ETH Zürich

Advanced Topics in Machine Learning and Data Science

Emergent Abilities of Large Language Models

Wei, Jason, et al. (2022)

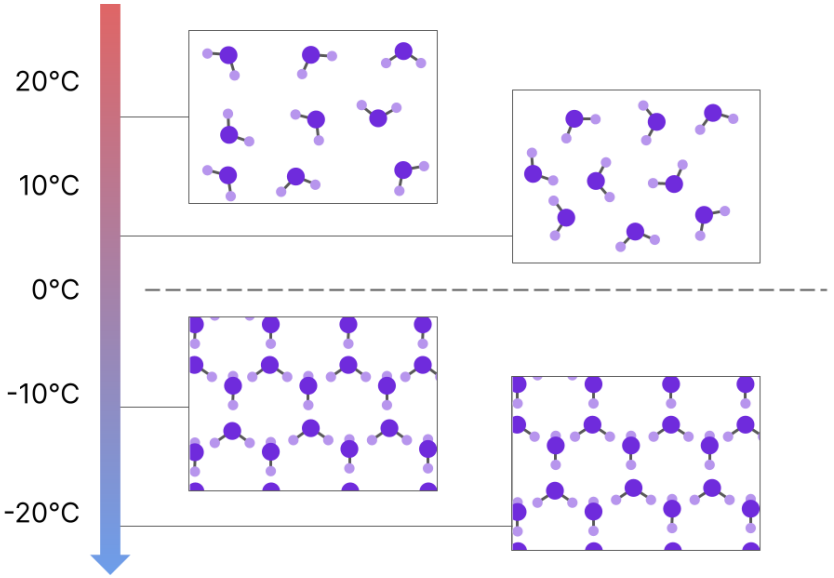
Presented by
Cyrill Püntener

Zürich, 10. April 2024

What are Emergent Abilities?

Emergent Abilities (Definition)

> Emergence is when quantitative changes in a system result in qualitative changes in behavior.
>
> (Wei, et al., 2022)



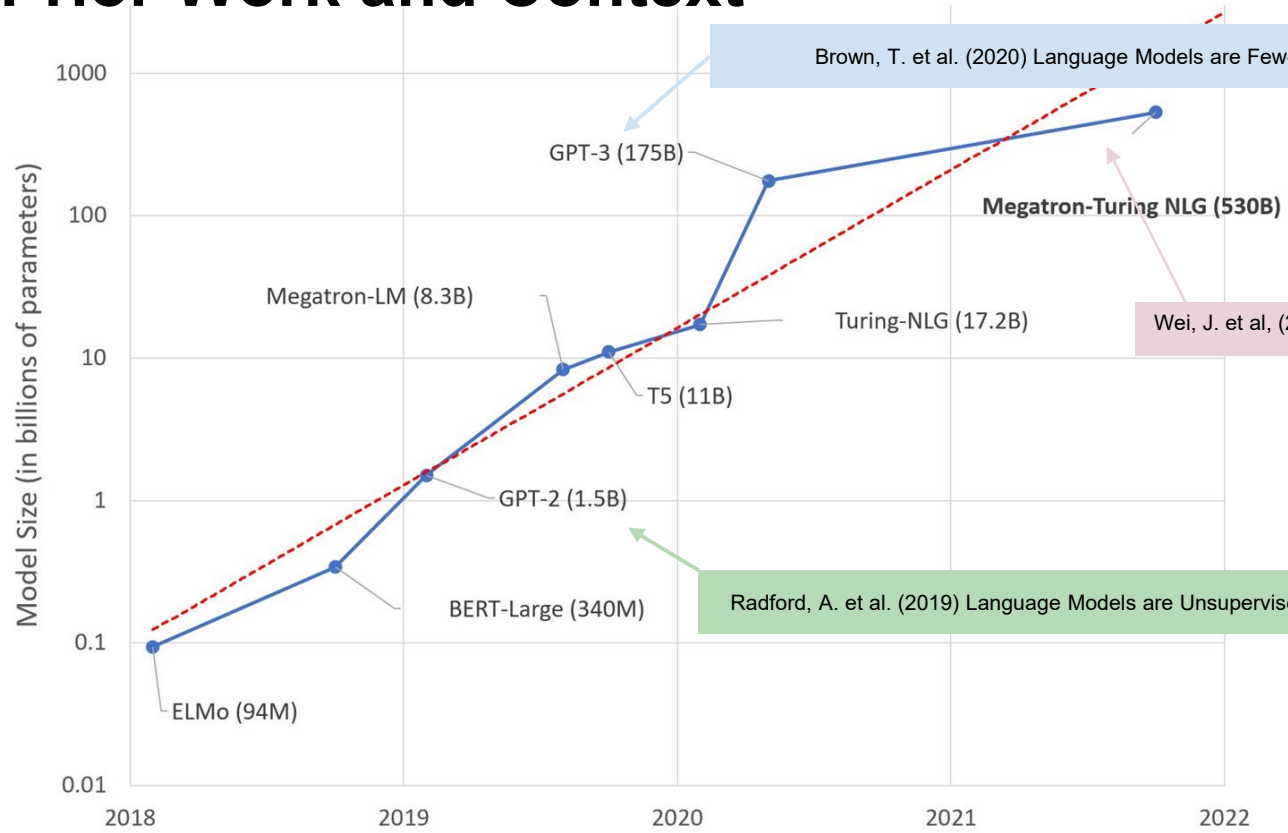
Source Image: Emergent Abilities of Large Language Models (2023) News, Tutorials, AI Research. Available at: https://www.assemblyai.com/blog/emergent-abilities-of-large-language-models/ (Accessed: 2 April 2024).

