Emergent Abilities of Large Language Models

Jason Wei,Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, William Fedus

Samarendra Dash

2025-01-23

Figure 1: What is emergence?

Emergence

What is emergence?

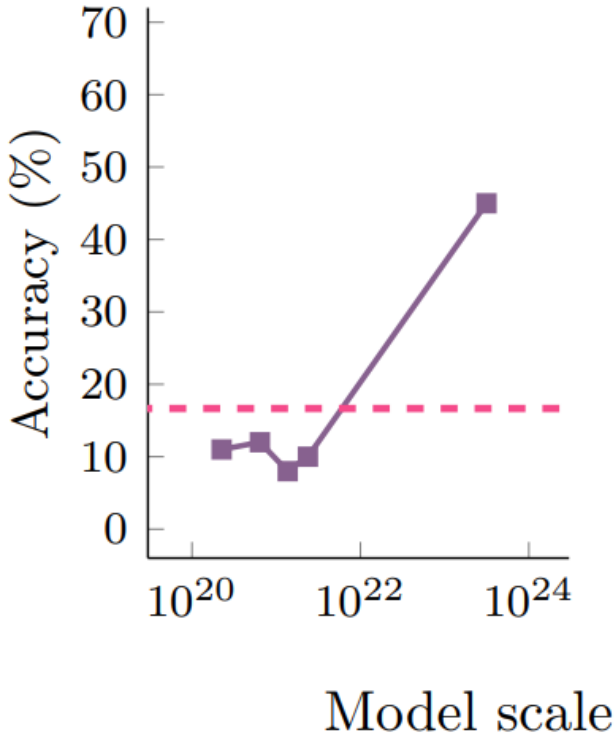


Figure 2: Example of Emergence in GPT3 model family

Emergence

- Emergence is defined as a property that is observed in an complex entity but not part of it's constituent. Examples may include consciousness, bird flocking.
- For the discussion pertaining to LLMs the authors of Wei et al. [9] define emergence as: **An ability that is not present in smaller model, but present in larger models in a way that couldn't have been predicted by extrapolating the scaling laws**.

Motivation

PaLM [3] beats GPT3 [2] on WiC benchmark, disproving the earlier hypothesis of Decoder-Only Models being bad for the benchmark.

	Word in Context (WiC)
Fine-tuned SOTA	76.1
Fine-tuned BERT-Large	69.6
GPT-3 Few-Shot	**49.4**
PaLM Few-Shot	**64.6**

Table 1: Performance of Models on WiC Benchmark

Motivation

Simply by scaling in number of parameters, training dataset size, and/or training compute, a massive performance boost is observed. A subset of those performances can't predicted by simply extrapolating the scaling laws.

Emergence in LLMs

- LLMs are scaled majorly in the following dimensions, i.e. Model Size (number of parameters), Total compute need for training (Total number of Floting Point Operations/FLOPs) and Training dataset size.
- Wei et al. [9] explored the emergence behaviors against training compute.**No claims are made about achieving emergence at a certain scale.** The aim is just to observe the behavior.

Emergence in LLMs

The following tasks were found to be emergent by Wei et al. [9].

