-\rho)\Big(1 - \tfrac{2}{\pi}\Big),$$

using $\mathbb{E}Z^2 = 1$ and $\mathbb{E}|Z| = \sqrt{2/\pi}$. ■

F.6.4 Conditional central limit theorem

Lemma E.4. For fixed g, as $M \to \infty$,

$$\mathrm{MAE}_M \mid G = g \;\approx\; \mathcal{N}\Big(c\sigma\, r_\rho(g),\; \frac{c^2\sigma^2}{M}\, v_\rho(g)\Big), \tag{E.7}$$

with an $O(1/\sqrt{M})$ skewness correction.

Proof. By Lemma E.1, $\mathrm{MAE}_M \mid g$ is the sample average of M independent and identically distributed variables $|E_m|$ with finite moments, since $|E_m| = c\sigma|\sqrt{\rho}g + \sqrt{1-\rho}Z_m|$ has finite moments of every order. The Lindeberg–Lévy central limit theorem gives

$$\sqrt{M}\Big(\frac{1}{M}\sum_{m=1}^{M} |E_m| - c\sigma r_\rho(g)\Big) \;\xrightarrow{d}\; \mathcal{N}\big(0,\, c^2\sigma^2 v_\rho(g)\big),$$

which is (E.7). The next Edgeworth term comes from the third moment of $|E_m|$ and is of order $O(1/\sqrt{M})$, affecting only the $O(1/M)$ remainder of the main result. ■

F.6.5 Limiting law and its lower endpoint

Proposition E.5. As $M \to \infty$, the inner variance in Lemma E.1 is proportional to $1/M$ and vanishes, so MAE_M converges to the limit random variable

$$Y := c\sigma\, r_\rho(G), \qquad G \sim \mathcal{N}(0,1), \tag{E.8}$$

whose support has infimum

$$\inf \operatorname{supp}(Y) = \sqrt{\tfrac{2}{\pi}}\, c\,\sigma\,\sqrt{1-\rho} =: y_0, \tag{E.9}$$

and $Y \geq y_0$ everywhere with equality only at $G = 0$.

Proof. First, the conditional mean in Lemma E.1 equals Y, so

$$\mathbb{E}[(\mathrm{MAE}_M - Y)^2] = \frac{c^2\sigma^2}{M}\,\mathbb{E}[v_\rho(G)] \leq \frac{c^2\sigma^2}{M} \longrightarrow 0,$$

where $v_\rho(g) \leq \mathbb{E}[(\sqrt{\rho}\,g + \sqrt{1-\rho}\,Z)^2] = \rho g^2 + 1 - \rho$. Thus $\mathrm{MAE}_M \to Y$ in mean square and hence in distribution. By Lemma E.2, $r''_\rho > 0$ everywhere and $r'_\rho(0) = 0$, so $g = 0$ is the unique global minimum and $r_\rho(g) \geq r_\rho(0) = \sqrt{2/\pi}\sqrt{1-\rho}$ for all g, with equality only at $g = 0$. Multiplying by $c\sigma > 0$ gives $Y \geq y_0$. Since $G \sim \mathcal{N}(0,1)$ places positive probability on every neighborhood of 0 and r_ρ is continuous, Y comes arbitrarily close to y_0, so y_0 is the infimum of its support. ■

The value y_0 is the lower endpoint of the limiting distribution in Theorem 2, and it relates to the mean (D.3) by $y_0 = \sqrt{2/\pi}\,c\sigma\sqrt{1-\rho} = \mathbb{E}[\mathrm{MAE}] \cdot \sqrt{1-\rho}$, smaller than the mean by a factor $\sqrt{1-\rho}$.

F.6.6 The square-root singularity of the limiting density

Lemma E.6. The density of $Y = c\sigma\, r_\rho(G)$ diverges to the right of y_0,

$$f_Y(y) \;\sim\; A_0\,(y - y_0)^{-1/2} \quad (y \downarrow y_0), \qquad A_0 = \frac{1}{\sqrt{2\pi\kappa}}, \quad \kappa = \tfrac{1}{2}c\sigma\, r''_\rho(0) > 0. \tag{E.10}$$

Proof. The infimum y_0 is given by Proposition E.5, so it remains to find the shape just above it. Let $\Psi(g) = c\sigma\, r_\rho(g)$, so $Y = \Psi(G)$. By the local expansion (E.4), near $g = 0$,

$$y = \Psi(g) = y_0 + \kappa g^2 + O(g^4), \qquad \kappa = \tfrac{1}{2}c\sigma\, r''_\rho(0) > 0,$$

an upward parabola with bottom at $g = 0$. Every y slightly above y_0 has two symmetric preimages $g_\pm(y) = \pm\sqrt{(y - y_0)/\kappa}\,(1 + o(1))$. The change-of-variables formula sums over preimages,

$$f_Y(y) = \sum_{i:\,\Psi(g_i) = y} \frac{\varphi(g_i)}{|\Psi'(g_i)|}.$$

Here $\Psi'(g) = c\sigma r'_\rho(g)$, and by (E.5) near zero $r'_\rho(g) = r''_\rho(0)g + o(g)$, so $\Psi'(g) = 2\kappa g + o(g)$ and $|\Psi'(g_\pm)| \sim 2\sqrt{\kappa(y - y_0)}$ at both roots. The numerators tend to $\varphi(0) = 1/\sqrt{2\pi}$, so summing the two equal contributions,

$$f_Y(y) \sim 2 \cdot \frac{1/\sqrt{2\pi}}{2\sqrt{\kappa(y - y_0)}} = \frac{1}{\sqrt{2\pi\kappa}}\,(y - y_0)^{-1/2},$$

which is (E.10). ■

The limiting density is zero below y_0 and diverges as $(y - y_0)^{-1/2}$ when approaching y_0 from above. The finite-M analysis below shows how conditional fluctuations of width $O(1/\sqrt{M})$ smooth this singularity and displace the mode. The endpoint y_0 belongs to the limiting law, not to the finite-M distribution, which can assign positive probability below y_0.

