MIT-IBM
Watson AI Lab

IBM Research

Evaluating and Routing LLMs Efficiently with Benchmarks

Mikhail Yurochkin

Evaluation of LLMs

Capability	Benchmark Higher is better	Description	Gemini Ultra
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	90.0% CoT@32*
Reasoning	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	83.6% 3-shot
	DROP	Reading comprehension (F1 Score)	82.4 Variable shots
	HellaSwag	Commonsense reasoning for everyday tasks	87.8% 10-shot*

LLM Benchmarks

HELM

116 Scenarios

BIG-bench

200+ Tasks

Open LLM Leaderboard

MMLU, HellaSwag, ...
Results for 1000+ models!

tinyBenchmarks: evaluating LLMs with fewer examples

Felipe Maia Polo, Lucas Weber, Leshem Choshen,
Yuekai Sun, Gongjun Xu, *Mikhail Yurochkin*

🤗 Open LLM Leaderboard

- MMLU: 14042 inputs
- HellaSwag: 10042 inputs
- …
- Total: ~30k inputs, 6 scenarios

Main Ideas

Estimate performance of a new LLM using much smaller number of inputs

Identify inputs most helpful to quantify LLM abilities

Demonstration

[Colab LINK]

Item Response Theory

does gre use item response theory

In irt | Images | Perspectives | Videos | News | Shopping | Books | Maps | Flights

About 2,720,000 results (0.39 seconds)

Today, all major educational tests, such as the Scholastic Aptitude Test (SAT) and Graduate Record Examination (GRE), are developed by using item response theory, because the methodology can significantly improve measurement accuracy and reliability while providing potentially significant reductions in assessment time...

Item Response Theory

- Represent each example as a vector of *required skills* and each LLM as a vector of *abilities*
- LLM performs well on an example if its abilities match the required skills
- A small subset of examples is enough to “test” LLM abilities

Item Response Theory

$$p_{il} \triangleq \mathbb{P}(Y_{il} = 1 | \theta_l, \alpha_i, \beta_i) = \frac{1}{1 + \exp(-\alpha_i^\top \theta_l + \beta_i)}$$