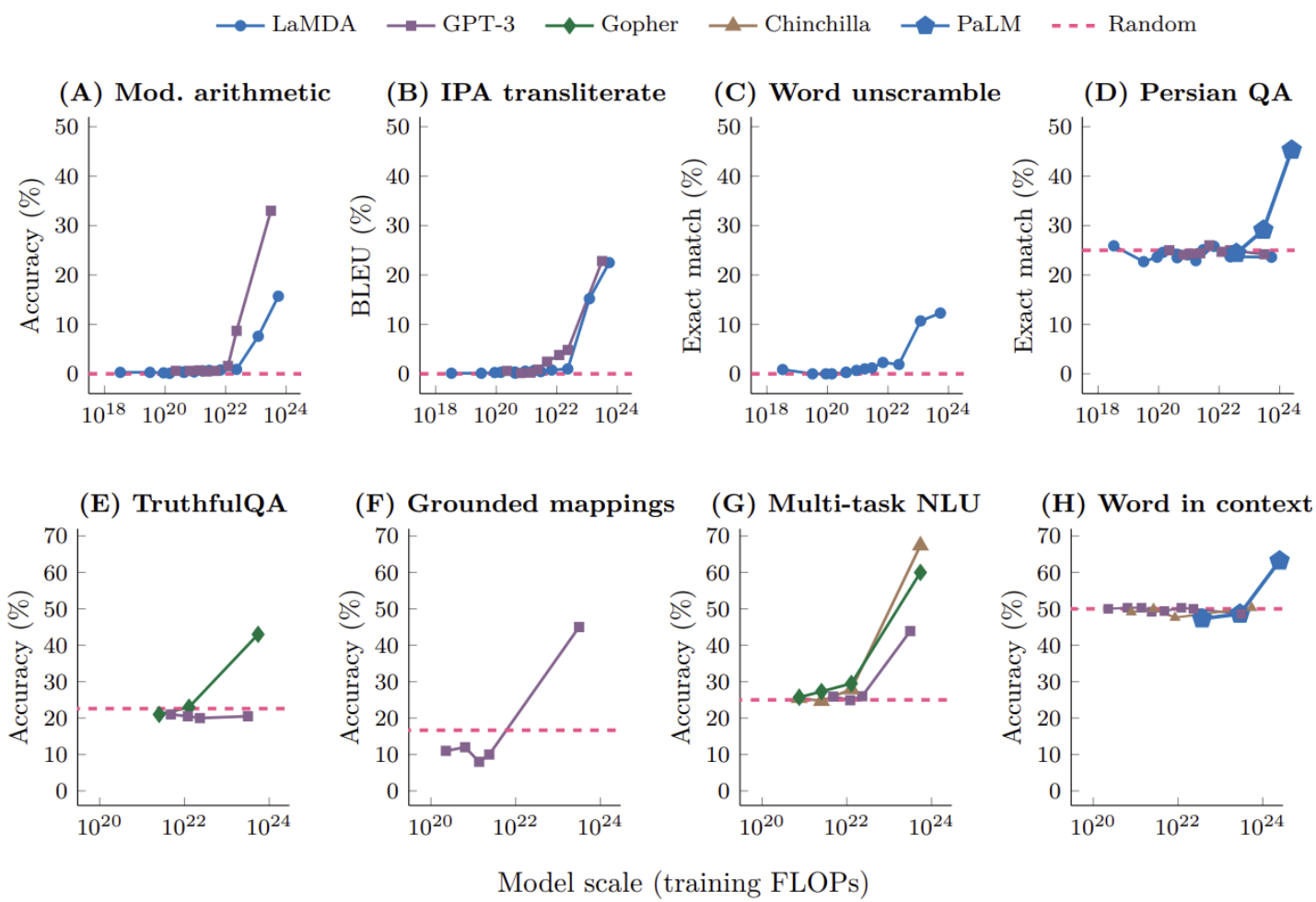


Figure 3: Emergent Tasks

Prompting and Finetuning as emergence

A prompting or finetuning technique that is only helpful at a certain scale is also considered emergent [9].

