Submodular Benchmark Selection

Choosing Which Benchmarks to Run via Entropy and Mutual Information

Alex Smola

Boson AI

RL-Eval Workshop · 2026

Data from 10 public leaderboards

Outline

1. Motivation

Why most benchmarks are redundant.

2. Framework

From regression to Gaussian model; entropy and mutual information.

3. Algorithm

Submodularity, greedy guarantees, EM.

4. Experiments

Three matrices, ten leaderboards, cross-validated imputation R^2.

5. Takeaways

Question. A new model arrives. Which k benchmarks should we run on it to predict the rest of its scores?

Answer. A simple Gaussian model + submodular optimization gives:

- $k = 5$ benchmarks $\Rightarrow R^2 \approx 0.91$ on MMLU
- $(1 - 1/e)$ approximation guarantee
- Connection to pivoted Cholesky

The Benchmark Landscape Has Exploded

Collection	Models	Tasks
MMLU	5,452	57
Open LLM v2	4,507	6
MTEB	263	56
AlpacaEval 2	223	3
LiveBench	195	3
BigCodeBench	155	14
WildBench	63	11
Arena-Hard	60	1
HELM Lite	30	11
MT-Bench	5	8
Total tasks		**170+**

Two competing pressures

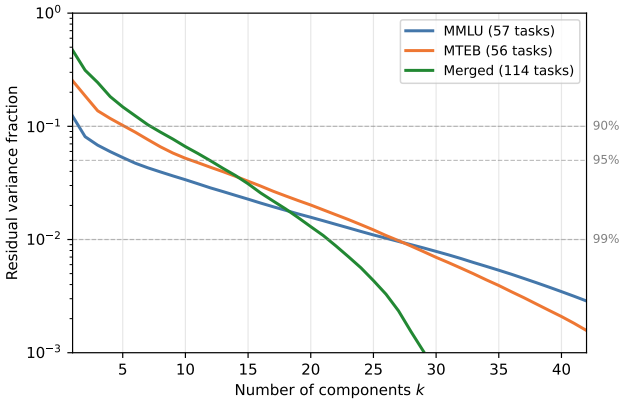
Completeness: cover math, code, knowledge, reasoning, safety, ...

Cost: every new model requires re-running *everything*. Days of GPU time.

Hypothesis: Many benchmarks measure the *same underlying capability*.

A small subset should suffice — if we know *which* subset.

Are More Benchmarks Really Better?



Eigenspectrum of the score covariance,
log-scale residual variance $1 - \rho(k)$.

Empirical decay

MMLU: 2 components $\rightarrow$ 90%
MTEB: 6 components $\rightarrow$ 90%
MERGED: 8 components $\rightarrow$ 90%

Conclusion. Benchmark scores live in a *low-dimensional* subspace.

A few well-chosen benchmarks should suffice.

The Imputation Problem

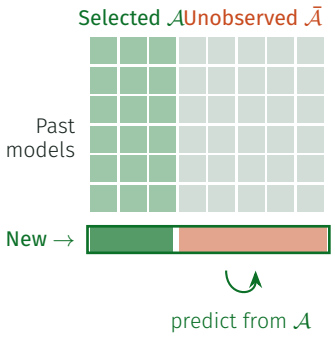
M models, N benchmarks, score matrix $B \in \mathbb{R}^{M \times N}$.

Pick a subset $\mathcal{A} \subseteq \{1, \ldots, N\}$. For a new model:

- Observe scores $B_{i,\mathcal{A}}$ on selected benchmarks
- *Impute* $B_{i,\bar{\mathcal{A}}}$ for the rest

Simplest predictor: linear regression of $B_{\bar{\mathcal{A}}}$ on $B_{\mathcal{A}}$, learned from past models.

Need: second-order statistics (μ, Σ) — estimable from leaderboard data.



Gaussian Model: Closed-Form Imputation

Assume each row of B is i.i.d. Gaussian:

$$B_{i\cdot} \sim \mathcal{N}(\mu, \Sigma)$$

Obviously wrong — but extremely useful.