Difficulty of example i

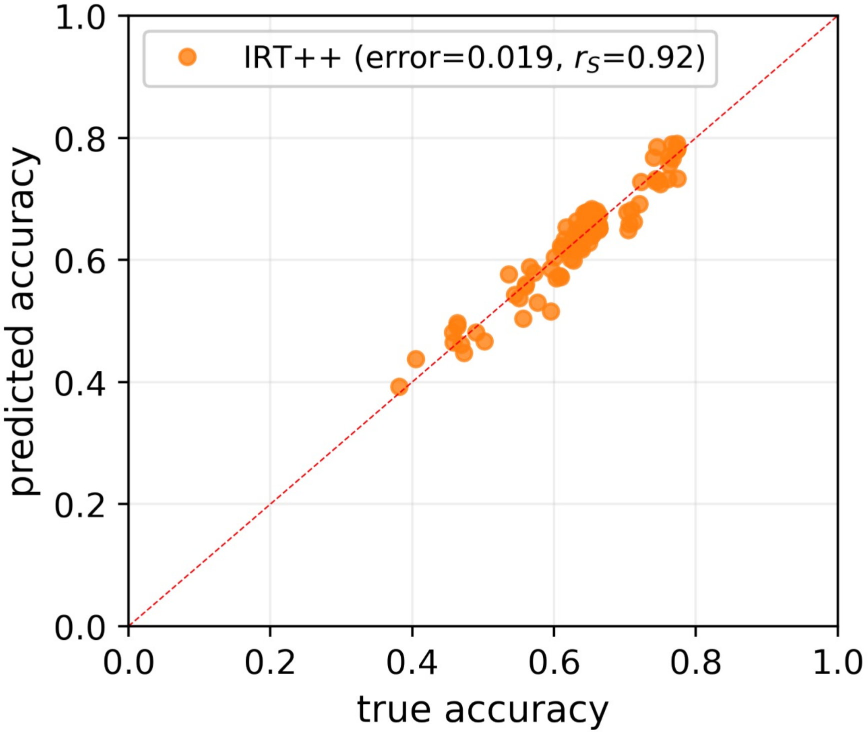
Required skills for example i

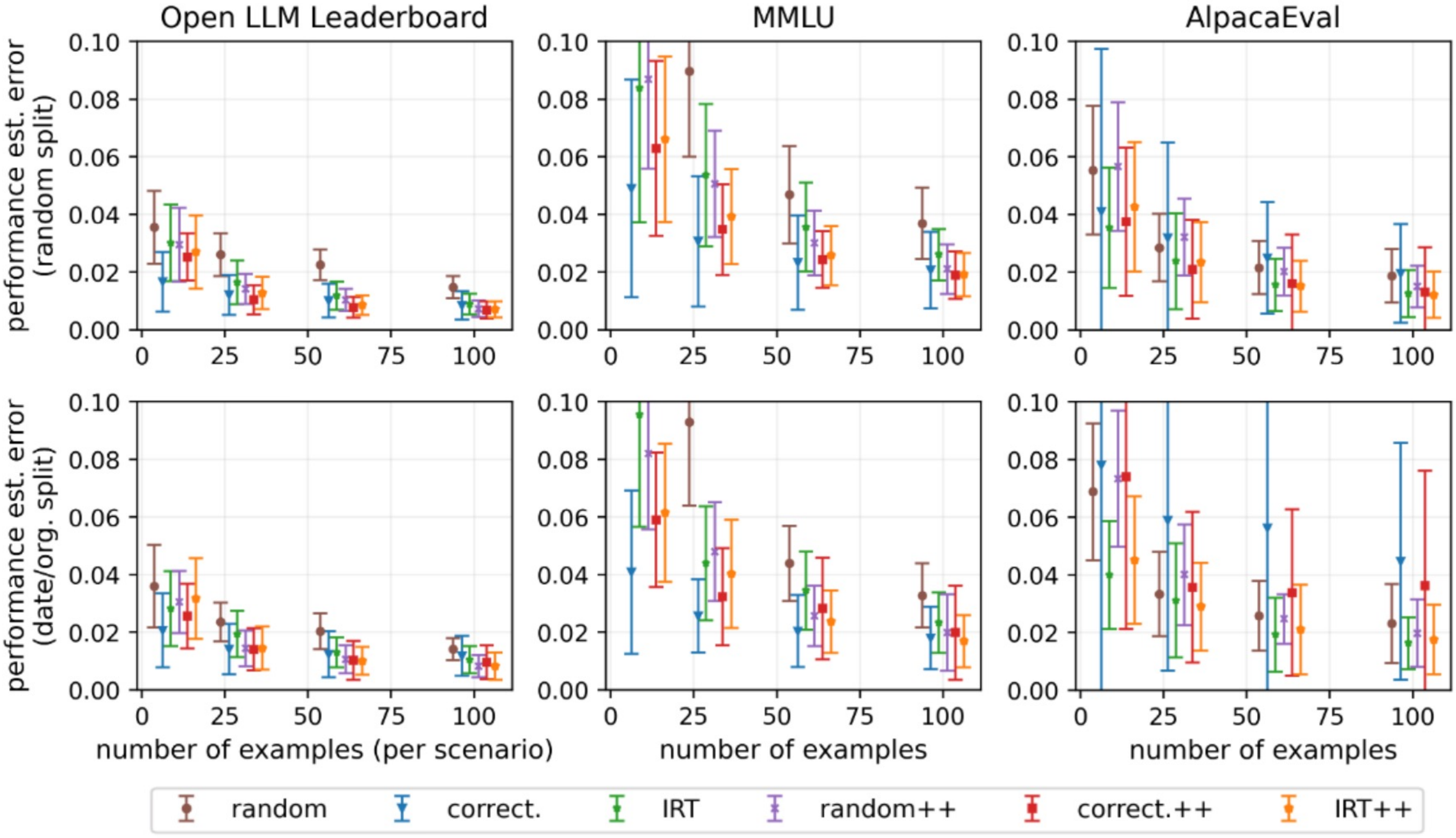
Abilities of model l

1. Cluster (α_i, β_i) to find a "tiny" subset of points to evaluate
2. Estimate θ_l for a new LLM using evaluations on the "tiny" subset
3. Predict correctness on the remaining examples
4. Combine observed and predicted correctness to estimate overall performance

