

# Emergent Abilities

## Large Language Models (LLMs)

### Learning goals

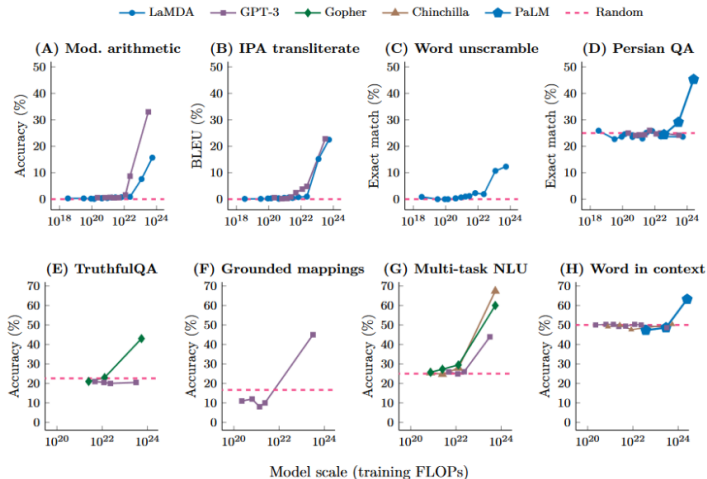
- illustrate emergent abilities that LLMs reveal when they are scaled up
- discuss a counterargument for the concept of emergence

# EMERGENT ABILITIES

An *emergent ability* is an ability that is not present in small models but is present in large models. (if everything else is held constant – not really possible)

- Is emergence a rare phenomenon?
- Are many tasks emergent?
- We need to observe scaling models
  - GPT-3
  - Chinchilla
  - PaLM

# EMERGENT ABILITIES AND MODEL SIZE



**Figure:** Examples of emergence in few-shot prompting. ► Wei et al., 2022

# EMERGENT TASKS IN BIG-BENCH

- “MODEL SIZE (TASK)” means: MODEL can do TASK with SIZE, but not with less than SIZE (hence emerging)
- GPT-3 13B (2 tasks): hindu knowledge, modified arithmetic
- GPT-3 175B (15 tasks): analytic entailment, codenames, phrase relatedness, question answer creation, self evaluation tutoring, ...
- LaMDA 137B (8 tasks): gender inclusive sentences german, repeat copy logic, sports understanding, ...
- PaLM 8B (3 tasks): auto debugging, sufficient information, parsinlu reading comprehension
- PaLM 64B (14 tasks): anachronisms, ascii word recognition, conceptual combinations, ...
- PaLM 540B (25 tasks): analogical similarity, causal judgment, code line description, crass ai, cs algorithms, ...

# EMERGENT TASKS IN MMLU

- Chinchilla 7B (7 tasks): Professional Medicine, High School Statistics, High School Macroeconomics, High School Psychology, Anatomy, High School Government And Politics, High School Microeconomics
- Chinchilla 70B (44 tasks): International Law, Human Aging, Sociology, Us Foreign Policy, High School World History, Marketing, Logical Fallacies, Miscellaneous, College Biology, High School Us History, Security Studies, High School European History, ...

# OTHER EMERGENT TASKS

- GPT-3 paper: 3 digit addition/subtraction (GPT-3 13B), 4-5 digit addition/subtraction (GPT-3 175B), leveraging few-shot examples for word denoising (GPT-3 13B)
- Gopher paper: Toxicity classification (Gopher 7.1B), TruthfulQA (Gopher 280B)
- Patel & Pavlick: grounded conceptual mappings (GPT-3 175B)
- PaLM paper: Word in Context benchmark (PaLM 540B)

# COUNTER ARGUMENT

► Source: Schaeffer et al., 2023

Two defining properties for emergent abilities in LLMs:

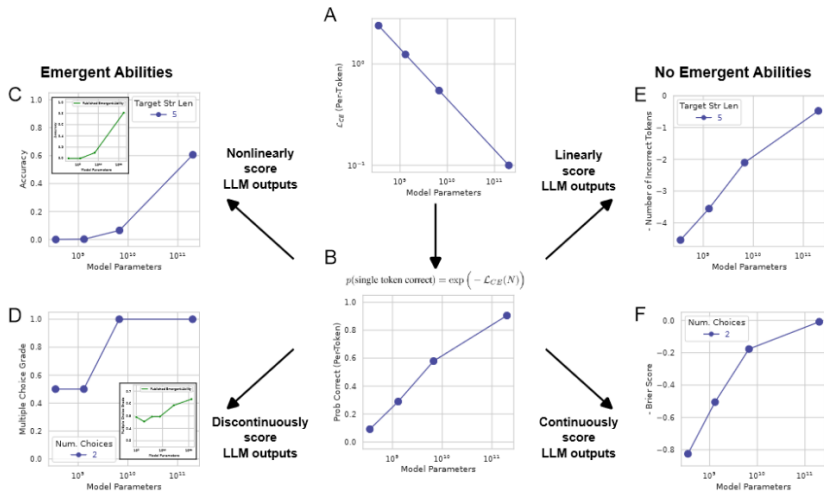
- ❶ **Sharpness:** transitioning seemingly instantaneously from not present to present
- ❷ **Unpredictability:** transitioning at seemingly unforeseeable model scales

Claim: Emergent abilities appear only under metrics that nonlinearly or discontinuously scale model's per-token error rate:

Multiple Choice Grade  $\stackrel{\text{def}}{=} \begin{cases} 1 & \text{if highest probability mass on correct option} \\ 0 & \text{otherwise} \end{cases}$

Exact String Match  $\stackrel{\text{def}}{=} \begin{cases} 1 & \text{if output string exactly matches target string} \\ 0 & \text{otherwise} \end{cases}$

# COUNTER ARGUMENT



► Source: Schaeffer et al., 2023



# COUNTER ARGUMENT

## Conclusions:

- If TASK is emergent for family MODEL at SIZE on METRIC, then it is often possible to choose another metric for which the task is not emergent.
- Emergent abilities can be induced in computer vision tasks as well
- A task and a metric are distinct and meaningful choices when constructing a benchmark
- When choosing a metric, one should consider the metric's effect on the per-token error rate; may require adapting the measuring process

# COUNTER COUNTER ARGUMENT

- Subjectively, there was clearly some form of “psychological” emergence when GPT3 came out – and probably with several other models as well.