How Reliable is Language Model Micro-Benchmarking?

Gregory Yauney Shahzaib Saqib Warraich Swabha Swayamdipta USC

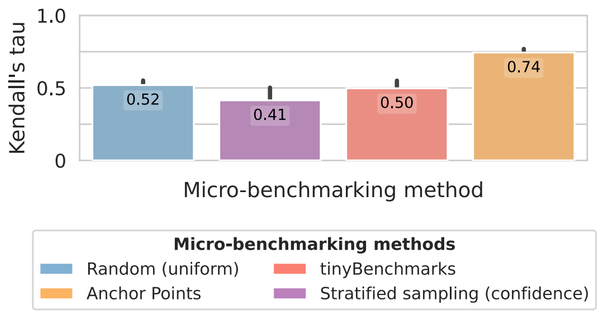
Micro-benchmarks reduce the size of evaluation datasets.

How reliably do micro-benchmarks echo the model performance judgments of full benchmarks?

We introduce a meta-evaluation measure to quantify how well micro-benchmarks preserve pairwise rankings of models as a function of the accuracy difference between models.

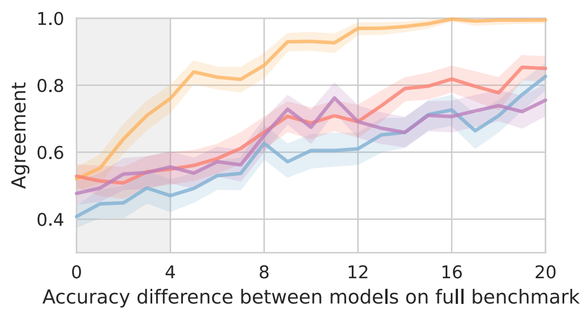
Aggregate correlation with full benchmark

Micro-benchmarks can yield model rankings that correlate with the full benchmark's rankings in the aggregate...

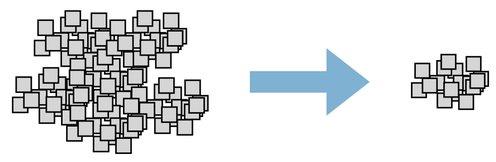


Fine-grained probability of correct pairwise rankings

...but micro-benchmarks often do not agree with full benchmarks when models differ by fewer than 4 points of accuracy.



Micro-benchmarking methods



Select a subset of examples using predictions from *source models* evaluated on the full dataset

Goal: predict *target model* accuracy on full dataset using just micro-benchmark

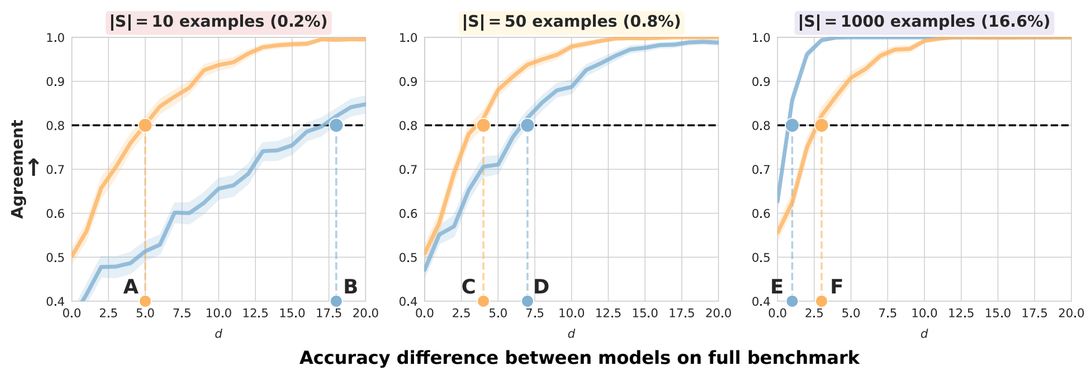
Random sampling: select examples uniformly at random

Anchor Points: cluster examples in space of model correlations (Vivek et al., EACL 2024)

tinyBenchmarks: train example embeddings with an auxiliary Item Response Theory model and cluster these (Polo et al., ICML 2024)

Probability micro-benchmark agrees with full benchmark

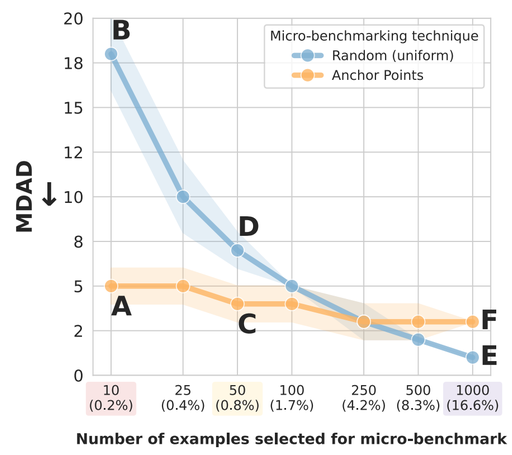
If model A outperforms model B on the full benchmark, does A also outperform B on the micro-benchmark?



Minimum Detectable Accuracy Difference (MDAD)

MDAD is the minimum accuracy difference at which a micro-benchmark agrees with the full benchmark ≥80% of the time.