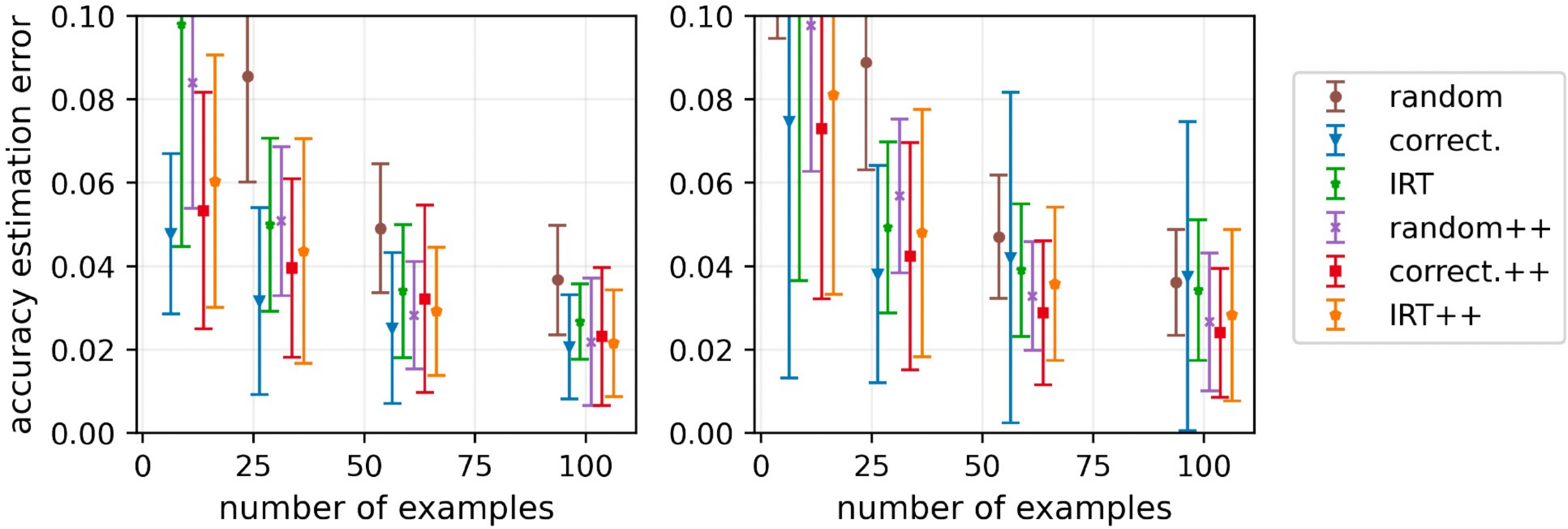
tinyMMLU = 100 examples

Saving 140x compute!




Open LLM Leaderboard
MMLU
AlpacaEval
performance est. error (random split)
performance est. error (date/org. split)
0.00
0.02
0.04
0.06
0.08
0.10
0
25
50
75
100
number of examples (per scenario)
number of examples
number of examples
random
correct.
IRT
random++
correct.++
IRT++

MMLU: Specialized Models



Demonstration

Colab LINK

Limitation or a Feature?

Mikhail Yurochkin 11:48 PM

I checked our method on a “flagged” model that has been confirmed to have MMLU in its train set.

