Emergent Abilities of Large Language Models

- Wei, Jason, et al.

Presented by - Malyaban Bal

Dept. of Computer Science and Engineering

PennState

High-level overview

- Large Language Models Scaling Factors
  - Number of parameters.
  - Amount of computation (e.g. FLOPs).
  - Training data size.

- Idea of “Emergent Abilities”
  - An ability is emergent if it is not present in smaller models but is present in larger models.
  - Random performance in smaller model => Performance increases beyond a threshold.

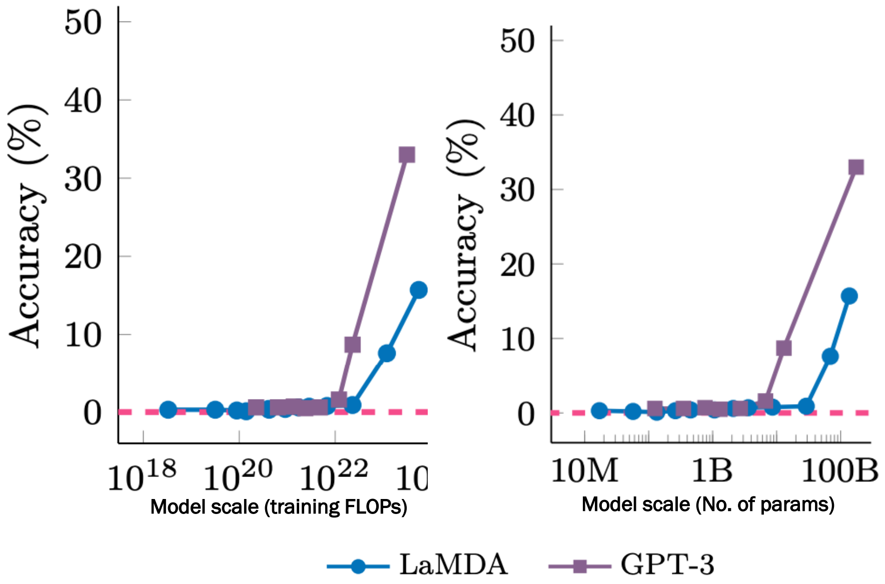
“Emergence is when quantitative changes in a system result in qualitative changes in behavior.” - Philip Anderson

Emergent Abilities

- Typical features
  - Cannot be extrapolated using a scaling law.
  - Concept of a threshold.
  - Function of multiple correlated features.

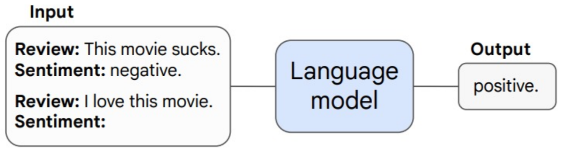
- Is the scale factor absolute?
  - Train data quality, model architecture, etc. can reduce number of params, or compute.
  - Open question: Are our LLMs trained optimally?

Mod. Arithmetic task



Few-Shot Prompted Tasks

- Prompting Technique
  - Perform a specific task by prompting the LM in inference time without any model fine-tuning.
  - Idea of Few-shot prompting (even 0 or 1-shot).



Input

Review: This movie sucks.
Sentiment: negative.

Review: I love this movie.
Sentiment:

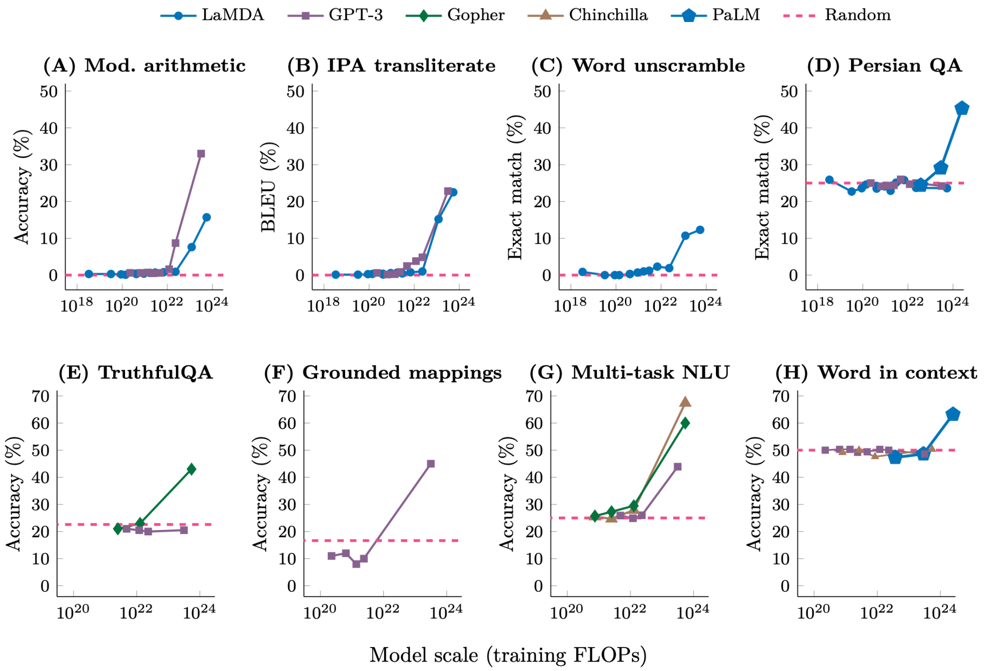
Language model

Output

positive.

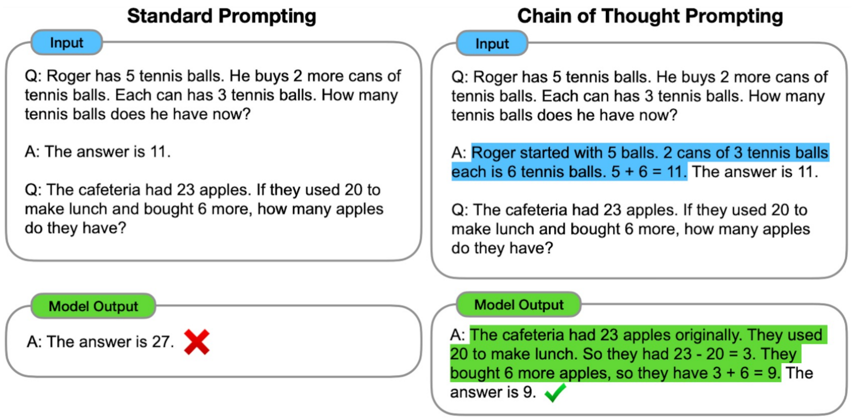
Figure 1: Example of an input and output for few-shot prompting.

- Few-shot prompting as an “Emergent Ability”
  - Few-shot prompting gives random performance until a threshold.
  - Beyond threshold multiple tasks can be solved using prompting. For examples, several tasks in Big-Bench (LM benchmarks), TruthfulQA, Multi-task language understanding, etc.



Augmented Prompting Strategies

- **Multi-step reasoning:** Guiding LMs to produce a sequence of intermediate steps before giving the final answer. For example, chain-of-thought prompting.



Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

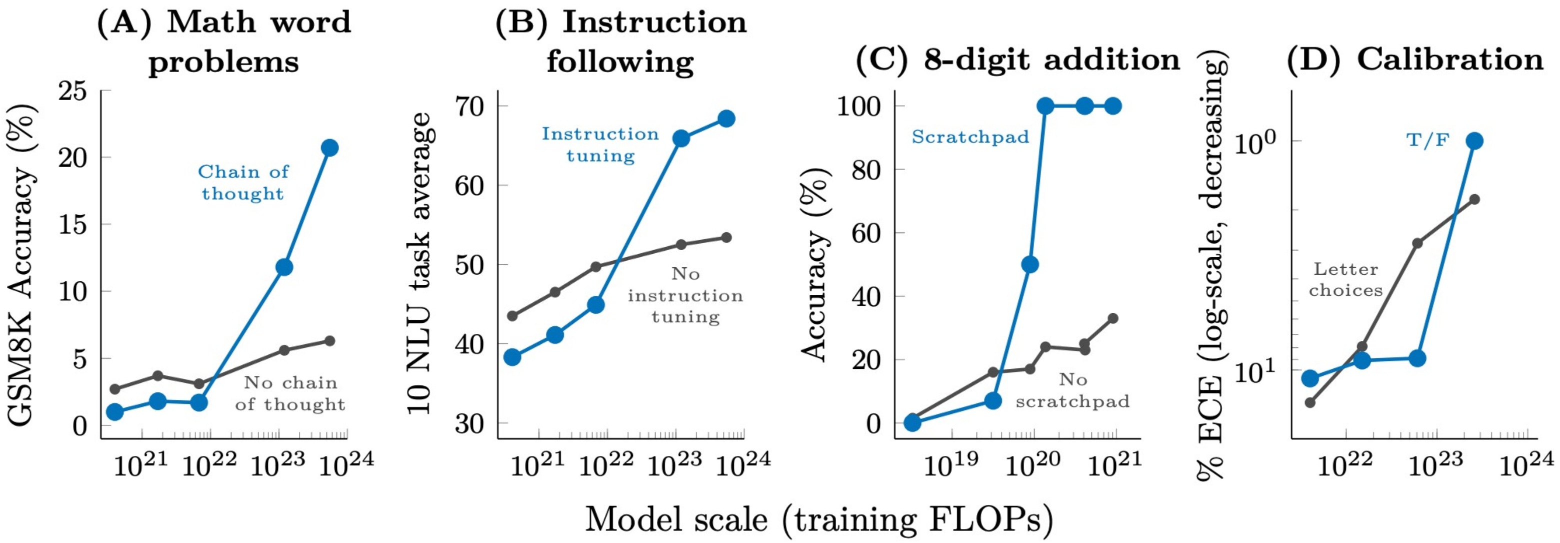
- **Instruction following:** Perform new tasks by reading instructions describing the task (without few-shot exemplars).

