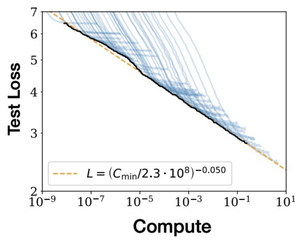
Observational Scaling Laws and the Predictability of Language Model Performance

Yangjun Ruan[123] Chris J. Maddison[23] Tatsunori Hashimoto[1]

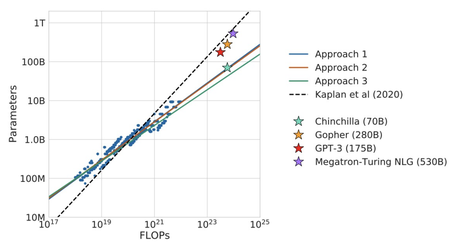
[1]Stanford University [2]University of Toronto [3]Vector Institute

Overview

Scaling laws are critical tools for understanding and predicting model capabilities and algorithmic development



Capability prediction [Kaplan, et al., 2020]



Resource allocation [Hoffmann, et al., 2022]

Limitations of conventional, compute-based scaling analyses:

☹ **substantial cost**: requires training a large model family across scales

☹ **restricted coverage**: requires controlled interventions & training recipes

Contribution: an **observational scaling** approach that is

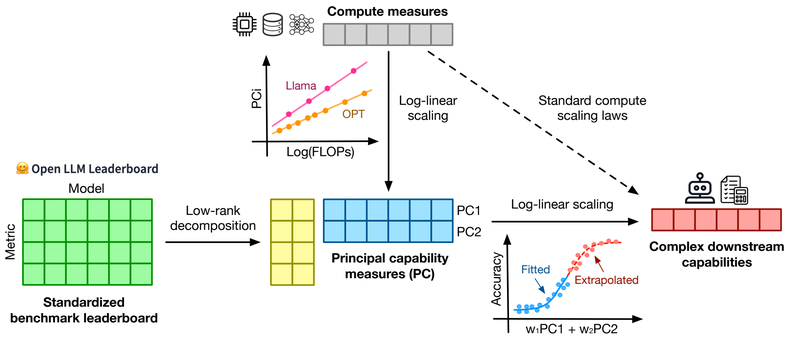
☺ **lower-cost**: no model training required

☺ **higher-resolution**: utilizing a large set of existing public models

☺ **broader-coverage**: covering different model families & scaling dims.

Observational Scaling Laws

Motivation: there are many public models & standard eval benchmarks



Idea: extract **low-rank capability measures** from observable simple metrics

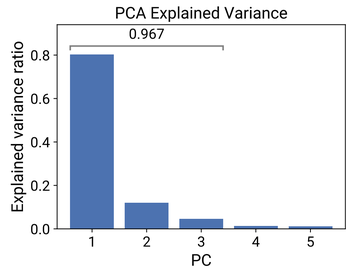
✔ unify heterogenous model families with varying compute efficiencies

✔ relate training compute to complex downstream capabilities

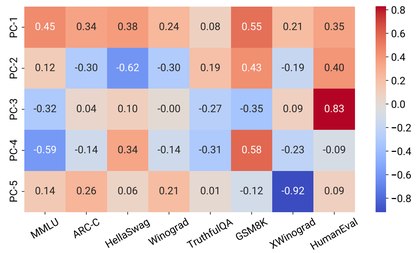
Principal Capabilities as Surrogate Scale

Approach: apply PCA to model-benchmark matrix using 100+ public models & diverse benchmarks