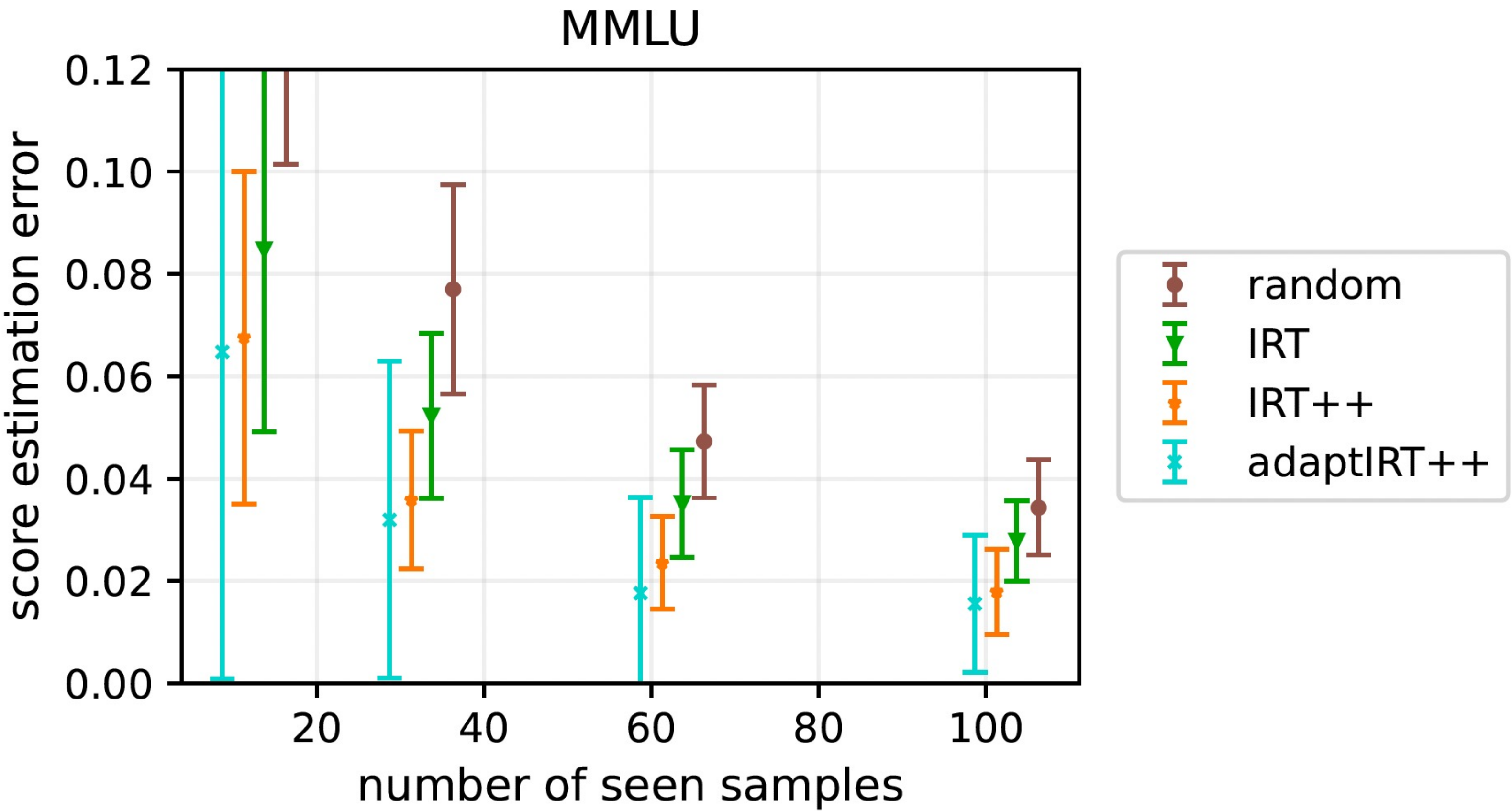
```
Model: open-llm-leaderboard/details_Aspik101__trurl-2-13b-pl-instruct_unload
- True accuracy:                                        0.787
- Predicted accuracy based on anchor points (IRT):      0.756
- Predicted accuracy based on p-IRT:                    0.716
- Predicted accuracy based on gp-IRT (IRT++):           0.721
```

timje Aug 25, 2023

Further clarification for anyone (like me) who missed the Voicelab discussion, the trurl-2-13b model's training *included* much of the MMLU test, so of course it scores exceedingly well on the test for a 13b model. The Voicelab team is re-training without the MMLU dataset but doesn't expect much difference from base llama-2-13b; their focus is on Polish knowledge.