Example: When selecting 10 examples from MMLU-Pro, Anchor Points has an MDAD of 5 (point A), but random sampling has an MDAD of 19 (point B).



Meta-evaluation experimental design

Datasets: MMLU, MMLU-Pro, BIG-bench Hard, GPQA

Models: >400 evaluated on the Open LLM Leaderboard

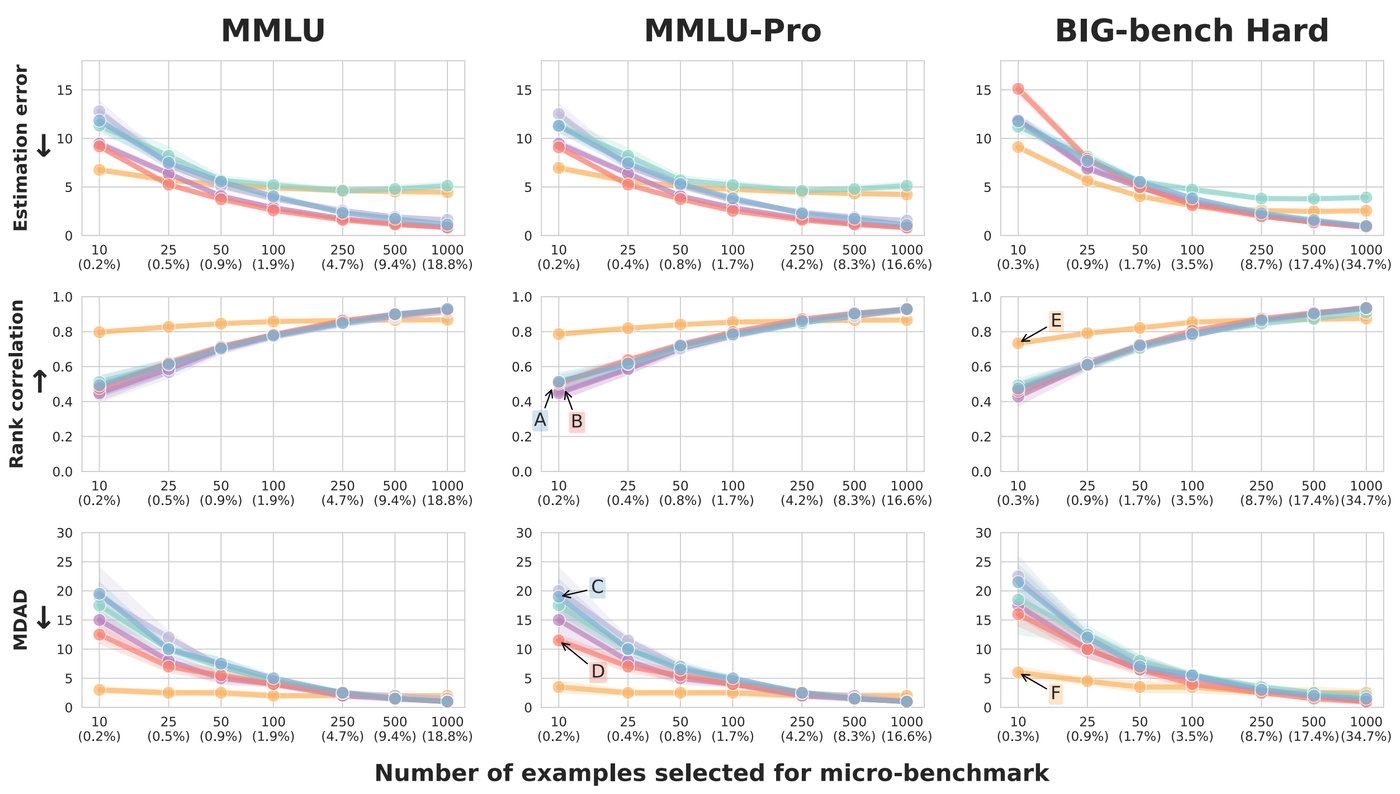
Existing measures: mean estimation error, rank correlation

Micro-benchmarking methods:

- Random (uniform)
- Random (subtask stratified, equal)
- Stratified sampling (confidence)
- Diversity
- Anchor Points
- tinyBenchmarks

For each trial: bootstrap a dataset, sample source and target models

Results



1. At extreme dataset reductions, methods struggle to distinguish models that differ by fewer than 3 points of accuracy.
2. Anchor Points can best distinguish models out of the methods evaluated.
3. When >200 examples are selected, random sampling can distinguish models as well as other methods.

MDAD offers complementary benefits to existing meta-evaluation measures:

→ Same rank correlations can map to different MDADs. (points A, B, C, D)

→ High rank correlation does not guarantee low MDAD. (points E, F)

Recommendations

1. Select a micro-benchmark size capable of distinguishing models at your desired resolution.
2. Larger micro-benchmarks are often needed to distinguish models with similar performances.
3. Random sampling with several hundred examples is often enough to reliably echo the full benchmark.

More in the paper!

- MDAD estimation error analysis
- Generalizing to new draws of datasets
- Explaining why top-ranked models can be predicted with few examples