PC Analysis: PC measures are **low-dim** and **interpetable**

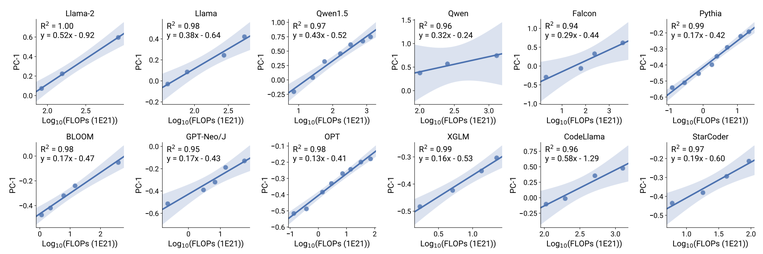


PCA explained variance



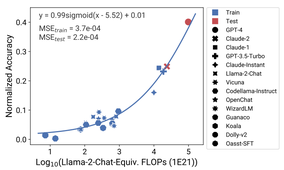
Principal component weights

PC Scaling: PC measures **linearly correlate** with log-FLOPs within each model family

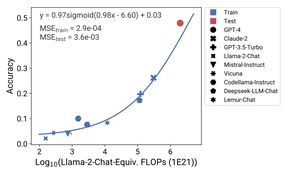


Predictability of Agentic Capabilities

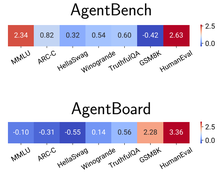
Question: how do LMs' capabilities as agents scale?



AgentBench



AgentBoard

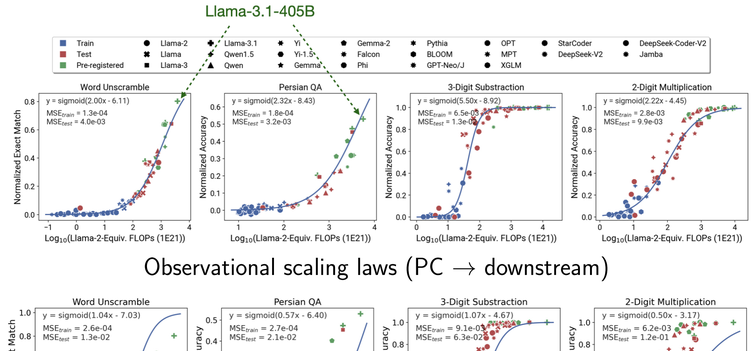


Fitted weights

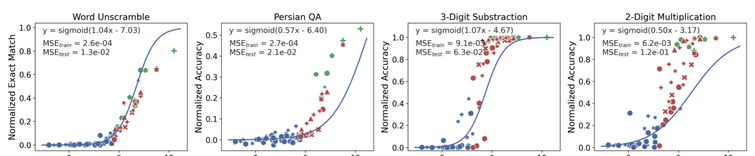
- LMs' agentic capabilities can be accurately predicted by their **simple benchmark metrics**
- **Programming capabilities** are essential for agents

Predictability of "Emergent" Capabilities

Question: is "emergence" an artifact of **low-resolution** data?



Observational scaling laws (PC → downstream)



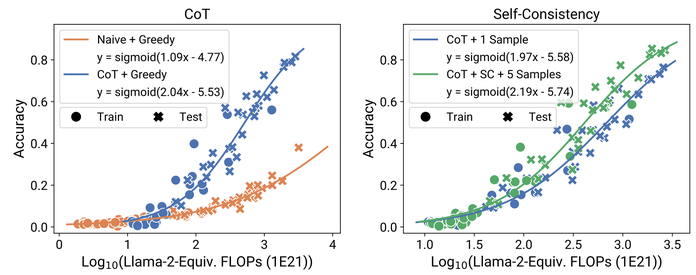
Compute scaling law (training FLOPs → downstream)

- **High-resolution** obs. scaling accurately predicts emergent capabilities from sub-Llama-2-7B → **Llama-3.1-405B**
- Compute-based scaling demonstrates **poor** performance

Impact of Post-Training Techniques

Question: can we predict which post-training techniques will maintain effective gains **at larger scales**?

Approach: fit & compare the scaling w/ & w/o the techniques



- LM performance with post-training methods are **predictable**
- Different techniques demonstrate different scaling properties