👍 12 +

MMLU: Choosing samples adaptively



Benchmarking Prompts

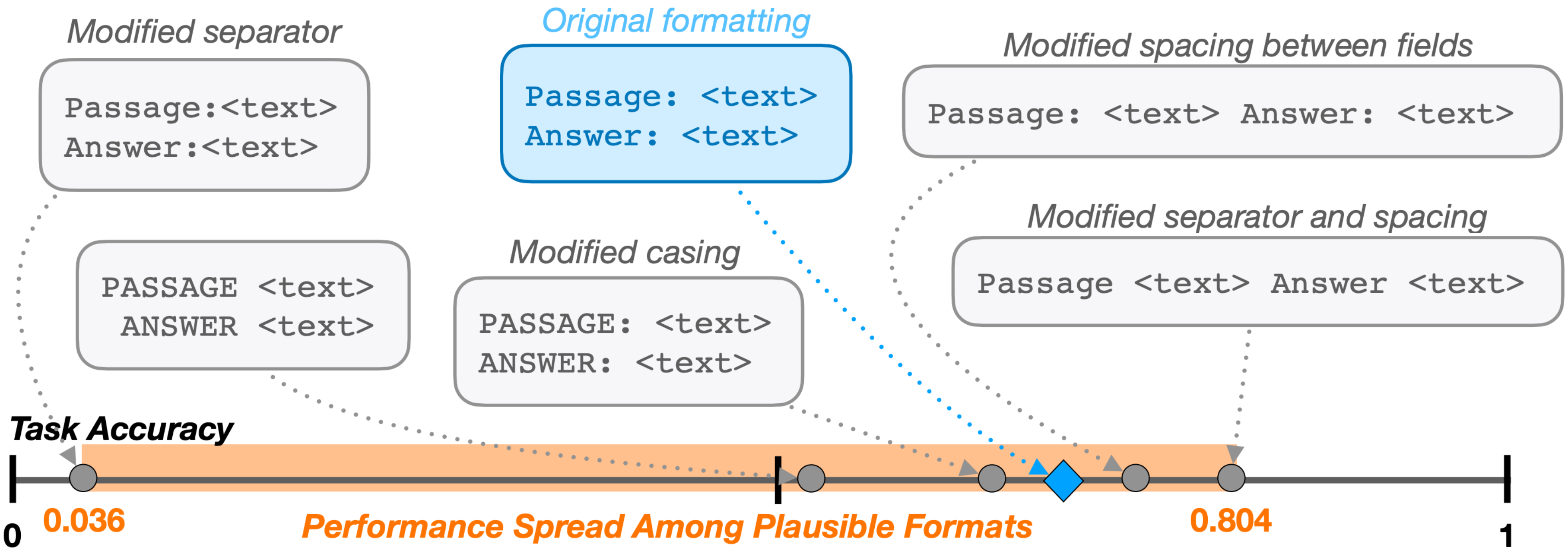
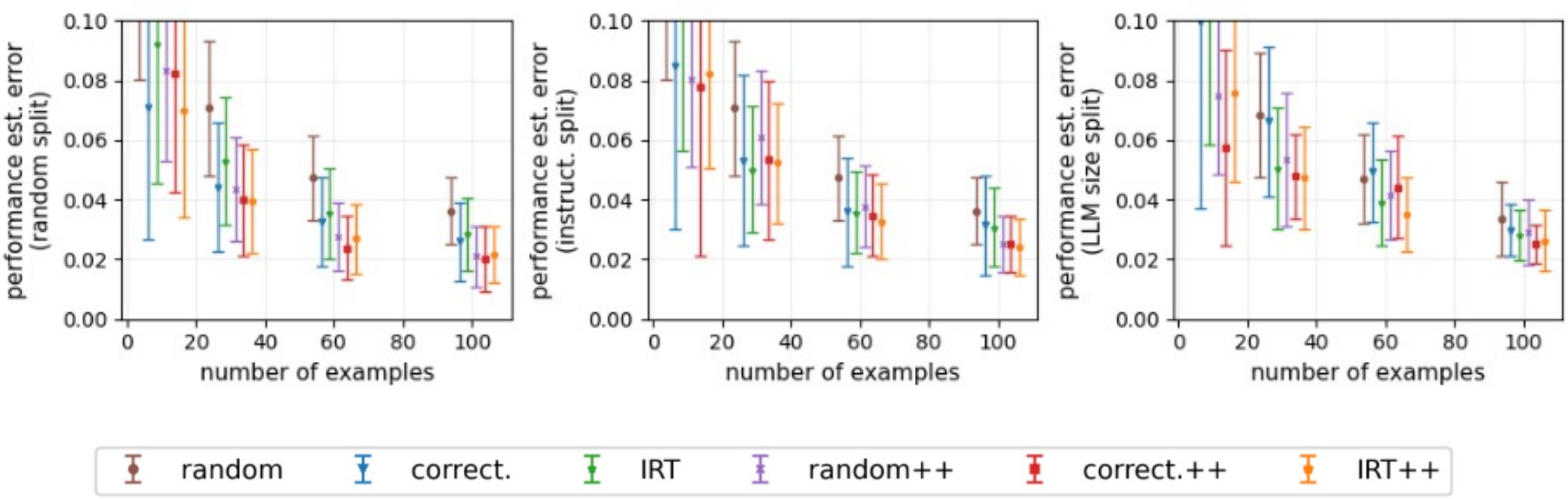