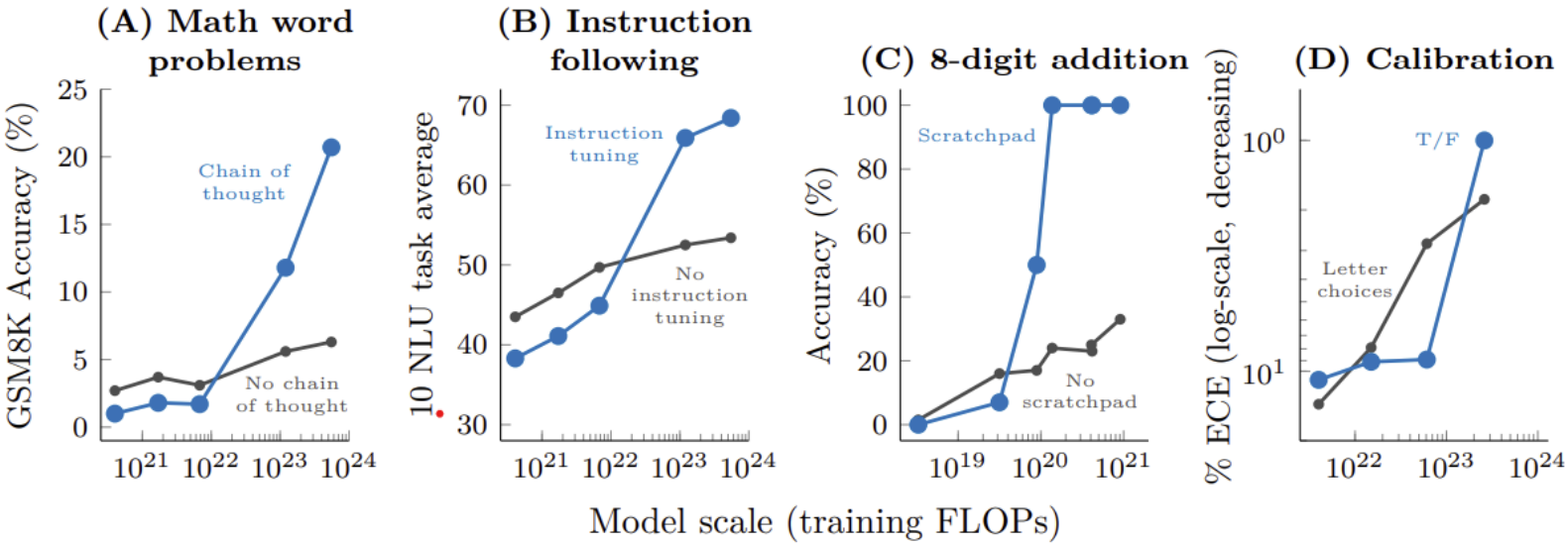


Figure 4: Specialized prompting or finetuning methods as emergence

Discussion

- Frontier Tasks
  - Models with their current scale give random performance on tasks such as, “checkmate in one”, “mathematical induction”, “multistep arithmetic” etc.
- Improving Architechture
  - Many abilities earlier only seen in larger models, is achieved at a much smaller scale in a later generation. Example LLAMA 3.2 models [8].
  - More study in model architecture and emergence could be the key to unlocking these abilities at much smaller scale.

So, is emergence real?

Limitation

- Srivastava et al. [7] suspect the emergence trends to be an artifact of the metrics being discrete in nature. Wei et al. [9] also have observed the difference between using discrete and continuous metrics, but fail to provide a theory to explain the discrepancy.
- Srivastava et al. [7] prove the hypothesis and provide an explanation. **Discrete metrics fail to reliably capture incremental improvements in the smaller models accurately.**

Limitation

Using a metric with higher resolution causes emergence trends to disappear [6].