Emergent Abilities (Definition)

> An ability is emergent if it is not present in smaller models but is present in larger models.
>
> (Wei, et al., 2022)

> Emergent abilities cannot be predicted simply by extrapolating the performance of smaller models.
>
> (Wei, et al., 2022)

Prior Work and Context



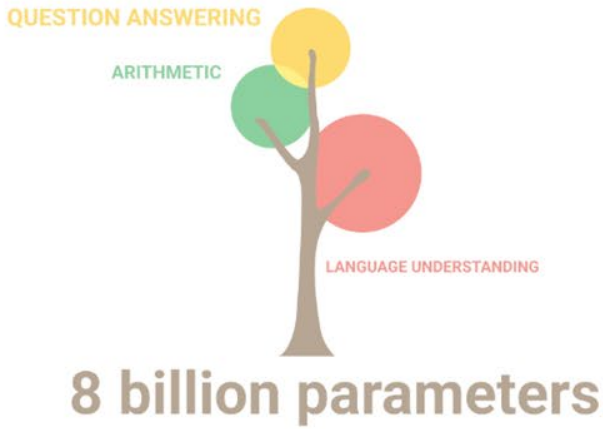
Brown, T. et al. (2020) Language Models are Few-Shot Learners.

Wei, J. et al, (2022) Emergent Abilities of Large Language Models

Radford, A. et al. (2019) Language Models are Unsupervised Multitask Learners.

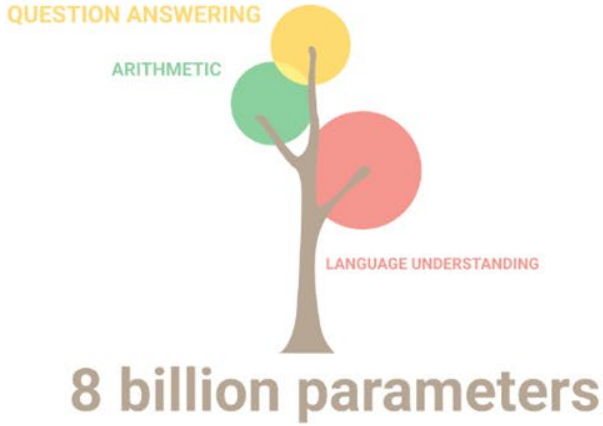
Large Language Models: A New Moore's Law? (no date). Available at: https://huggingface.co/blog/large-language-models (Accessed: 8 April 2024).

Emergent Abilities (Examples from PaLM)