F.6.7 Finite M as a convolution and the universal peak

Lemma E.7. Let the smoothing width be

$$\delta_M = c\sigma\sqrt{\frac{v_\rho(0)}{M}} = c\sigma\sqrt{\frac{(1-\rho)(1 - 2/\pi)}{M}}. \tag{E.11}$$

In an $O(\delta_M)$ neighborhood of y_0,

$$f_{\mathrm{MAE}_M}(y) \;\approx\; \big(f_Y * \mathcal{N}(0, \delta_M^2)\big)(y) \approx A_0\,\delta_M^{-1/2}\,h\Big(\frac{y - y_0}{\delta_M}\Big), \tag{E.12}$$

where

$$h(u) = \frac{1}{\sqrt{2\pi}} \int_0^\infty v^{-1/2}\, e^{-(u-v)^2/2}\, dv = \int_0^\infty v^{-1/2} \varphi(u-v)\, dv, \tag{E.13}$$

so the leading-order location of the mode of f_{MAE_M} is

$$\hat{y}(M) = y_0 + u^*\, \delta_M + O(M^{-1}), \qquad u^* = \arg\max_{u>0} h(u). \tag{E.14}$$

Proof. For the convolution approximation, integrate the approximate conditional density of Lemma E.4 over the outer G,

$$f_{\mathrm{MAE}_M}(y) \approx \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\,\varsigma(g)} \exp\Big(- \frac{(y - c\sigma r_\rho(g))^2}{2\varsigma(g)^2} \Big) \varphi(g)\, dg, \qquad \varsigma(g) = c\sigma \sqrt{v_\rho(g)/M}.$$

Near y_0 the dominant contribution is from $g \approx 0$, since only $c\sigma r_\rho(g) \approx y$ keeps the inner Gaussian from being suppressed and $y \approx y_0$ corresponds to $g \approx 0$. In this neighborhood the slowly varying $\varsigma(g)$ is frozen at $\varsigma(0) = c\sigma\sqrt{v_\rho(0)/M} = \delta_M$, and the map $g \mapsto c\sigma r_\rho(g)$ becomes the image of the limiting law. This is exactly the statement that G is first mapped to $Y = c\sigma r_\rho(G)$ and then blurred by an independent $\mathcal{N}(0, \delta_M^2)$, so $f_{\mathrm{MAE}_M} \approx f_Y * \mathcal{N}(0, \delta_M^2)$.

For the scaling, insert the near-boundary form $f_Y(t) \approx A_0 (t - y_0)^{-1/2} \mathbf{1}_{t>y_0}$ of Lemma E.6,

$$f_{\mathrm{MAE}_M}(y) \approx \int_{y_0}^\infty A_0 (t - y_0)^{-1/2}\, \frac{1}{\sqrt{2\pi}\,\delta_M}\, e^{-\frac{(y-t)^2}{2\delta_M^2}}\, dt.$$

Substituting $v = (t - y_0)/\delta_M$ and $u = (y - y_0)/\delta_M$, so that $t - y_0 = \delta_M v$, $y - t = \delta_M(u - v)$ and $dt = \delta_M dv$,

$$f_{\mathrm{MAE}_M}(y) \approx A_0\, \delta_M^{-1/2} \cdot \frac{1}{\sqrt{2\pi}} \int_0^\infty v^{-1/2} e^{-(u-v)^2/2}\, dv = A_0\, \delta_M^{-1/2}\, h(u),$$

which is (E.12), and the weight $v^{-1/2}$ in (E.13) is precisely the square-root singularity of Lemma E.6.

For the leading-order peak, the dependence on y enters only through $u = (y - y_0)/\delta_M$, while $A_0 \delta_M^{-1/2}$ is free of y. Maximizing this profile therefore amounts to maximizing h. To establish uniqueness, differentiation under the integral and integration by parts give

$$h'(u) = \int_0^\infty v^{-1/2}(v - u)\varphi(u - v)\, dv, \qquad h''(u) + u h'(u) + \tfrac{1}{2} h(u) = 0.$$

The first identity gives $h'(u) > 0$ for $u \le 0$. Since $h(u) \to 0$ at both infinities, it has an interior maximum at some $u^* > 0$. At every stationary point, the second identity gives $h''(u) = -h(u)/2 < 0$, so every stationary point is a strict maximum and there can be only one. Thus the leading profile peaks at $y_0 + u^* \delta_M$. The remainder estimate in (E.14) follows from the local expansion detailed in Theorem E.8. ■

F.6.8 Assembling the main result

Theorem E.8. Suppose (D.5) holds with fixed $c\sigma > 0$ and $0 < \rho < 1$. Then for large M the dominant mode $\hat{y}(M)$ of the density of MAE_M satisfies

$$\hat{y}(M) = y_0 + u^*\, \delta_M + O\Big(\frac{1}{M}\Big), \qquad u^* = \arg\max_u h(u), \tag{T}$$

and equivalently the relative gap is the universal quantity

$$\frac{\hat{y}(M)}{y_0} - 1 = \frac{u^* \sqrt{\pi/2 - 1}}{\sqrt{M}} + O\Big(\tfrac{1}{M}\Big) = \frac{C_0}{\sqrt{M}} + O\Big(\tfrac{1}{M}\Big), \qquad C_0 = \sqrt{\tfrac{\pi}{2} - 1}\, u^*. \tag{T′}$$

Proof. Expanding δ_M in the mode formula (E.14) with $v_\rho(0) = (1-\rho)(1 - 2/\pi)$ from Lemma E.3,

$$\hat{y}(M) = y_0 + u^*\, c\sigma \sqrt{\frac{(1-\rho)(1-2/\pi)}{M}} + O\Big(\frac{1}{M}\Big),$$

which gives the leading term in (T). For the remainder, set $y = y_0 + \delta_M u$ and $g = \sqrt{\delta_M/\kappa}\, w$. For bounded u, w, the $O(g^4)$ term in $c\sigma r_\rho(g)$ contributes $O(M^{-1/2})$ to the standardized conditional mean. Evenness gives $v_\rho(g) = v_\rho(0) + O(g^2)$ and $\varphi(g) = \varphi(0) + O(g^2)$, also producing relative $O(M^{-1/2})$ corrections. The conditional Edgeworth correction has the same order. The folded-normal summands have Gaussian tails and smooth parameter dependence near $g = 0$; their local density expansion and its first two derivatives can therefore be integrated over the boundary region. This gives, uniformly for u in a fixed neighborhood of u^*,

$$\frac{\delta_M^{1/2}}{A_0} f_{\mathrm{MAE}_M}(y_0 + \delta_M u) = h(u) + O(M^{-1/2}),$$

with the same error order after one or two derivatives with respect to u. Since $h''(u^*) = -h(u^*)/2 < 0$, the stationary point is displaced by $O(M^{-1/2})$ in u, hence by $O(\delta_M M^{-1/2}) = O(M^{-1})$ in y. The boundary peak has height of order $\delta_M^{-1/2}$, whereas the limiting density is bounded away from y_0; the decaying tails of h localize the dominant peak to this boundary region. These observations justify the mode expansion without asserting that a central limit theorem alone controls density maxima.