Image is from "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting" (Sclar et al., 2023)

Benchmarking Prompts



tinyBenchmarks on Hugging Face

Datasets 7

Sort: Recently updated

- tinyBenchmarks/tinyMMLU
  Viewer • Updated Mar 12 • 2.91k • 15
- tinyBenchmarks/tinyWinogrande
  Viewer • Updated Mar 7 • 2.63k • 1
- tinyBenchmarks/tinyAI2_arc
  Viewer • Updated Mar 7 • 2.12k • 3
- tinyBenchmarks/tinyHellaswag
  Viewer • Updated Mar 7 • 3.2k • 2
- tinyBenchmarks/tinyGSM8k
  Viewer • Updated Mar 7 • 1.73k • 3
- tinyBenchmarks/tinyTruthfulQA
  Viewer • Updated Mar 7 • 2.11k • 2
- tinyBenchmarks/tinyAlpacaEval
  Viewer • Updated Mar 7 • 20 • 2

~15k downloads last month

Questions?

LLM Routing with Benchmark Datasets

Tal Shnitzer*, Anthony Ou*, Mírian Silva, Kate Soule, Yuekai Sun,
Justin Solomon, Neil Thompson, *Mikhail Yurochkin*

LLM Benchmarks

HELM

116 Scenarios

BIG-bench 🪑

200+ Tasks

🤗 **Open LLM Leaderboard**

MMLU, HellaSwag, ...

Results for 1000+ models!