Conditional distribution gives the imputer:

$$\boxed{\hat{B}_{i,\bar{\mathcal{A}}} = \mu_{\bar{\mathcal{A}}} + \Sigma_{\bar{\mathcal{A}}\mathcal{A}} \Sigma_{\mathcal{A}\mathcal{A}}^{-1} (B_{i,\mathcal{A}} - \mu_{\mathcal{A}})}$$

$$\Sigma_{\bar{\mathcal{A}}|\mathcal{A}} = \Sigma_{\bar{\mathcal{A}}\bar{\mathcal{A}}} - \Sigma_{\bar{\mathcal{A}}\mathcal{A}} \Sigma_{\mathcal{A}\mathcal{A}}^{-1} \Sigma_{\mathcal{A}\bar{\mathcal{A}}}$$

Two key properties

1. The conditional covariance $\Sigma_{\bar{\mathcal{A}}|\mathcal{A}}$ depends only on *which* benchmarks are in $\mathcal{A}$ — not on the observed values.

2. The conditional mean is the *best linear predictor*, regardless of whether B is actually Gaussian.

$\Rightarrow$ We can optimize the *choice of* $\mathcal{A}$ *before* seeing any new data.

Two Objectives

Question 1. How much information do we *gain* by adding benchmark v to $\mathcal{A}$?
$\rightarrow$ **Entropy:**

$$f_1(\mathcal{A}) = H(X_{\mathcal{A}}) = \tfrac{1}{2} \log\det\left(2\pi e\, \Sigma_{\mathcal{A}\mathcal{A}}\right)$$

Tests that are diverse from each other.

Question 2. How much do we learn about the *remaining* benchmarks?
$\rightarrow$ **Mutual Information:**

$$\begin{aligned} f_2(\mathcal{A}) &= I(X_{\mathcal{A}}; X_{\bar{\mathcal{A}}}) \\ &= \tfrac{1}{2} \left[\log\det \Sigma_{\mathcal{A}\mathcal{A}} + \log\det \Sigma_{\bar{\mathcal{A}}\bar{\mathcal{A}}} - \log\det \Sigma\right] \end{aligned}$$

Tests coupled with the complement.

Marginal MI gain of adding v:

$$\Delta(v \mid \mathcal{A}) = \tfrac{1}{2} \underbrace{\log \sigma^2_{v|\mathcal{A}}}_{\text{grows: diversity}} - \tfrac{1}{2} \underbrace{\log \sigma^2_{v|\bar{\mathcal{A}}_v}}_{\text{shrinks: coupling}}$$

Both are closed-form under the Gaussian model: log-determinants and Schur complements only.

Surrogate gap. Entropy minimizes *residual variance*; MI maximizes *predictive coupling*. For imputation we want MI.

Submodularity and Greedy Selection

Submodularity = diminishing returns:

$$\underbrace{f(\mathcal{A} \cup \{v\}) - f(\mathcal{A})}_{\text{early gain}} \geq \underbrace{f(\mathcal{B} \cup \{v\}) - f(\mathcal{B})}_{\text{later gain}}$$

for all $\mathcal{A} \subseteq \mathcal{B}$.

Greedy: at each step, add the v with largest marginal gain.

Nemhauser et al. 1978

For *monotone submodular* f:

$$f(\mathcal{A}_{\text{greedy}}) \geq (1 - 1/e) \cdot \max_{|\mathcal{A}| = k} f(\mathcal{A})$$

And Feige (1998): *no polynomial-time algorithm* can do strictly better.

For Gaussians (Krause, Guestrin, Singh 2005):

Entropy $H(X_{\mathcal{A}})$ is submodular. After a constant shift, it's also monotone $\Rightarrow$ $(1 - 1/e)$ **guarantee** applies.

The shift doesn't change greedy's choices.

Mutual information is submodular but *not monotone* in general — once $\mathcal{A}$ covers most of the signal, adding v shrinks the complement.

But for small $k \ll N$, MI gains stay positive in all our experiments. We use greedy as a *heuristic* here.

Algorithms: Entropy and Mutual Information

Greedy entropy = pivoted Cholesky.

At each step, select the benchmark with largest residual variance:

$$j^* = \underset{j \notin \mathcal{A}}{\arg\max} \text{ with } d_j, \qquad d_j = \sigma^2_{j|\mathcal{A}}$$

Maintain d_j via rank-one Cholesky:

$$\ell_{j,t} = \left(\Sigma_{j,j^*} - \boldsymbol{\ell}_j^\top \boldsymbol{\ell}_{j^*}\right) / \sqrt{d_{j^*}} \text{ and } d_j \leftarrow d_j - \ell^2_{j,t}$$

Cost: $O(k^2 N)$ total.
Residual variance *tracked for free*

$$\operatorname{tr}(\Sigma_{\bar{\mathcal{A}}|\mathcal{A}}) = \operatorname{tr}(\Sigma - L_k L_k^\top)$$

Greedy mutual information.