For the relative gap, divide by $y_0 = \sqrt{2/\pi}\, c\sigma\sqrt{1-\rho}$,

$$\frac{\hat{y}(M) - y_0}{y_0} = \frac{u^*\, c\sigma\sqrt{(1-\rho)(1-2/\pi)/M}}{\sqrt{2/\pi}\, c\sigma\sqrt{1-\rho}} = \frac{u^*}{\sqrt{M}}\sqrt{\frac{1-2/\pi}{2/\pi}},$$

and since $\frac{1-2/\pi}{2/\pi} = \frac{\pi(1-2/\pi)}{2} = \frac{\pi-2}{2} = \frac{\pi}{2} - 1$, this is (T′). The factors c, σ, ρ all cancel, so the relative gap is universal. ■

Collecting the definitions, the universal constant of Theorem 2 is

$$C_0 = \sqrt{\tfrac{\pi}{2} - 1}\, u^*, \qquad u^* = \arg\max_{u>0} \int_0^\infty v^{-1/2} \varphi(u - v)\, dv. \tag{E.15}$$

The leading-order relative gap $0.578/\sqrt{M}$ is universal, giving approximately 5.8% at $M = 100$, 1.8% at $M = 1000$, and 0.9% at $M = 4000$. These estimates describe how the mode approaches the limiting lower endpoint y_0. Figure 15 provides validation on real benchmarks. The empirical MAE distributions closely match the theoretical predictions in both shape and mode even with only five models, and this agreement persists as M increases. Together with Figure 1, these results show that the behavior predicted by Theorem 2 is already visible with a small set of models.

A non-vanishing mean–mode gap is worth noting. Within the Gaussian model, (D.3) gives $\mathbb{E}[\mathrm{MAE}_M] = \sqrt{2/\pi}\, c\sigma$, independent of M and ρ, whereas the mode tends to $y_0 = \sqrt{2/\pi}\, c\sigma\sqrt{1-\rho}$. Thus $(\mathbb{E}[\mathrm{MAE}_M] - \hat{y}(M))/\mathbb{E}[\mathrm{MAE}_M] \to 1 - \sqrt{1-\rho} > 0$. This separation is consistent with the low-error mode and persistent high-error draws discussed in the main text. A mode near the lower endpoint does not, by itself, specify the probability of meeting a given error tolerance; that probability must be obtained from the distribution.

F.7 Proof of Theorem 3

Theorem 3. Under Assumptions 1 and 2, least-squares orthogonality decomposes the common shift into $\bar{E} = \bar{E}_\parallel + \bar{E}_\perp$ with $\bar{E}_\parallel = \beta^\top(\bar{x}_S - \bar{x}_N)$ and $\bar{E}_\perp = \bar{e}_S$, uncorrelated with variances $R^2 s^2$ and $(1-R^2)s^2$. For the full-benchmark-reference rule $S^* = \arg\min_{j \le m} \|\bar{x}_{S_j} - \bar{x}_N\|$, write $D^* = \|\bar{x}_{S^*} - \bar{x}_N\|$. Within the joint Gaussian approximation specified below, the selected scale satisfies $s_{S^*}^2 = \mathrm{Var}(\bar{E}_\parallel(S^*)) + (1-R^2)s^2$, with $0 \le \mathrm{Var}(\bar{E}_\parallel(S^*)) \le \|\beta\|^2 \mathbb{E}[D^{*2}]$. Consequently, if $\mathbb{E}[D^{*2}] \to 0$, the scale tends to $\sqrt{1-R^2}\, s < s$.

Throughout write $c := c(k)$. The orthogonal decomposition and the pre-selection variance identities are exact under uniform sampling without replacement and do not require Assumption 1. The post-selection residual-variance identity additionally uses a joint Gaussian approximation for the probe and residual imbalances, together with independent candidate draws from the full benchmark. Least-squares orthogonality alone gives zero covariance, not independence. This distinction specifies the scope of the selection formula in Theorem 3; the pool-based implementation is discussed below.