- **Program Execution:** Finetune language models to predict intermediate outputs (“scratchpad”) which enables them to successfully execute multi-step computations.

- **Model Calibration:** Calibration measures whether models can predict which questions they will be able to answer correctly.

Augmented Prompting Strategies as Emergent Abilities

- Why Emergent?
  - Technique doesn't improve performance or degrade it for small scale models.
  - Actual advantage can be realized at larger scale.



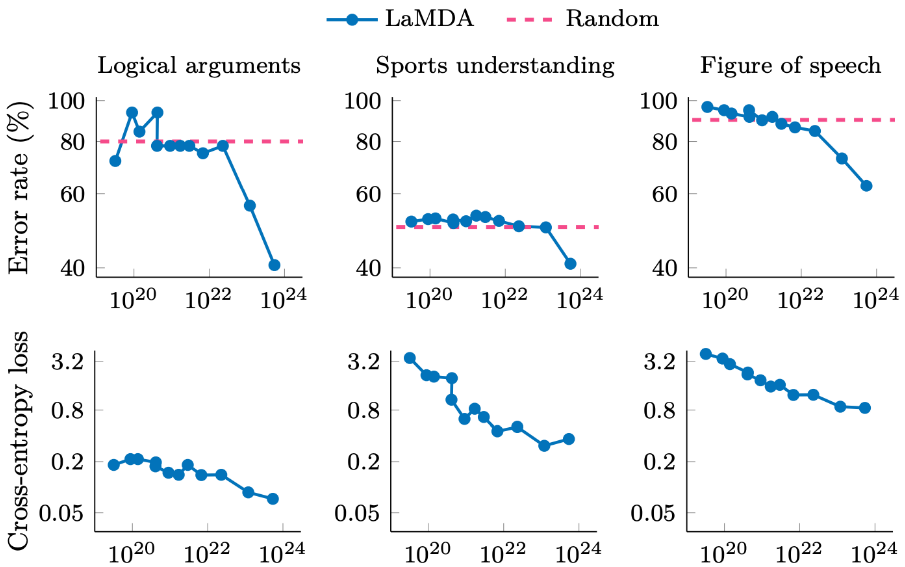
Explanation of “Emergence”

- Intuition
  - Some tasks inherently require deeper networks, eg. Multi-step reasoning.
  - Task requiring world knowledge benefit from more parameters and training data. (“Memorization?”). For example, closed-book question-answering task may require a model with enough parameters to capture the compressed knowledge base.