Pick benchmark whose log-determinant gain is largest on both sides:

$$j^* = \underset{j \in \bar{\mathcal{A}}}{\arg\max} \ \tfrac{1}{2}\left[\log d_j + \log P_{jj}\right]$$

where $P_{jj} = \left[(\Sigma_{\bar{\mathcal{A}},\bar{\mathcal{A}}})^{-1}\right]_{jj}$

Computed from a *fresh* Cholesky of the complement block:

$$L_{\bar{\mathcal{A}}} L_{\bar{\mathcal{A}}}^\top = \Sigma_{\bar{\mathcal{A}},\bar{\mathcal{A}}}, \qquad P_{jj} = \left\| \left(L_{\bar{\mathcal{A}}}^{-1}\right)_{:,j} \right\|^2$$

Cost: $O(kN^3)$. Precision downdates are unstable; fresh Cholesky is fast enough.

The Gaussian Model Buys Us EM "For Free"

Real leaderboards are *not* fully observed. $\Rightarrow$ Need $\hat{\Sigma}$ from *incomplete* data.

EM (Dempster, Laird, Rubin 1977) under MAR:

- **E-step**: impute missing entries with $\mathbb{E}[B_{i,\bar{\mathcal{B}}_i} \mid B_{i,\mathcal{B}_i}]$ — same formula as our imputer.
- **M-step**: re-estimate (μ, Σ) from completed data + a per-row uncertainty correction C_i that keeps Σ PSD.

The Gaussian model wears two hats:

1. Drives *selection* via H and I.
2. Drives *imputation* via the conditional mean.

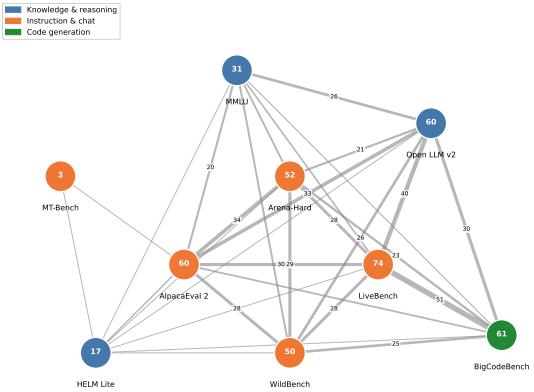
Once you commit to the model, everything else is closed-form.

Rank-deficient case ($M < N$): Ledoit–Wolf shrinkage with $\alpha = (N - M)/N$, PSD floor after each M-step.

Three Score Matrices, Ten Leaderboards

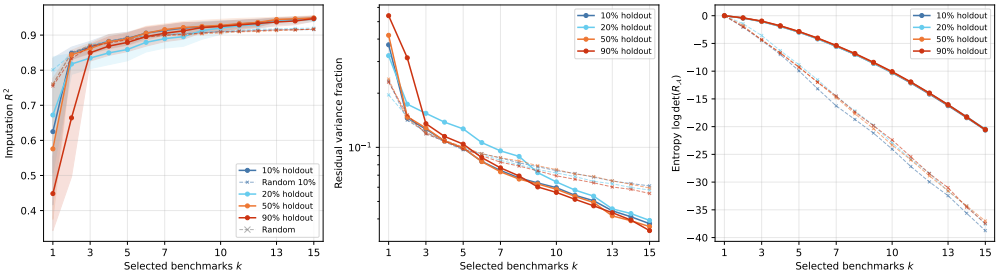
Matrix	*M*	*N*	% obs.
MMLU	5 452	57	100%
MTEB	263	56	77%
MERGED	118	114	31%
BENCHPRESS	83	49	34%

MERGED canonicalizes 118 model names across 9 leaderboards: heterogeneous, sparse. **BENCHPRESS** (Papailiopoulos et al. 2026): the sparsest matrix in our study.



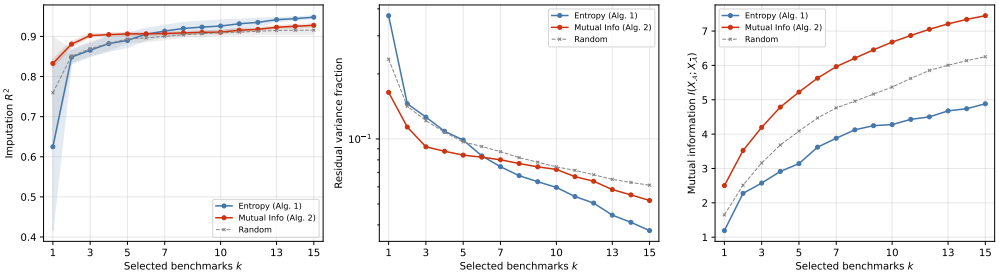
Pairwise overlap between leaderboards (MTEB omitted: embedding-only).