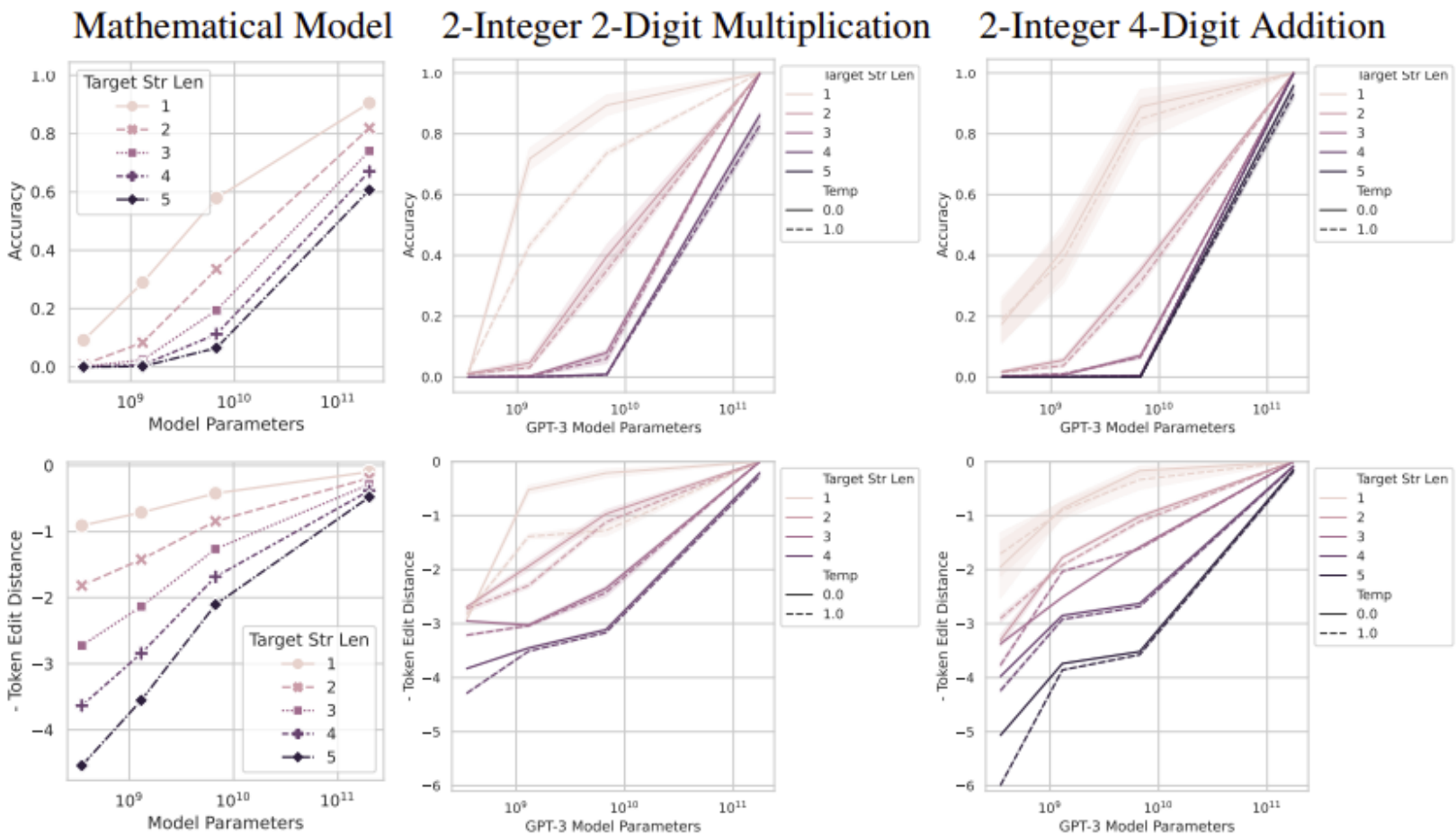


Figure 5: Using Accuracy vs Token-Edit-Distance metric

Limitation

Using more test points (i.e. providing more resolution) also causes emergence trends to disappear [6].