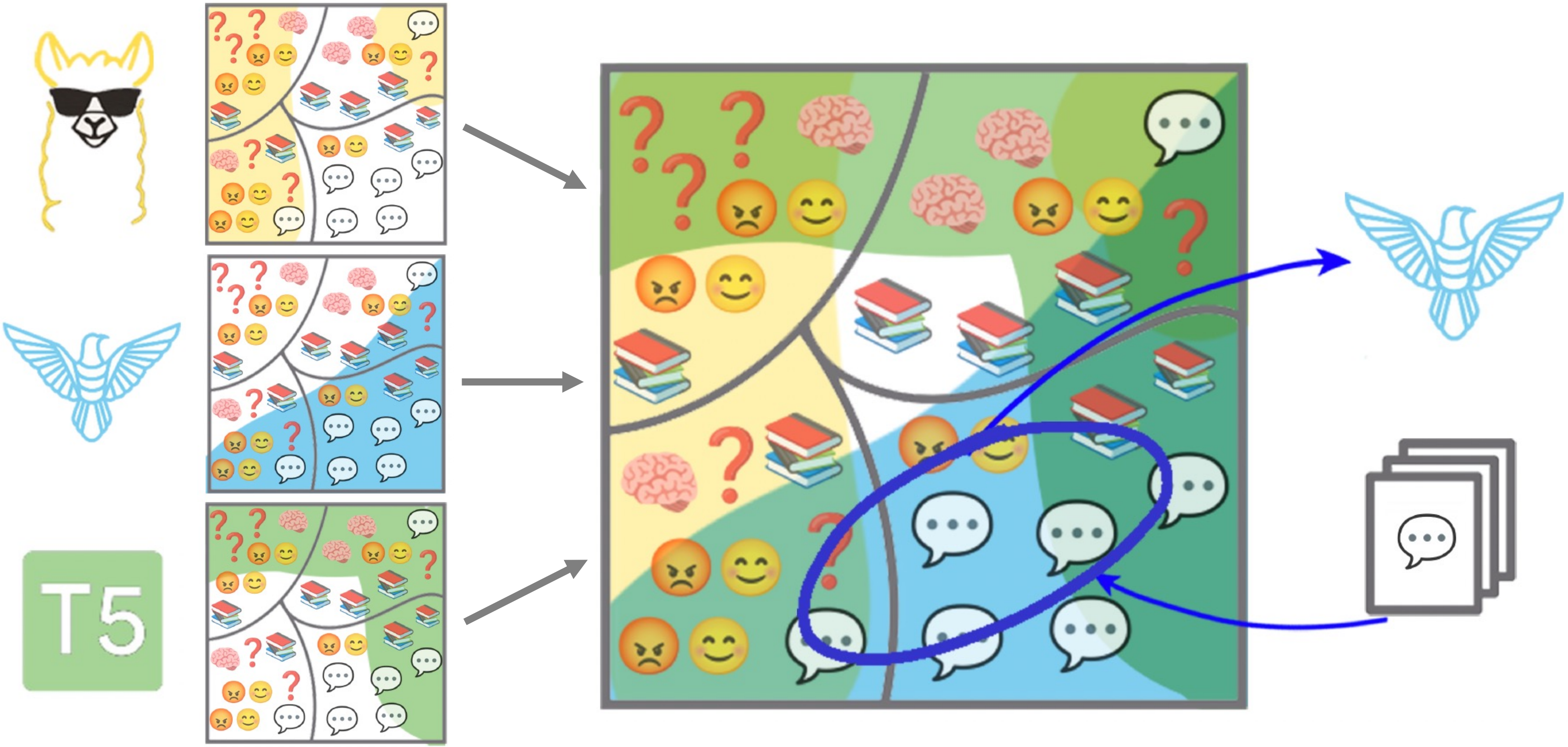
Main Ideas

Use LLM Benchmark evaluations
to Learn their Strengths:
simple classification problem

Choose LLM for a New Task using
"Correctness Predictors"

Learning "correctness" of LLMs

LLM Routing



Relation to OOD Generalization

- New tasks are likely to differ from Benchmark Datasets
- If we can predict correctness of LLMs OOD accurately, we are good to go!
- But it is hard...
- LLM routing with imperfect predictions