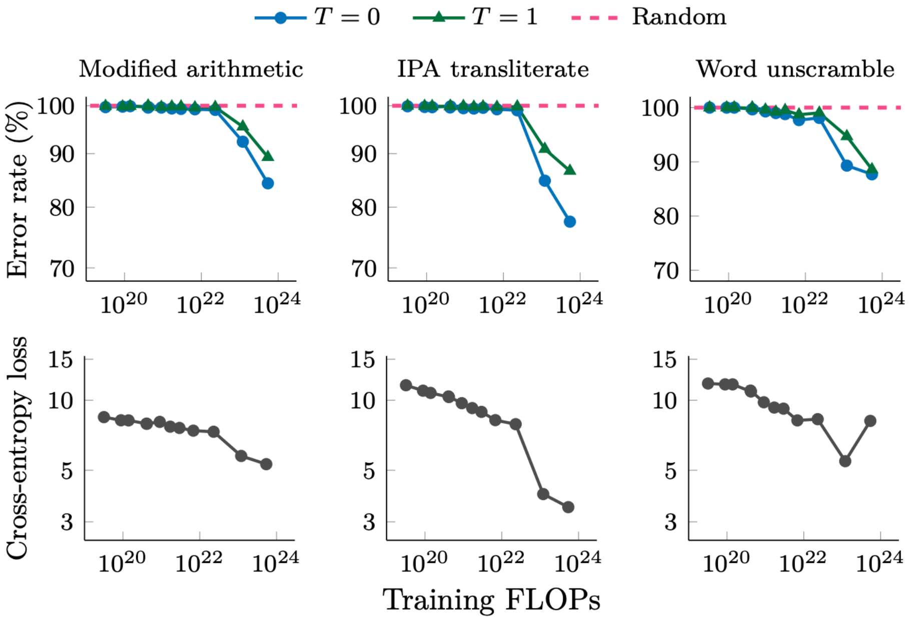
- Evaluation Metrics
  - Choices such as Exact string match doesn’t identify incremental advances.
  - Metric should give partial credits to understand progress.

Cross-Entropy Loss

- Cross-entropy loss improves (decreases) with increase in scale.
- Cross-entropy loss captures the distance between the predicted distribution and ground truth.



Classification Tasks: Metric Used to compute error rate is accuracy.



Generative Tasks: T = 0 is greedy decoding and T = 1 Random Sampling. Metric used for modified arithmetic and word unscramble is Exact match and for IPA transliterate is BLEU.

PennState

Emergence beyond scaling

- Novel Architectures & Improved training
  - PaLM(62B) achieves above-random performance in 14 Big-Bench Tasks (ascii word recognition, metaphor understanding, etc) in which LaMDA(137B) and GPT-3(175B) models perform at near-random.
    - Possible reasons?: Better quality training data and architectural differences

- Novel Pre-training objective.
  - Example: Mixture-of-denoisers objective (Tay et al., 2022a) enabled emergent performance on several BIG-Bench tasks.

- Porting Emergent Abilities to models with smaller scale
  - Various finetuning approaches (Ouyang et al. 2022) can be used to reduce the scale of the model.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. Unifying language learning paradigms. arXiv preprint arXiv:2205.05131, 2022a. URL https://arxiv.org/abs/2205.05131.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155, 2022. URL https://arxiv.org/abs/2203.02155.

PennState

Emergent Risks

- Societal Risks
  - Serious implications such as truthfulness, bias, toxicity, etc.
  - Bias can increase with scaling for ambiguous contexts (BBQ bias benchmark (Parrish et al., 2022)).
  - LLMs can produce more toxic responses from the RealToxicityPrompts dataset (mitigation possible with selective prompting)
  - TruthfulQA benchmark (Lin et al., 2021) showed that GPT-3 models were more likely to mimic human falsehoods as they got larger.
- Future Risks
  - Risks that might be in futures LLMs or haven't been categorized yet. Example: Backdoor vulnerabilities, inadvertent deception, or harmful content synthesis.
    - Data filtering, forecasting, auto-discovery of harmful behaviors, etc. are approaches proposed for discovering and mitigating emergent risks.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In Findings of ACL, 2022. URL https://arxiv.org/abs/2110.08193.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958, 2021. URL https://arxiv.org/abs/2109.07958.

Sociological Impact

- Shift from task specific model research to general models that can perform multiple tasks.
    - Few-shot Prompting techniques (on ChatGPT) even bested sota methods for various tasks such as TriviaQA & PiQA question-answering benchmarks, etc.