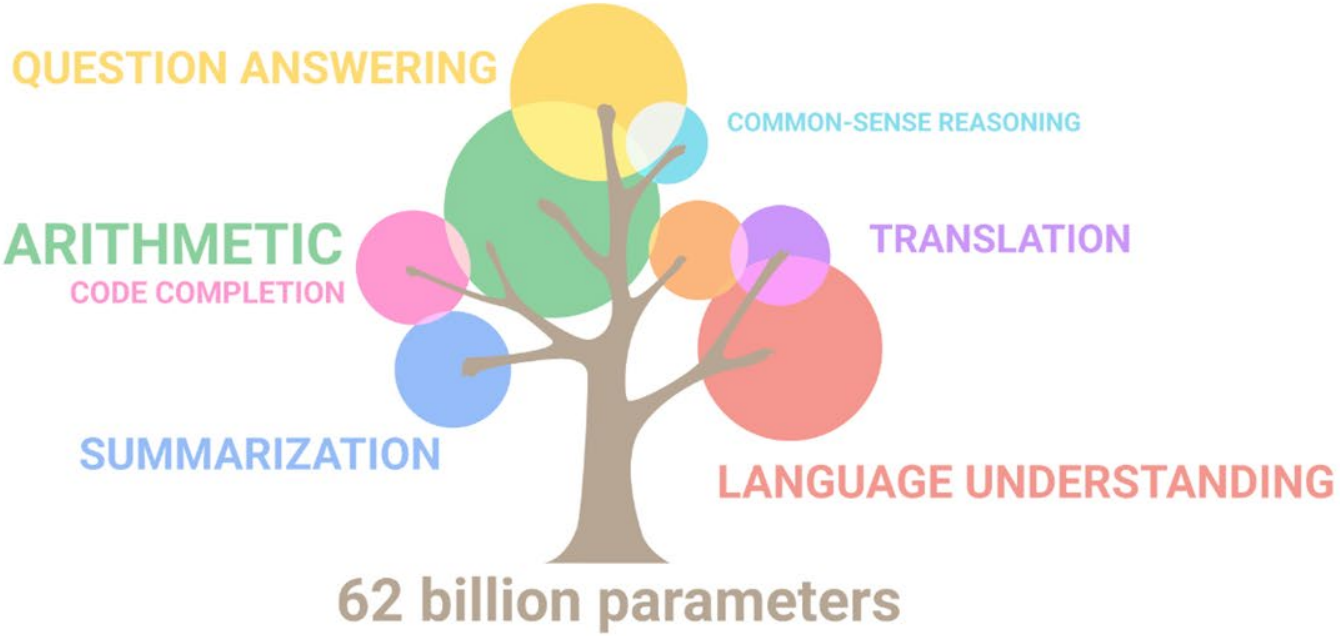
Source Animation: Pathways language model (PaLM): scaling to 540 billion parameters for breakthrough performance (2022). Available at: https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html?ref=assemblyai.com (Accessed: 2 April 2024).

Emergent Abilities (Examples from PaLM)



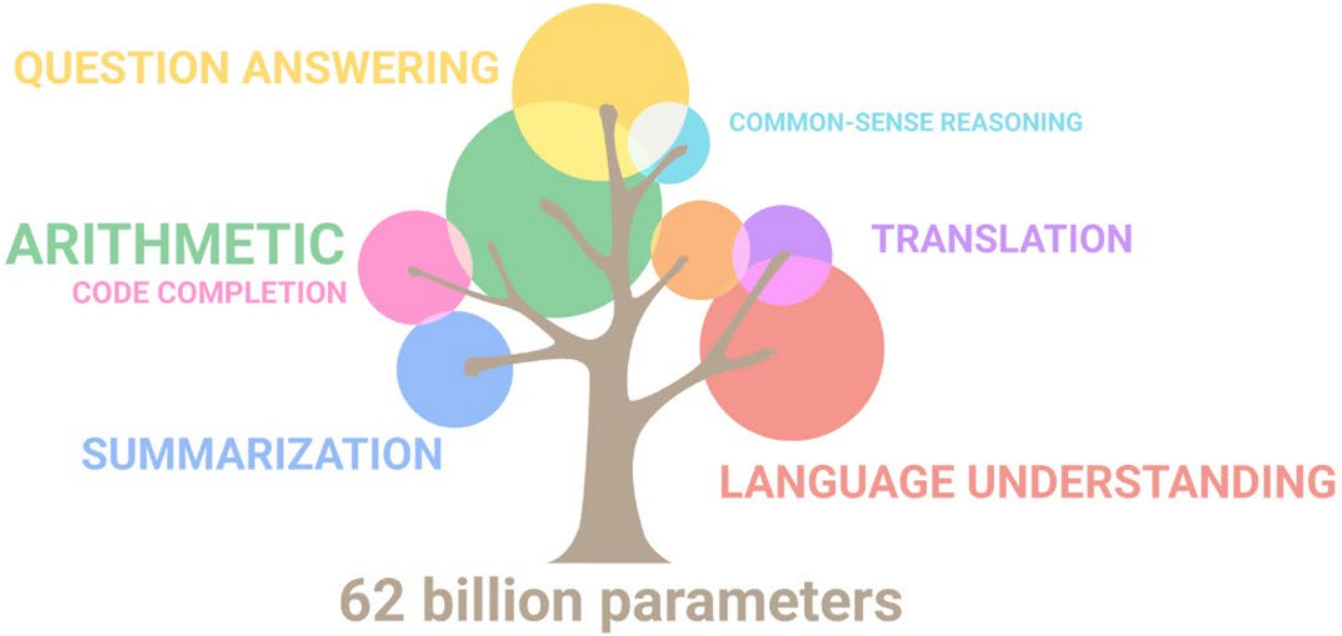
Source Animation: Pathways language model (PaLM): scaling to 540 billion parameters for breakthrough performance (2022). Available at: https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html?ref=assemblyai.com (Accessed: 2 April 2024).