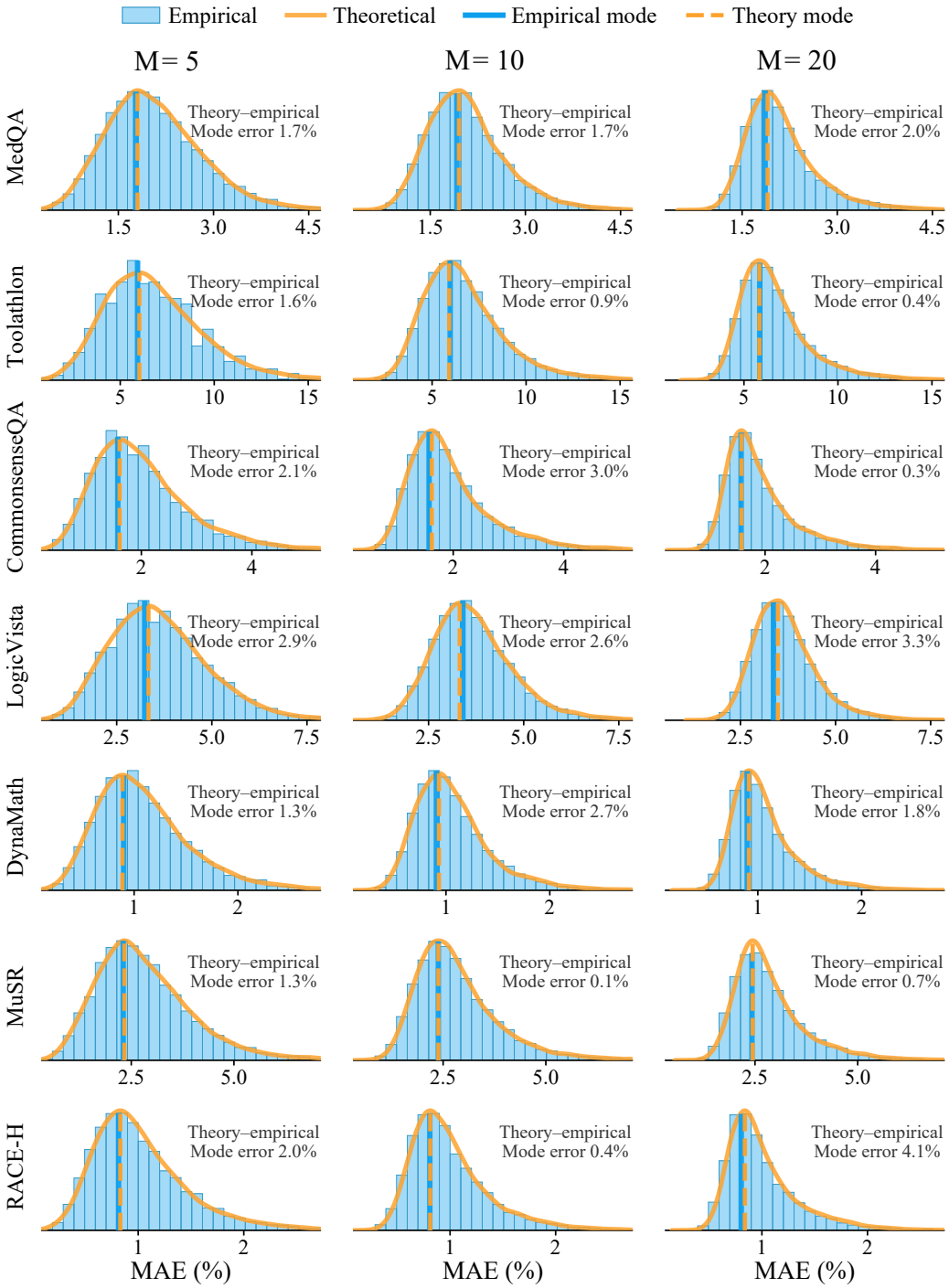

Figure 15: Empirical MAE distributions closely match the theoretical predictions of Theorem 2 across representative benchmarks. Their shapes and empirical modes agree with theory even when only a small number of models is available. Blue histograms show empirical distributions, orange curves show theoretical distributions, and the solid blue and dashed orange lines indicate the empirical and theoretical modes. The three panels in each row use the same horizontal scale.

F.7.1 Normal equations and in-sample orthogonality

Assumption 2 regresses difficulty on the p probe scores by least squares, $g_s = \alpha + \beta^\top x_s + e_s$, where α, β minimize $\sum_s (g_s - \alpha - \beta^\top x_s)^2$ and e_s is the residual.

Lemma F.1. The least-squares solution satisfies the normal equations

$$\bar{e}_N = \frac{1}{N}\sum_{s=1}^{N} e_s = 0, \qquad \mathrm{Cov}_N(x, e) = \frac{1}{N}\sum_{s=1}^{N} x_s e_s - \bar{x}_N \bar{e}_N = 0. \tag{F.2}$$

Proof. At the optimum the partial derivatives of the residual sum of squares vanish. Differentiating with respect to α,

$$\frac{\partial}{\partial \alpha}\sum_s (g_s - \alpha - \beta^\top x_s)^2 = -2\sum_s e_s = 0,$$

so $\sum_s e_s = 0$ and $\bar{e}_N = 0$. Differentiating with respect to the i-th component of β,

$$\frac{\partial}{\partial \beta_i}\sum_s (g_s - \alpha - \beta^\top x_s)^2 = -2\sum_s x_{s,i} e_s = 0,$$

so $\sum_s x_{s,i} e_s = 0$ for every i, that is $\sum_s x_s e_s = 0$. With $\bar{e}_N = 0$ this gives $\mathrm{Cov}_N(x, e) = \frac{1}{N}\sum_s x_s e_s - \bar{x}_N \bar{e}_N = 0$. ■

The orthogonality (F.2) is an algebraic identity of least squares and holds for any data.

F.7.2 The Pythagorean split of difficulty variance and R^2

Lemma F.2. The empirical difficulty variance splits orthogonally,

$$\mathrm{Var}_N(g) = \mathrm{Var}_N(\beta^\top x) + \mathrm{Var}_N(e), \tag{F.3}$$

so $R^2 := \mathrm{Var}_N(\beta^\top x)/\mathrm{Var}_N(g) \in [0, 1]$ and $\mathrm{Var}_N(e) = (1 - R^2)\mathrm{Var}_N(g)$. Under Assumption 2, $R^2 \in (0, 1]$.

Proof. From $g_s = \alpha + \beta^\top x_s + e_s$, and since a constant contributes no variance,

$$\mathrm{Var}_N(g) = \mathrm{Var}_N(\beta^\top x) + 2\,\mathrm{Cov}_N(\beta^\top x, e) + \mathrm{Var}_N(e).$$

The cross term is $\mathrm{Cov}_N(\beta^\top x, e) = \beta^\top \mathrm{Cov}_N(x, e) = 0$ by Lemma F.1, so only two terms remain, giving (F.3). Dividing by $\mathrm{Var}_N(g) > 0$ gives $R^2 + \mathrm{Var}_N(e)/\mathrm{Var}_N(g) = 1$, hence $\mathrm{Var}_N(e) = (1 - R^2)\mathrm{Var}_N(g)$ and $R^2 \in [0, 1]$. The case $R^2 = 1$ is allowed: it occurs when the probes fit difficulty exactly, leaving zero residual variance. Assumption 2 excludes only $R^2 = 0$. ■

Here R^2 measures how strongly items that are hard overall are also hard for the probes, and a larger value means the probes see more of the difficulty shift.

F.7.3 The visible and invisible parts of the common shift

Proposition F.3. The common shift decomposes as

$$\bar{E} = \bar{E}_\parallel + \bar{E}_\perp, \qquad \bar{E}_\parallel = \beta^\top(\bar{x}_\mathcal{S} - \bar{x}_N), \quad \bar{E}_\perp = \bar{e}_\mathcal{S}, \tag{F.4}$$

with uncorrelated parts of variance

$$\mathrm{Var}(\bar{E}_\parallel) = R^2 s^2, \qquad \mathrm{Var}(\bar{E}_\perp) = (1 - R^2)s^2, \qquad \mathrm{Cov}(\bar{E}_\parallel, \bar{E}_\perp) = 0. \tag{F.5}$$