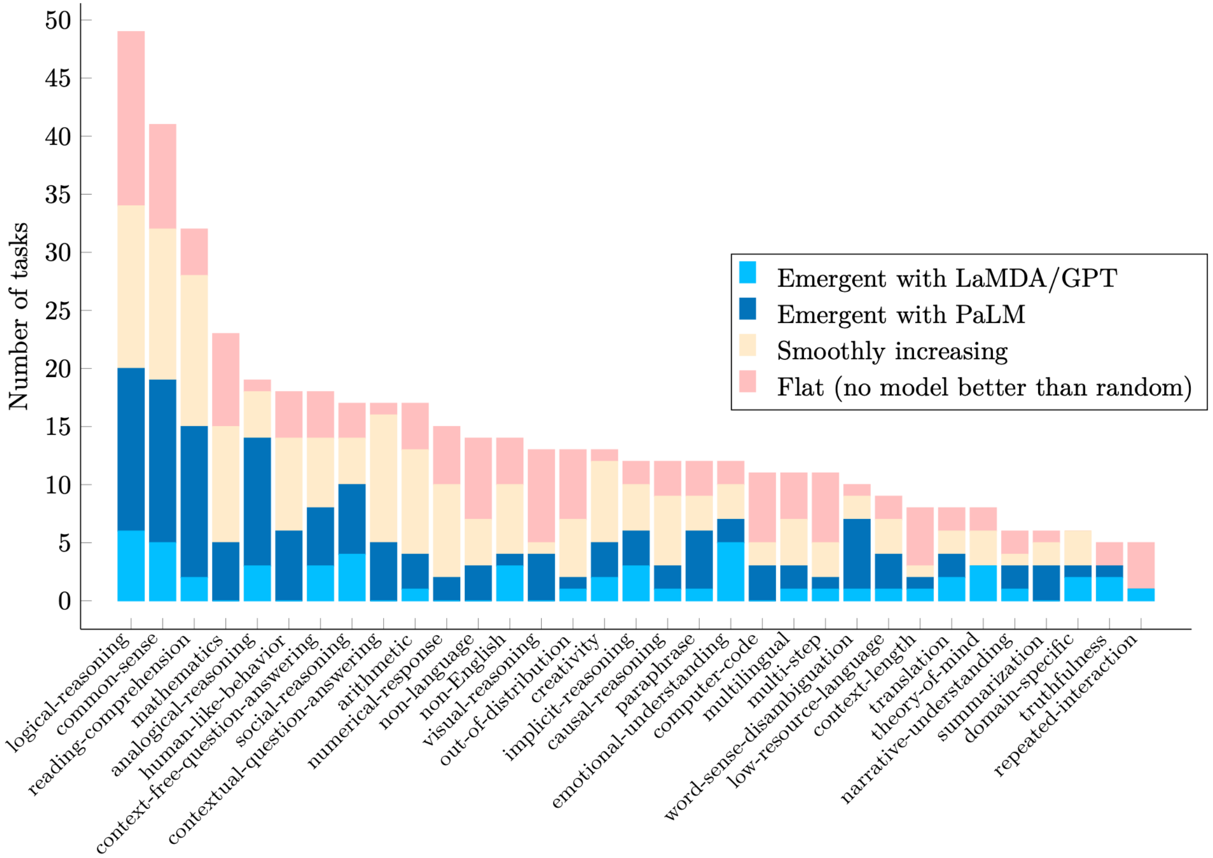
- Applications of LLMs outside NLP community
    - LLM used via prompting to translate natural language instructions into actions executable by, interact with users , etc.
    - LLMs deployed in real-world products, such as GitHub CoPilot, ChatGPT, etc. Search engines like BING have started using it!

Future Research

- Model Scaling
  - Scaling is expensive and constrained by hardware limitations.
- Improved model architectures and training
  - Developing better models on large high quality training data can be a promising path forward. Example: sparse mixture-of-experts architectures (Lepikhin et al., 2021), which scale up the number of parameters in a model while maintaining constant computational costs for an input.
- Improving on prompting techniques
  - Techniques like those discussed previously to improve few-shot prompting.
- Unsolved tasks
  - There are multiple benchmark tasks on which even LLMs like GPT-3 has near to random performance. Abilities such as abstractive reasoning is yet to found. Multilingual Emergence is another interesting domain.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. ICLR, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Abilities to Explore



Proportion of emergent tasks for keywords in BIG-Bench (each task can be associated with multiple keywords). Smoothly increasing performance improved predictably as model scale increased. Emergent with LaMDA/GPT: performance was near-random until used with LaMDA 137B or GPT-3 175B. Emergent with PaLM: performance was near-random for all previous models, until using a PaLM model (8B, 62B, or 540B). Flat: no model performs better than random.

Conclusion

- Emergent abilities are recently discovered outcome of scaling up language models, enabling rich and diverse applications of LLMs.

- Open questions
    - How exactly such abilities emerge?
    - Will further scaling enable even more emergent abilities not previously known?
    - How can we achieve optimal LLM training?

PennState

Questions?