How Predictable Are Large Language Model Capabilities? A Case Study on BIG-bench

Qinyuan Ye, Harvey Yiyun Fu, Xiang Ren, Robin Jia {qinyuany, harveyfu, xiangren, robinjia}@usc.edu

USC University of Southern California

TL;DR

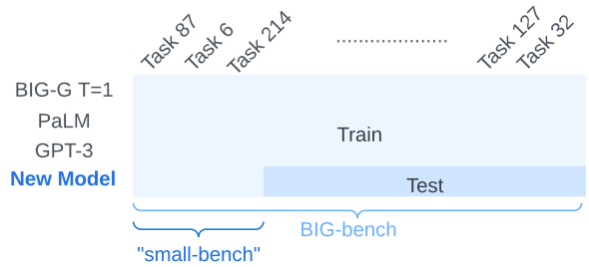
We gathered 56k LLM experiment records in BIG-bench.

Performance Prediction

Model Family	# param	Task	# shot	Perf.
GPT-3	3B	strategy_qa	0	0.48
BIG-G T=1	8B	elementary_math	3	0.19
PaLM	64B	code_line_desc	2	0.23
GPT-3	6B	elementary_math	1	?

We trained models to predict LLM performance on unseen experiment configurations.

Searching for "small-bench"



We searched for a subset of BIG-bench, from which the full BIG-bench performance can be maximally recovered.

Motivation

LLM User

What model scale should I use?

LLM Developer

What tasks should I prioritize in evaluation?

LLM Researcher

Which capabilities are hard to predict?

Dataset

https://github.com/google/BIG-bench

# Experiment Records	56,143
# Model Families	6
	BIG-G T=0, BIG-G T=1, BIG-G Sparse, PaLM GPT-3, Gopher
# Models†	51
# BIG-bench Tasks	134
# BIG-bench Subtasks‡	313
$\{n_{shot}\}$	$\{0, 1, 2, 3, 5\}$

56k records

diverse models

diverse tasks

Part 1: Performance Prediction on BIG-bench

Problem Definition

Normalized Performance $\hat{y} = f(l, n_{param}, t, n_{shot})$

l: Model Family; n_{param}: # Parameters; t: Tasks; n_{shot}: # In-context Examples

Regression Problem. Evaluated with **RMSE** and **R^2 score**.

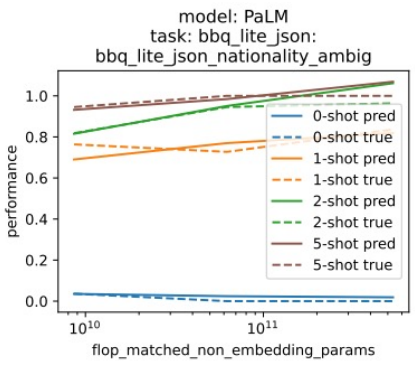
Results and Findings

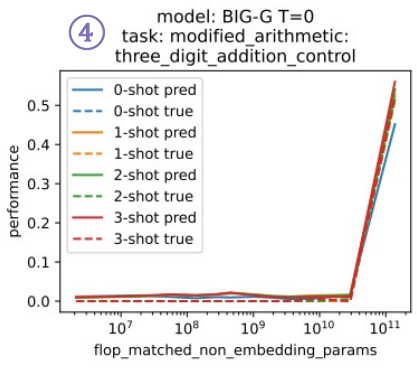
① Gradient boosted trees and MLPs can achieve RMSE < 5%, R^2 > 95% on the random train-test split.

② Prediction performance drops when the train-test split becomes more challenging.

③ Zero-shot performance and experiments with larger models are harder to predict.

④ Emergent abilities are harder to predict in general but can be predicted accurately in some cases.





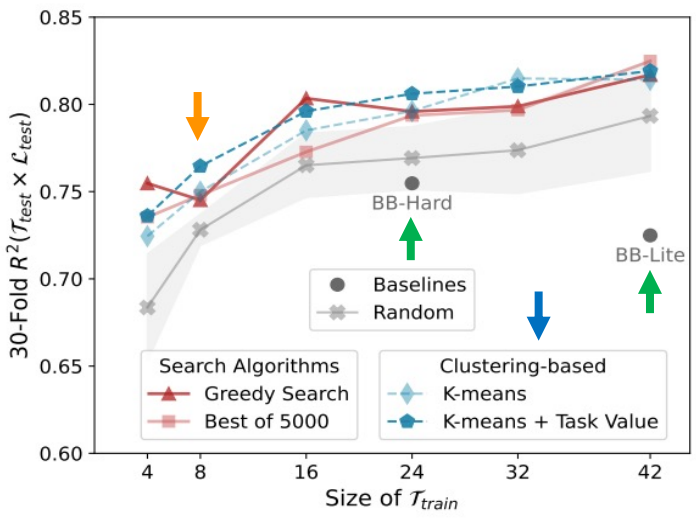
Part 2: Searching for "small-bench"

Problem Definition

Given an evaluation budget of *b* select *b* tasks such that performance on remaining tasks are *maximally* recovered.

$$\arg\max_{\mathcal{T}_{train}} R^2(\mathcal{T}_{test} \times \mathcal{L}_{test}) \quad \text{s.t.} \quad \mathcal{T}_{train} \subseteq \mathcal{T}, |\mathcal{T}_{train}| = b$$

Results and Findings



① BIG-bench Lite and BIG-bench Hard are suboptimal if the goal is to recover the performance on remaining tasks.

② We are able to find subsets that are as informative as BIG-bench Hard while being 3x smaller.

③ Task diversity and task value are important factors in constructing "small-bench."

Broader Discussions + Future Work

Rethinking LLM Evaluation

There is a lack of consensus regarding LLM evaluation. Task selection is often heuristic and following past practices. How to evaluate LLMs efficiently, reliably and rigorously?

Broadening Observations on LLM Capability Landscape

Integrating LLM experiment records from other evaluation efforts (e.g., HELM). Adding more dimensions of experiment configurations (e.g., instruction tuning, RLHF, prompting).