Proof. By Proposition C.3, $\bar{E} = \bar{g}_\mathcal{S} - \bar{g}_N$. Substituting $g_s = \alpha + \beta^\top x_s + e_s$ and cancelling the constant α between the two averages,

$$\bar{E} = \big(\alpha + \beta^\top \bar{x}_\mathcal{S} + \bar{e}_\mathcal{S}\big) - \big(\alpha + \beta^\top \bar{x}_N + \bar{e}_N\big) = \beta^\top(\bar{x}_\mathcal{S} - \bar{x}_N) + (\bar{e}_\mathcal{S} - \bar{e}_N),$$

and $\bar{e}_N = 0$ from Lemma F.1 gives $\bar{e}_{\mathcal{S}} - \bar{e}_N = \bar{e}_{\mathcal{S}}$, which is (F.4). For the variances, $\bar{E}_{\parallel} = \beta^\top(\bar{x}_{\mathcal{S}} - \bar{x}_N)$ is the sample-mean deviation of the per-item quantity $\beta^\top x_s$, so Lemma B.2 with $h = h' = \beta^\top x$ gives

$$\mathrm{Var}(\bar{E}_{\parallel}) = c^2\,\mathrm{Var}_N(\beta^\top x) = c^2\,R^2\,\mathrm{Var}_N(g) = R^2 s^2,$$

using Lemma F.2 and $s^2 = c^2\mathrm{Var}_N(g)$ from (C.8). Likewise $\bar{E}_{\perp} = \bar{e}_{\mathcal{S}} = \Delta\bar{e}$ since $\bar{e}_N = 0$, so

$$\mathrm{Var}(\bar{E}_{\perp}) = c^2\,\mathrm{Var}_N(e) = c^2(1-R^2)\mathrm{Var}_N(g) = (1-R^2)s^2.$$

The cross term uses Lemma B.2 with $h = \beta^\top x$, $h' = e$ and the normal equation $\mathrm{Cov}_N(x, e) = 0$,

$$\mathrm{Cov}(\bar{E}_{\parallel}, \bar{E}_{\perp}) = c^2\,\mathrm{Cov}_N(\beta^\top x, e) = c^2\,\beta^\top \mathrm{Cov}_N(x, e) = 0.$$

■

Before selection, the common-shift variance splits into the probe-explained block $R^2 s^2$ and the residual block $(1-R^2)s^2$. Probe matching directly controls the former. Whether selection leaves the latter unchanged requires the additional distributional argument below.

F.7.4 Selecting the probe-matched subset contracts the visible block

Building on the decomposition above, we analyze an idealized full-reference selection rule for benchmark compression and show that probe matching can reduce the probe-explained component of common-shift variance. Covariate balance has been studied in other settings (Morgan & Rubin, 2012; Li et al., 2018; Wang & Li, 2025; Deville & Tillé, 2004). This rule draws $S_1, \ldots, S_m$ independently, each uniformly among the k-subsets of the full benchmark, and selects $S^* = \arg\min_{j \le m} \|\bar{x}_{S_j} - \bar{x}_N\|$ with $D^* = \|\bar{x}_{S^*} - \bar{x}_N\|$. Sampling is without replacement within each candidate, but candidates may overlap or repeat. Ties are broken by a fixed index rule that does not inspect residuals. Write $\bar{E}_{\parallel}(S^*) = \beta^\top(\bar{x}_{S^*} - \bar{x}_N)$ and $\bar{E}_{\perp}(S^*) = \bar{e}_{S^*}$.

Lemma F.4. Under Assumptions 1 and 2, and within the joint Gaussian approximation for independent candidate probe and residual imbalances, the common-shift scale on the selected subset is

$$s_{S^*}^2 = \mathrm{Var}\big(\bar{E}_{\parallel}(S^*)\big) + (1-R^2)s^2, \qquad 0 \le \mathrm{Var}\big(\bar{E}_{\parallel}(S^*)\big) \le \|\beta\|^2\,\mathbb{E}\big[D^{*2}\big], \tag{F.7}$$

If $\mathbb{E}[D^{*2}] \to 0$, then $s_{S^*} \to s\sqrt{1-R^2} < s$. The variance bound on the probe-explained part is exact even without the Gaussian approximation; invariance of the residual variance is a Gaussian-model conclusion.

Proof. For a candidate subset S, write $\Delta\bar{x} := \bar{x}_S - \bar{x}_N \in \mathbb{R}^p$ and $\Delta\bar{e} := \bar{e}_S - \bar{e}_N = \bar{e}_S$ using $\bar{e}_N = 0$, so the two parts are $\bar{E}_{\parallel} = \beta^\top\Delta\bar{x}$ and $\bar{E}_{\perp} = \Delta\bar{e}$. The proof is in three steps.

First, the visible block is squeezed by the probe imbalance, and the selection presses on it directly. By the Cauchy–Schwarz inequality, $|\bar{E}_{\parallel}| = |\beta^\top\Delta\bar{x}| \le \|\beta\|\,\|\Delta\bar{x}\| = \|\beta\|\,D$ with $D = \|\Delta\bar{x}\|$. The selection rule takes $\min_j \|\Delta\bar{x}_j\|$, driving D to D^*, so the selected visible block obeys $|\bar{E}_{\parallel}(S^*)| \le \|\beta\|D^*$. Squaring and taking expectation,

$$\mathrm{Var}\big(\bar{E}_{\parallel}(S^*)\big) \le \mathbb{E}\big[\bar{E}_{\parallel}(S^*)^2\big] \le \|\beta\|^2\,\mathbb{E}\big[D^{*2}\big] \xrightarrow{D^* \to 0} 0.$$

The bound holds for any m and rests only on Cauchy–Schwarz. Increasing the candidate count makes D^* and this upper bound non-increasing under a coupling that retains previous candidates. For a fixed finite benchmark, however, the attainable minimum is $\min_{|S|=k} \|\bar{x}_S - \bar{x}_N\|$, which need not equal zero. Thus the limit $D^* \to 0$ is conditional, not guaranteed merely by taking $m \to \infty$. The sufficient condition used here is $\mathbb{E}[D^{*2}] \to 0$; for bounded probe scores, convergence in probability to zero also implies this condition.

Second, the invisible block is untouched by the selection. The rule reads only the probe imbalances $\{\Delta\bar{x}_j\}$ and never the residual imbalances $\{\Delta\bar{e}_j\}$. By Lemma B.2 and the normal equation $\mathrm{Cov}_N(x, e) = 0$,

$$\mathrm{Cov}(\Delta\bar{x}, \Delta\bar{e}) = c^2\,\mathrm{Cov}_N(x, e) = \mathbf{0}.$$

Apply the finite-population central limit argument (Li & Ding, 2017) to the augmented per-item vector (x_s, e_s), with fixed probe dimension and the corresponding boundedness and covariance-limit conditions. This motivates a joint Gaussian approximation for $(\Delta\bar{x}, \Delta\bar{e})$. Within that model,

zero cross-covariance implies independence. Independence across candidate draws then makes the entire residual vector independent of the entire collection of probe imbalances. Conditional on all probe imbalances, the selected index is fixed and its residual retains the same zero-mean Gaussian marginal as every candidate. Therefore, within this approximation,

$$\mathrm{Var}\big(\bar{E}_\perp(S^*)\big) = \mathrm{Var}(\Delta\bar{e}) = c^2\,\mathrm{Var}_N(e) = (1-R^2)s^2.$$