Emergent Abilities (Examples from PaLM)



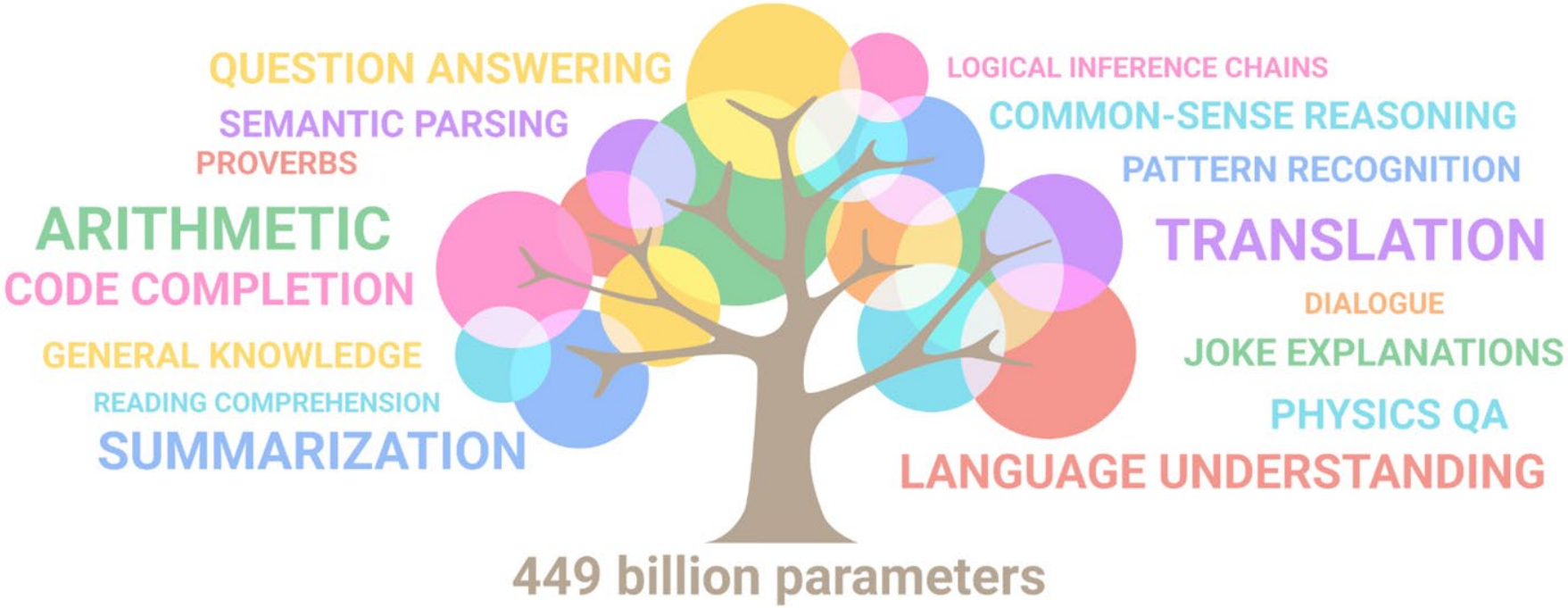
Source Animation: Pathways language model (PaLM): scaling to 540 billion parameters for breakthrough performance (2022). Available at: https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html?ref=assemblyai.com (Accessed: 2 April 2024).

Emergent Abilities (Examples from PaLM)



Source Animation: Pathways language model (PaLM): scaling to 540 billion parameters for breakthrough performance (2022). Available at: https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html?ref=assemblyai.com (Accessed: 2 April 2024).