Entropy on MMLU: A Few Benchmarks Suffice



10-fold CV. $k = 5$: $R^2 = 0.89$. $k = 10$: $R^2 > 0.92$. $k = 15$: $R^2 = 0.95$. Stable across holdout fractions, even at 90% holdout (∼545 training models).

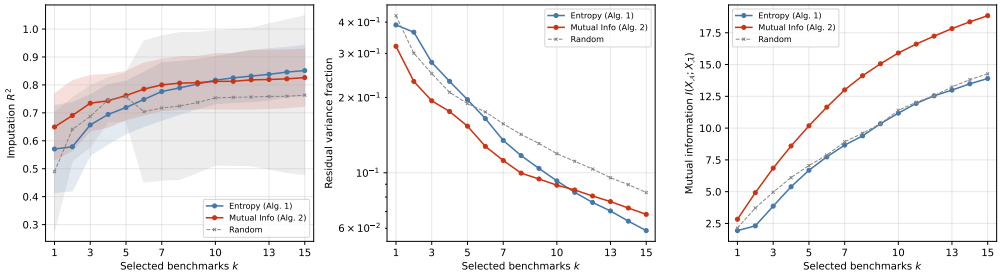
Entropy vs. Mutual Information: MMLU



MI wins at small k. $R^2 = 0.83$ (MI) vs. 0.65 (entropy) at $k = 1$; crossover near $k \approx 7$.

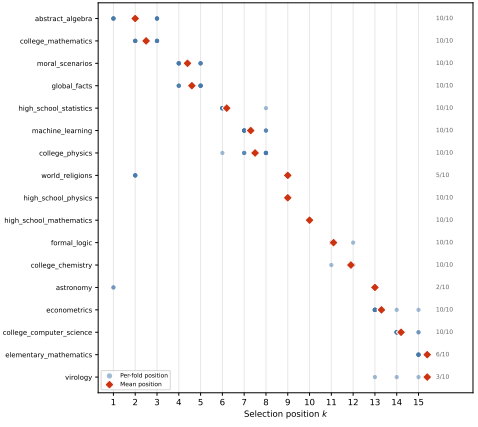
Surrogate gap: entropy gives lower residual variance (center) yet *worse* imputation — it maximizes diversity *within* $\mathcal{A}$, not coupling with $\bar{\mathcal{A}}$.

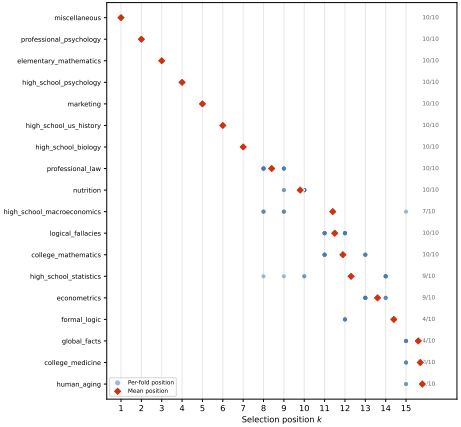
Entropy vs. Mutual Information: MTEB



56 embedding tasks: MI leads entropy by $\sim$10 R^2 points for $k \leq 3$; entropy catches up around $k \approx 10$ as diversity becomes sufficient. Same pattern on MERGED — MI advantage persists out to $k = 15$ when the covariance is sparser.

Which Benchmarks Get Selected?





Entropy: “outlier” subjects (abstract_algebra, moral_scenarios)

MI: “hub” subjects (miscellaneous, prof._psychology, elementary_math)

MI’s first 9 benchmarks are **identical across all 10 folds**.

Real-World Application: Discovering New Benchmarks

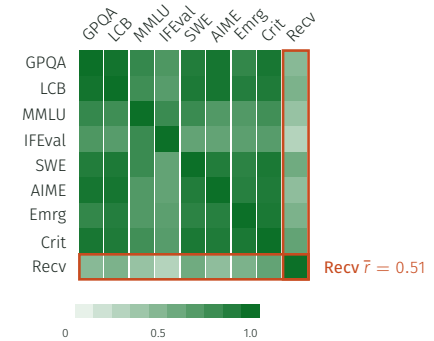
ProactBench
(Harfi et al. 2026, `arXiv:2605.09228`)
New benchmark measuring *conversational proactivity*.