Third, combine the two blocks. Since $\bar{E}_\parallel(S^*)$ is a function of $\{\Delta\bar{x}_j\}$ while $\bar{E}_\perp(S^*)$ is independent of them with zero mean,

$$\mathbb{E}\big[\bar{E}_\parallel(S^*)\,\bar{E}_\perp(S^*)\big] = \mathbb{E}\big[\bar{E}_\parallel(S^*)\big]\cdot 0 = 0,$$

so the two variances add,

$$s_{S^*}^2 = \mathrm{Var}\big(\bar{E}_\parallel(S^*)\big) + (1-R^2)s^2.$$

The first term lies in $[0, \|\beta\|^2\mathbb{E}[D^{*2}]]$ and vanishes if $\mathbb{E}[D^{*2}] \to 0$, giving $s_{S^*} \to s\sqrt{1-R^2}$. For the original finite-population distribution, zero covariance alone does not ensure that selection preserves the residual variance or the cross-covariance, so the displayed post-selection equality is approximate rather than an exact sampling identity. Nor does a fixed-candidate Gaussian approximation by itself justify arbitrarily increasing m. ■

The bound on the probe-explained variance shrinks with $\mathbb{E}[D^{*2}]$, while the residual contribution stays unchanged within the Gaussian selection model. Here β is the coefficient vector for the panel-average difficulty g_s, not a coefficient separately defined for each target model. Since $|\bar{E}_\parallel| \le \|\beta\|D$, reducing D provides a selection criterion that does not require knowing β. This motivates BestRand without asserting that every finite candidate count achieves the limiting contraction factor.

In the centered Gaussian selection model, simultaneous sign reversal of all candidate probe imbalances leaves the selected index unchanged and reverses the selected probe vector. Its mean is therefore zero. Together with the independent zero-mean residual, this gives $\mathbb{E}[\bar{E}(S^*)] = 0$ within that model. This is not an exact unbiasedness guarantee for every target model under finite-population or pool-based selection.

Figures 16 and 17 compare the predictions of Lemma F.4 with experimental results. As m increases, $\sqrt{\mathbb{E}[D^{*2}]}$ decreases under full-reference selection, while finite-pool BestRand reduces both its pool-reference objective and its full-reference imbalance. The common-shift scale ratio s_{S^*}/s decreases accordingly, and the Gaussian and empirical full-reference results approach the predicted limit $\sqrt{1-R^2}$. Finite-pool results follow the same trend, with small deviations because H only approximates $\bar{x}_N$. Overall, the experimental results closely match the predictions of Theorem 3.

F.7.5 Implications for high-error risk

For every selection rule, the pointwise inequality from Proposition C.3 remains valid:

$$\Pr\big(\mathrm{MAE}(S^*) > t\big) \ge \Pr\big(|\bar{E}(S^*)| > t\big). \tag{14}$$

Within the independent-candidate Gaussian model of Lemma F.4, if $\mathbb{E}[D^{*2}] \to 0$ and $0 < R^2 < 1$, the probe-explained part vanishes in mean square while the residual remains $\mathcal{N}(0, (1-R^2)s^2)$. Thus, for fixed $t > 0$,

$$\Pr\big(|\bar{E}(S^*)| > t\big) \longrightarrow 2\Phi\Big(-\frac{t}{s\sqrt{1-R^2}}\Big).$$

For $R^2 = 1$, the limiting common shift is zero and this probability tends to zero. At finite probe imbalance, the selected common shift need not be Gaussian, so replacing its law by a normal distribution solely on the basis of its variance is not justified.

The contraction reduces the common-shift contribution to high-error risk. A smaller lower bound alone does not prove that the actual MAE tail probability decreases, nor that its variance or quantiles contract by the same factor, or that its mode stays fixed. The empirical reduction in high-error outcomes is reported in Figure 2; the theorem provides the common-shift mechanism motivating the search.