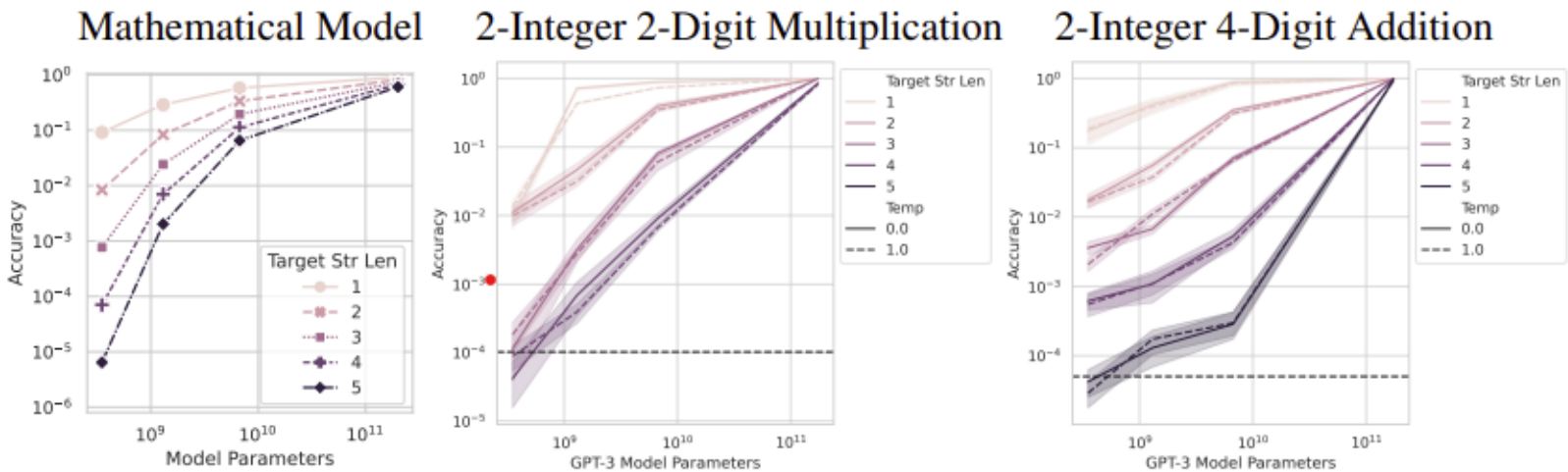
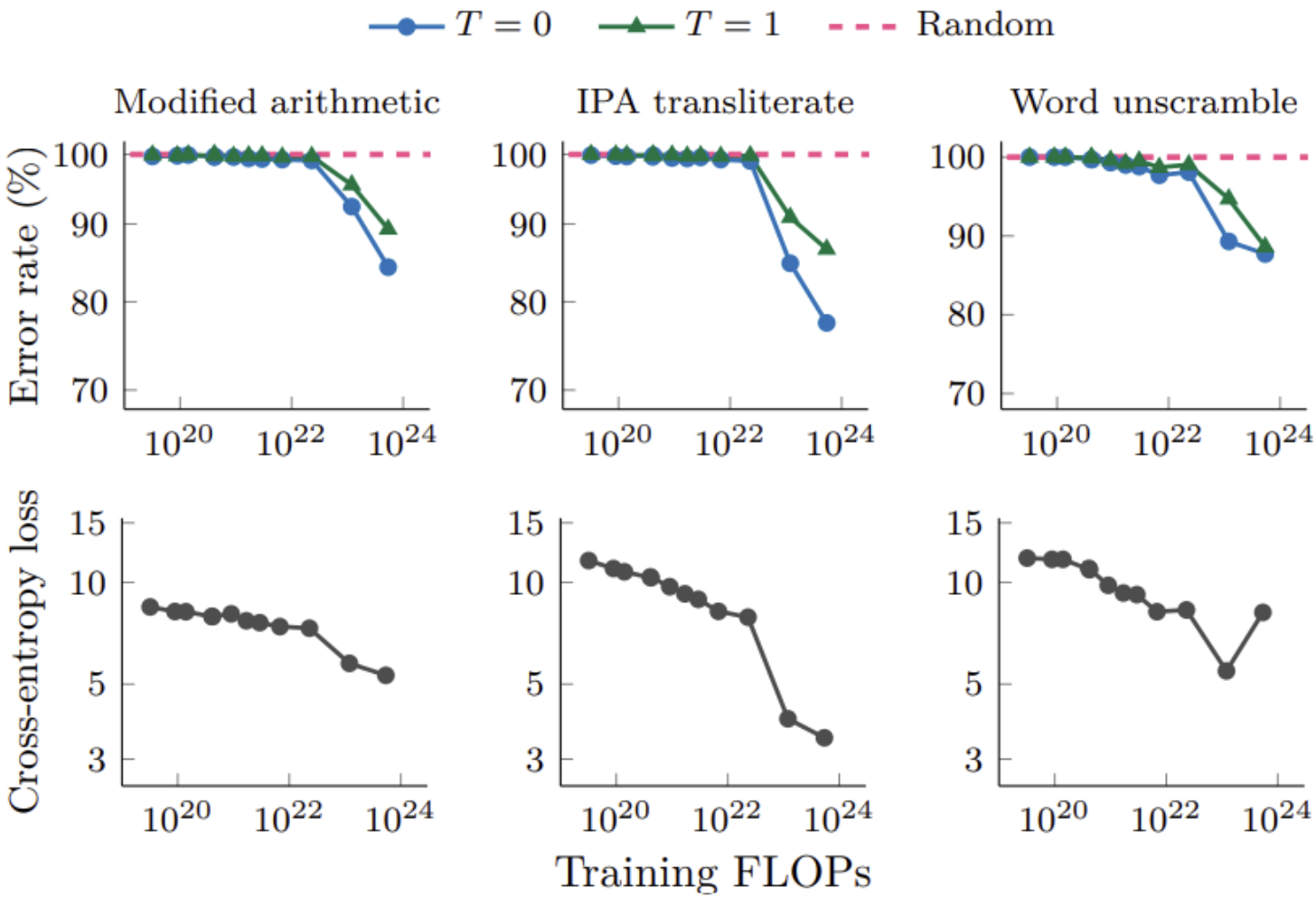