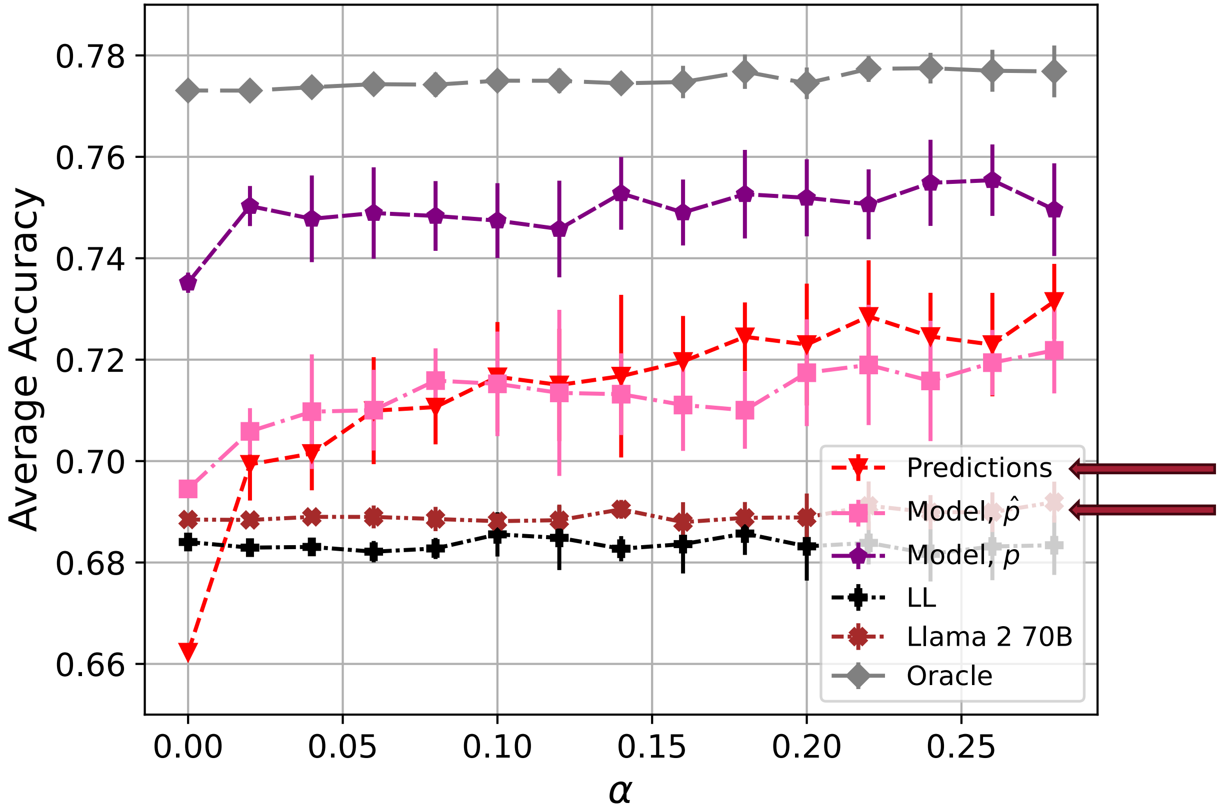
Modeling Correctness

- Correctness predictor for an LLM $\bar{g} : \textit{Input Text} \to \{0, 1\}$
- Correctness of an LLM $y : \textit{Input Text} \to \{0, 1\}$
- Standard goal: achieve high accuracy $\sum_i \mathcal{I}(\bar{g}(x_i) = y(x_i))$
- LLM routing: estimate $\sum_i y(x_i)$ for a new task d'

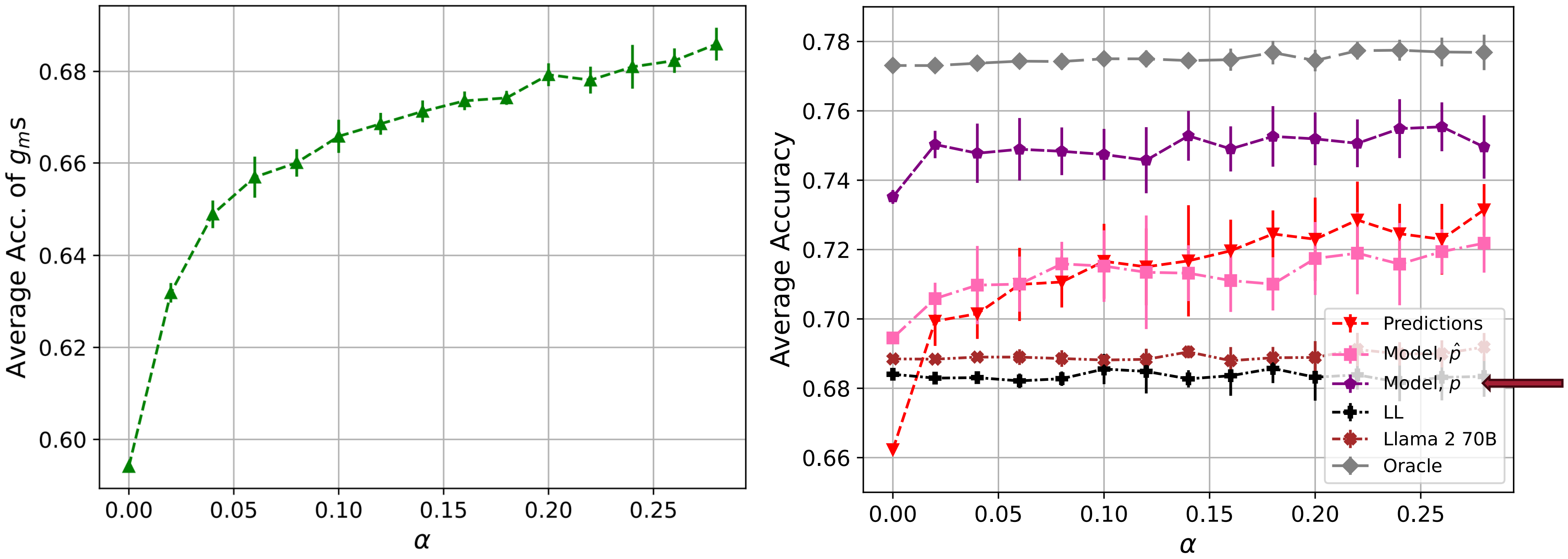
$$y(x)|x, d' = \begin{cases} \bar{g}(x) & \text{with probability } p(d') \\ 1 - \bar{g}(x) & \text{with probability } 1 - p(d'), \end{cases}$$

where $p(d')$ is accuracy of $\bar{g}$ on d'.

LLM Routing outperforms largest model (Llama 2 70B) using fewer (~40B-50B) parameters on average



Making decisions OOD with Imperfect Predictors



Thank You!

Questions?