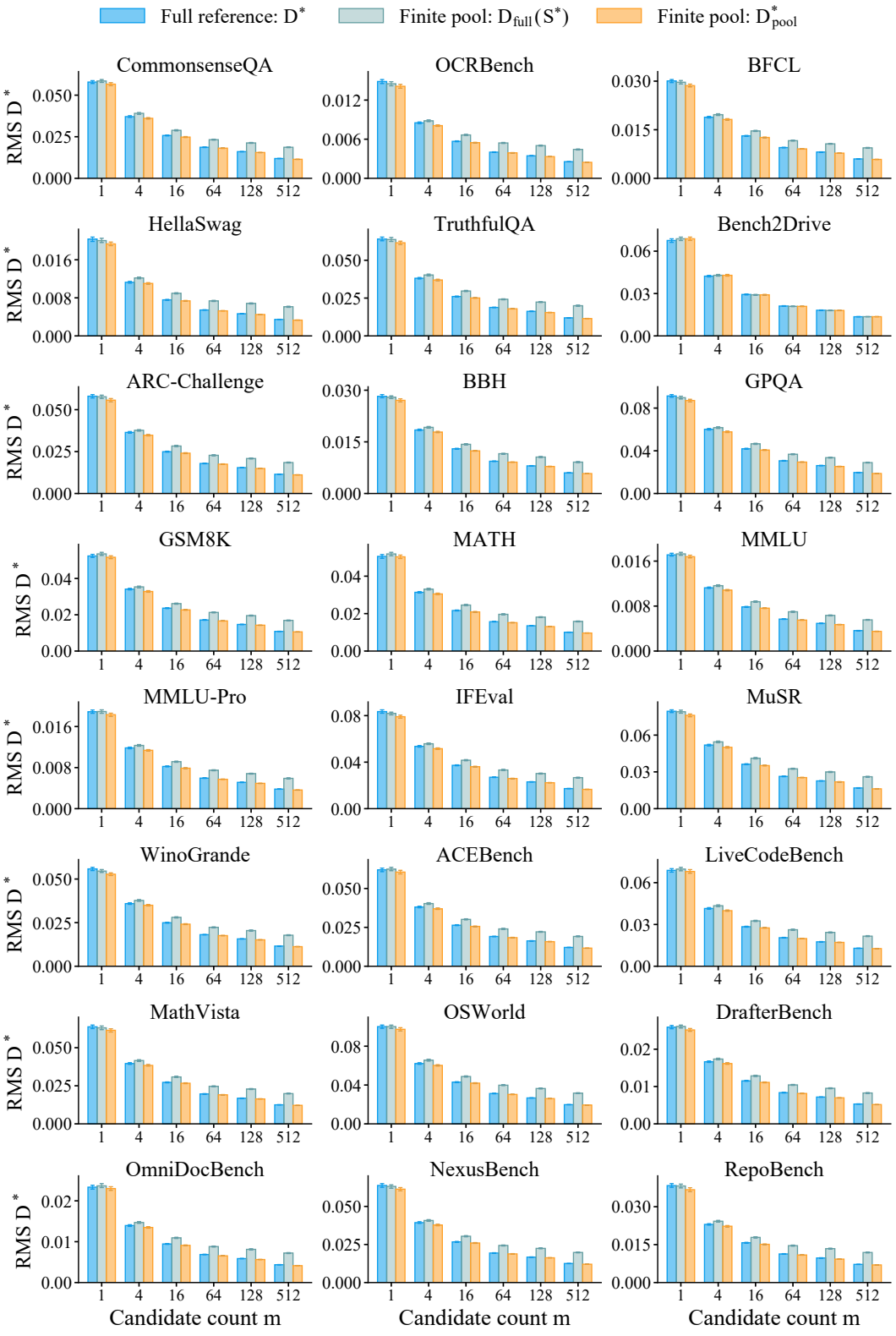

Figure 16: RMS probe imbalance as the candidate count increases. Full-reference selection reports $\sqrt{\mathbb{E}[D^{*2}]}$ for the rule in Theorem 3. For finite-pool BestRand, D^*_{pool} is the optimized imbalance and $D_{\text{full}}(S^*)$ is the imbalance relative to the full benchmark probe mean.

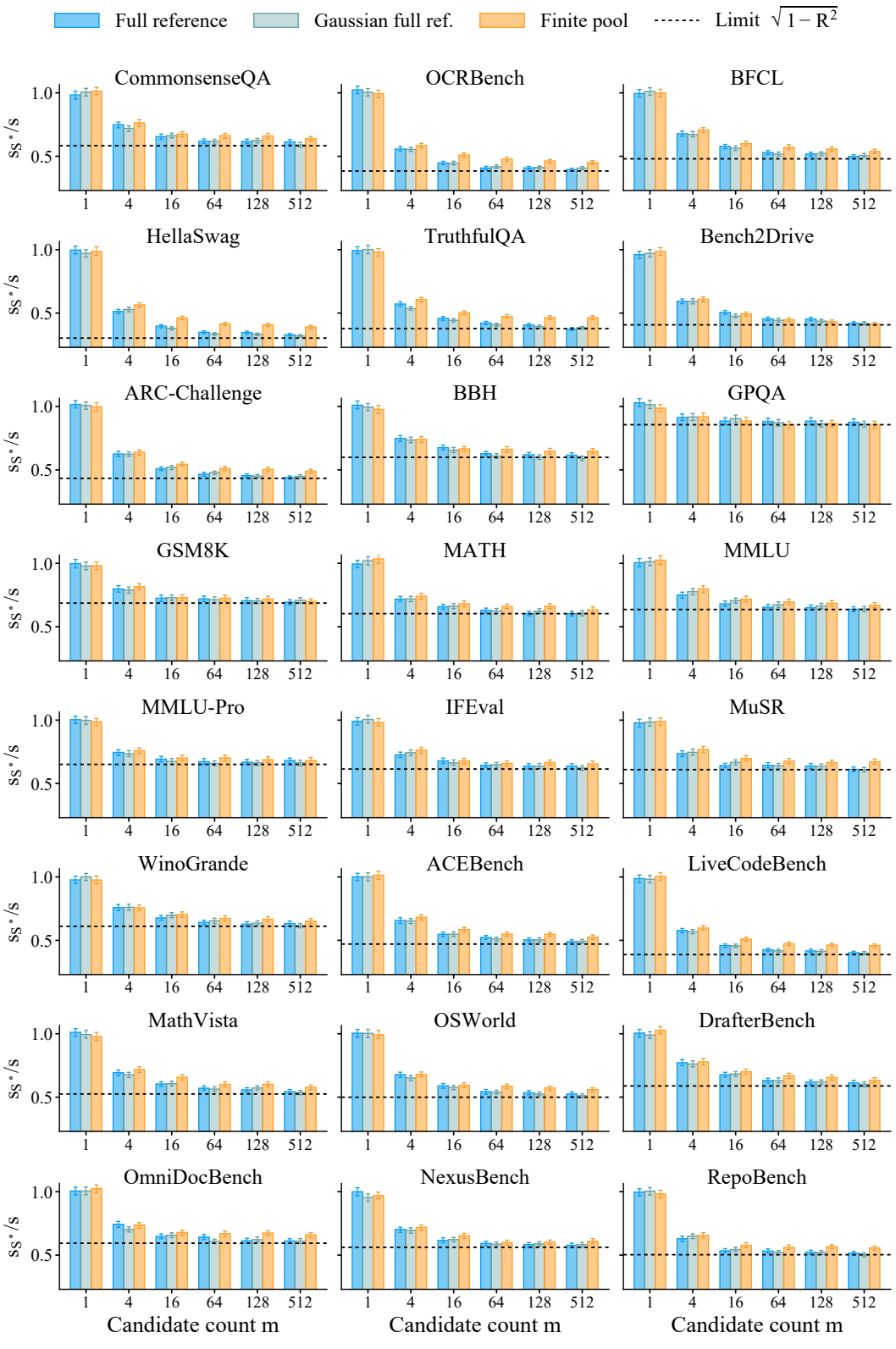

Figure 17: Common-shift scale ratio s_{S^*}/s as the candidate count increases. Curves compare the joint Gaussian full-reference model, empirical full-reference selection, and finite-pool BestRand. The horizontal line marks the Gaussian full-reference limit $\sqrt{1-R^2}$ as $\mathbb{E}[D^{*2}] \to 0$.