Figure 6: Generating the accuracy graph with more data points

Limtation

Using a continuous metric instead of a discrete metric also causes the trends to disappear (Observed by Wei et al. [9])



So, emergence is just mirage?

Possible mechanism for emergence

- Wei et al. [9] hypothesize, certain tasks may require a model to have layers and parameters beyond a particular threshold (E.g. computation steps, memorization requirements)
- Tasks that are bottlenecked by a particular skill requirement (**Monogenic** [4]) will appear emergent [4], [5], [1], even when using a continuous metric [4].
- Michaud et al. [4] report most LLM tasks to be **Polygenic**[1] in nature.

[1] i.e. dependent on multiple skils who compose additively

Conclusion

- Emergence in LLMs is defined as abilities that appear at scale but unexplainable by scaling laws.
- Scaling Language Models is essential to discover new abilities of models
  - Further study on model architecture would help achieve the emergent abilities at a smaller scale.
- Current evidence for emergence is not enough.
  - The apparent "emergence" can be explained by scaling laws by using a continuous metric.
  - There does exist a strong mathematical model to explain emergence if sighted.
  - Schaeffer, Miranda, and Koyejo [6], while criticizing the findings, also encourage further studies on emergence.

[1] Sanjeev Arora and Anirudh Goyal. "A theory for emergence of complex skills in language models". In: *arXiv preprint arXiv:2307.15936* (2023).

[2] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[3] Aakanksha Chowdhery et al. "Palm: Scaling language modeling with pathways". In: *Journal of Machine Learning Research* 24.240 (2023), pp. 1–113.

[4] Eric Michaud et al. "The quantization model of neural scaling". In: *Advances in Neural Information Processing Systems* 36 (2024).

[5] Maya Okawa et al. "Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task". In: *Advances in Neural Information Processing Systems* 36 (2024).

[6] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. "Are emergent abilities of large language models a mirage?" In: *Advances in Neural Information Processing Systems* 36 (2024).

[7] Aarohi Srivastava et al. "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models". In: *arXiv preprint arXiv:2206.04615* (2022).

[8] Hugo Touvron et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).

[9] Jason Wei et al. "Emergent abilities of large language models". In: *arXiv preprint arXiv:2206.07682* (2022).