Six standard benchmarks correlate at $r = 0.64$–0.97 (GPQA, LCB, MMLU, …).

Recovery breaks the pattern $\bar{r} = 0.51$, 95% CI $[0.29, 0.71]$.

By submodular selection [our paper], greedy entropy on the 9×9 correlation matrix **ranks Recovery #2**, ahead of every existing benchmark (top three in 97% of bootstrap resamples).

Pairwise Pearson r — sketch of ProactBench Fig. 3:



Recovery row+column outlined: low correlations with everything else.

Takeaways

You don't need all the benchmarks

5 MI-selected MMLU subjects $\rightarrow R^2 = 0.91$. Run the selected ones, impute the rest.

The dual works too

Transpose $B \rightarrow$ select reference *models* that characterize a new benchmark.

Discover new benchmarks

Rank benchmarks by their information *about rest*. Proactivity/Recovery ranks #2.

Caveat: gaming

The method works *until* a benchmark is gamed — and is the right tool to *detect* the gaming (anomalous decoupling) and find a replacement.

Calibrate data per benchmark

Conditional variance $\sigma^2_{j|\mathcal{A}}$ tells you how many samples a benchmark needs.
Weak-residual benchmarks deserve less compute.

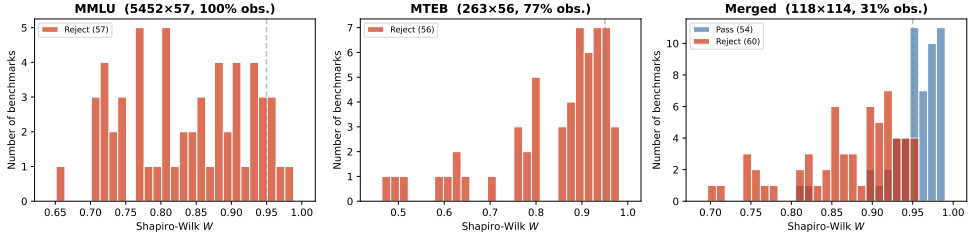
Thank You

Questions?

Alex Smola · smola@boson.ai

Submodular Benchmark Selection · arXiv:2605.02209

Backup: Normality Diagnostics



All 57 MMLU and all 56 MTEB benchmarks reject Gaussianity (Shapiro–Wilk, BH-corrected). Mardia $\hat{\beta}_{2,N} = 6\,374$ vs. null $3\,363$.

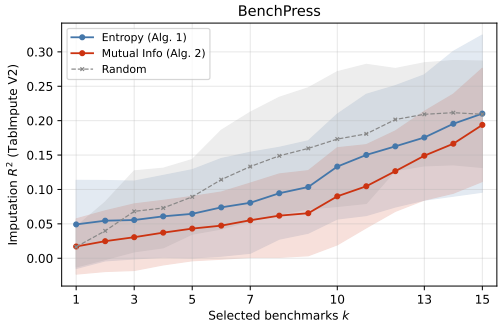
Doesn't matter: the conditional mean is the best *linear* predictor regardless of the true marginal distribution.

Backup: TabImpute — Nonlinear Imputation

Drop-in replacement: swap the Gaussian conditional mean for TabImpute V2 (Feitelberg et al. 2026), a pre-trained zero-shot transformer for tabular imputation.

R^2 at $k = 5$, MI selection:

Dataset	Gaussian	TabImpute
MMLU	**0.91**	0.61
MTEB	**0.76**	0.33
Merged	**0.51**	0.03
BenchPress	**−0.11**	0.04



BenchPress: neither imputer reaches $R^2 = 0.3$.

Gaussian *adapts* to the observed covariance; zero-shot cannot.

Backup: Logit-Space Transform

Motivation. Benchmark scores are bounded; a logit transform brings them closer to Gaussian.

Pipeline. Normalize to $[0, 1]$ via $s/s_{\max}$, clip to $[\varepsilon, 1-\varepsilon]$, apply logit; estimate, select, impute, invert with σ.

R^2 at $k = 5$, MI selection:

Dataset	Raw	Logit
MMLU	**0.91**	0.91
MTEB	**0.76**	0.72
MERGED	**0.51**	0.39
BENCHPRESS	**−0.11**	−0.22

Logit hurts (slightly).

The inverse map caps predictions at the training-set maximum, penalizing extrapolation to stronger validation models. Logit space is also slightly *less* compressible (higher effective rank).

Raw Gaussian framework already captures the dominant linear structure of benchmark score matrices, bound-aware transform doesn't add signal.