G Related Work

G.1 LLM, VLM, and VLA Benchmarks

Benchmarks for LLMs, VLMs, and VLAs evaluate a broad range of capabilities across text, multimodal, and interactive tasks (Hendrycks et al., 2020; Chen et al., 2021; Lu et al., 2024; Dauner et al., 2024; Ling et al., 2026; Zhong et al., 2024; Liu et al., 2025). Examples include LongBench v2, MMMU, and Bench2Drive (Bai et al., 2025; Yue et al., 2024; Jia et al., 2024). Evaluating modern foundation models can be expensive even for individual items, particularly when models process long contexts, generate extended reasoning chains, or perform repeated inference during interaction with an environment (Huang et al., 2025b; Bai et al., 2025; Jia et al., 2024). This cost accumulates across benchmarks containing hundreds to tens of thousands of evaluation items (Polo et al., 2024; Perlitz et al., 2024). Many benchmarks also contain redundancy, as different items can provide overlapping information about model capabilities (Zhang et al., 2025b; Vivek et al., 2024). Prior studies show that small subsets can closely approximate full-benchmark scores or preserve model rankings, suggesting that much of the information needed for aggregate evaluation can be retained with fewer items (Vivek et al., 2024; Polo et al., 2024; Huang et al., 2025b; Saranathan et al., 2025; Zhao et al., 2024; Wang et al., 2026). BestRand exploits this redundancy to reduce evaluation cost across these settings while improving the accuracy and stability of evaluation on compact subsets.

G.2 Low-Cost Deep Learning

Improving efficiency and reducing training and inference costs have long been central goals in deep learning, with methods operating on both models and data (Wan et al., 2023). On the model side, pruning and quantization improve inference efficiency (Han et al., 2015; Frantar & Alistarh, 2023; Xiao et al., 2023), while parameter-efficient fine-tuning reduces the resources required to adapt pretrained models (Houlsby et al., 2019; Hu et al., 2021). On the data side, coreset selection and data pruning retain informative training examples (Mirzasoleiman et al., 2020; Paul et al., 2021), while dataset condensation synthesizes compact training sets that approximate the learning signal of larger datasets (Zhao et al., 2020; Cazenavette et al., 2022). BestRand applies this data-efficiency perspective to model evaluation by selecting a reusable subset of existing test items that closely preserves full-benchmark results at low construction cost. By reducing the number of evaluated items, BestRand complements methods that reduce the cost of processing each item.

G.3 Benchmark Compression Methods

Benchmark compression reduces the number of evaluated items while preserving full-benchmark scores or model rankings. For constructing reusable subsets, a useful distinction is whether methods use historical item-level model responses. Methods without such records use random sampling or select representative items through prompt embeddings and other content-derived features, typically using clustering or diversity-based selection (Song et al., 2025a; Bean et al., 2025). Response-based methods instead exploit evaluation results from previously tested models to identify informative items through response-pattern clustering, item response modeling, or learned score prediction (Vivek et al., 2024; Polo et al., 2024; Zhang et al., 2026). Many such methods rely on large collections of model evaluations, making subset construction expensive for new benchmarks without public evaluation logs (Bean et al., 2025). Efforts to reduce this dependence include learning from small anchor sets, augmenting observed responses with synthetic records, and calibrating random subset sizes using a few models (Saranathan et al., 2025; Huang et al., 2026; Das et al., 2026).

Beyond method development, prior analyses examine benchmark redundancy (Zhang et al., 2025b; Sha & Zhao, 2026) and the statistical reliability of model comparisons (Miller, 2024; Heineman et al., 2025). Studies of compressed evaluation further highlight sensitivity to calibration data and model shifts, the limitations of very small subsets, and the competitiveness of random sampling (Zhang et al., 2025a; Yauney et al., 2026). Our theory complements these perspectives by characterizing the distribution of compression errors, explaining persistent high-error risk and providing a principled basis for reducing it. BestRand offers improved accuracy and stability over random sampling with lower construction cost than methods that rely on extensive historical evaluation records. Detailed descriptions of the baselines considered in our experiments are provided in Appendix A.3.

H THE NOVELTY AND CONTRIBUTION OF BESTRAND

H.1 NOVELTY OF BESTRAND

Our novelty follows a clear progression. We first derive a theoretical characterization of benchmark compression and explain its behavior, then use these insights to improve its reliability and stability.

An accurate characterization of the error distribution in benchmark compression. We derive an accurate analytical approximation to the MAE distribution when a benchmark subset is used to estimate full-benchmark scores. This distribution-level view goes beyond average error and ranking fidelity by describing both typical outcomes and high-error outcomes. It therefore provides a principled basis for understanding the accuracy and reliability of benchmark compression. Numerical results across LLM, VLM, and VLA benchmarks closely match the predicted distributions.

An explanation of why random sampling is almost enough. Our analysis shows that the MAE distribution has a dominant low-error region. As the number of evaluated models grows, its mode approaches the lower endpoint of the limiting distribution at a universal rate. Together with the observed concentration around this region, this result explains why random sampling produces accurate compressed benchmarks with high probability and is often a strong baseline. The locations of the empirical modes are consistent with the theoretical prediction.

An explanation of the persistent instability of random sampling. We identify the common shift, caused by a subset being easier or harder than the full benchmark, as a key source of the upper error tail. Its scale grows with behavioral correlation across models and remains positive as model count grows under positive correlation. Thus, evaluating more models cannot average away shared errors. This persistent instability reflects correlated behavior among foundation models, consistent with relationships among model correlation, common-shift scale, and high-error risk.

A theoretically grounded, stable, and low-cost method. This analysis motivates BestRand. We decompose the common shift into probe-explained and residual components. BestRand draws multiple random subsets and selects the one whose mean probe scores best match a pool reference to control the explained component. Under our selection model, theory guarantees that the common-shift variance converges to a value strictly below that of random sampling as the probe imbalance vanishes, thereby improving stability. Experiments consistently confirm this theoretical prediction.

H.2 CONTRIBUTIONS OF BESTRAND

Together, these theoretical insights and methodological advances yield three main contributions.

A theoretical framework for benchmark compression. We derive exact sampling moments and an accurate analytical approximation to the MAE distribution. The framework connects compression error and high-error risk to the subset size, item difficulty variation, and model correlation. The theory further shows that matching probe-model scores can reduce common-shift variance. Extensive numerical validation shows that the predicted distributions, modes, and tail-related quantities closely match empirical measurements, supporting both the validity and practical value of the theory.

A stable and low-cost compression method. We introduce BestRand, which uses a few probe models to select from random subsets. These probe models evaluate only a small subset of benchmark items rather than the full benchmark, so construction cost is low. Our theory shows that sufficiently accurate probe matching gives BestRand lower common-shift variance than Random and thus greater stability. Experiments consistently confirm this stability gain.

Extensive empirical validation and strong results. We evaluate BestRand on 70 text, multimodal, and agent benchmarks across various retention ratios, and further extend it to VLA evaluation. The empirical error distributions closely match our theoretical analysis, and BestRand consistently suppresses the high-error region predicted by the theory. It improves estimation accuracy at low construction cost while maintaining strong ranking fidelity. These gains remain consistent under score-range, model-family, and temporal shifts, supporting its robustness to unseen target models.