Emergent Abilities (Examples from PaLM)



Source Animation: Pathways language model (PaLM): scaling to 540 billion parameters for breakthrough performance (2022). Available at: https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html?ref=assemblyai.com (Accessed: 2 April 2024).

How to measure the performance of Large Language Models?

BIG-Bench

Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (Srivastava, et al. 2023)

What is BIG-Bench?

- comprehensive benchmark for language models
- composed by 400 researchers across 132 research institutes
- tries to solve the cap between benchmarks and LLM performance
- contains 204 diverse tasks across multiple domains

BIG-Bench Emoji Movie Test

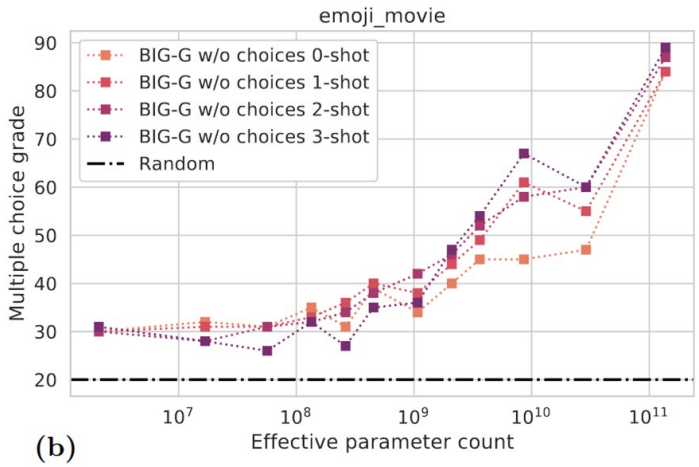
Q: What movie does this emoji describe? 👧🐟🐠🐡

Examples: Srivastava, A. et al. (2023) 'Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models'. arXiv. Available at: http://arxiv.org/abs/2206.04615.

BIG-Bench Emoji Movie Test

Q: What movie does this emoji describe? 👧🐟🐠🐡

2m: i'm a fan of the same name, but i'm not sure if it's a good idea
16m: the movie is a movie about a man who is a man who is a man …
53m: the emoji movie 🐟🐠🐡
125m: it's a movie about a girl who is a little girl
244m: the emoji movie
422m: the emoji movie
1b: the emoji movie
2b: the emoji movie
4b: the emoji for a baby with a fish in its mouth
8b: the emoji movie
27b: the emoji is a fish
128b: finding nemo



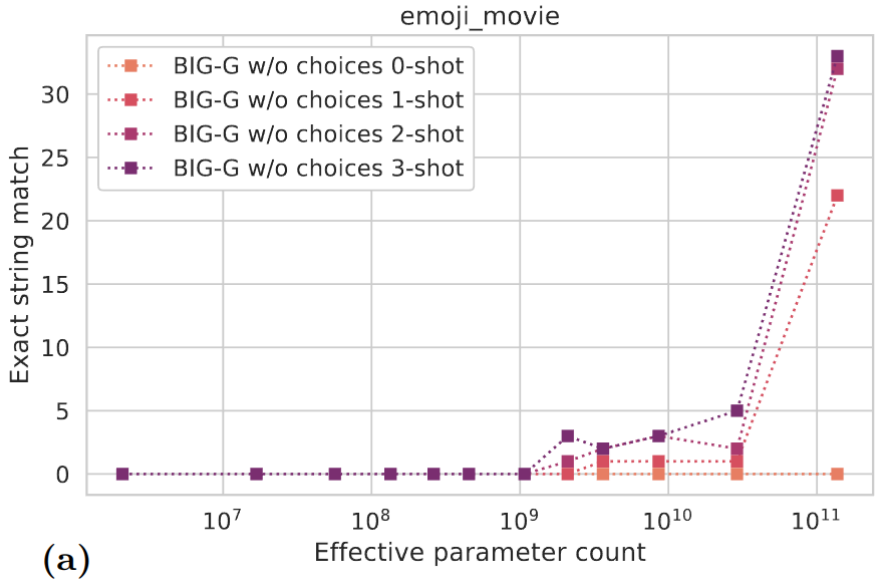


Figure: Srivastava, A. et al. (2023) 'Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models'. arXiv. Available at: http://arxiv.org/abs/2206.04615.

Abilities at different Model Scales

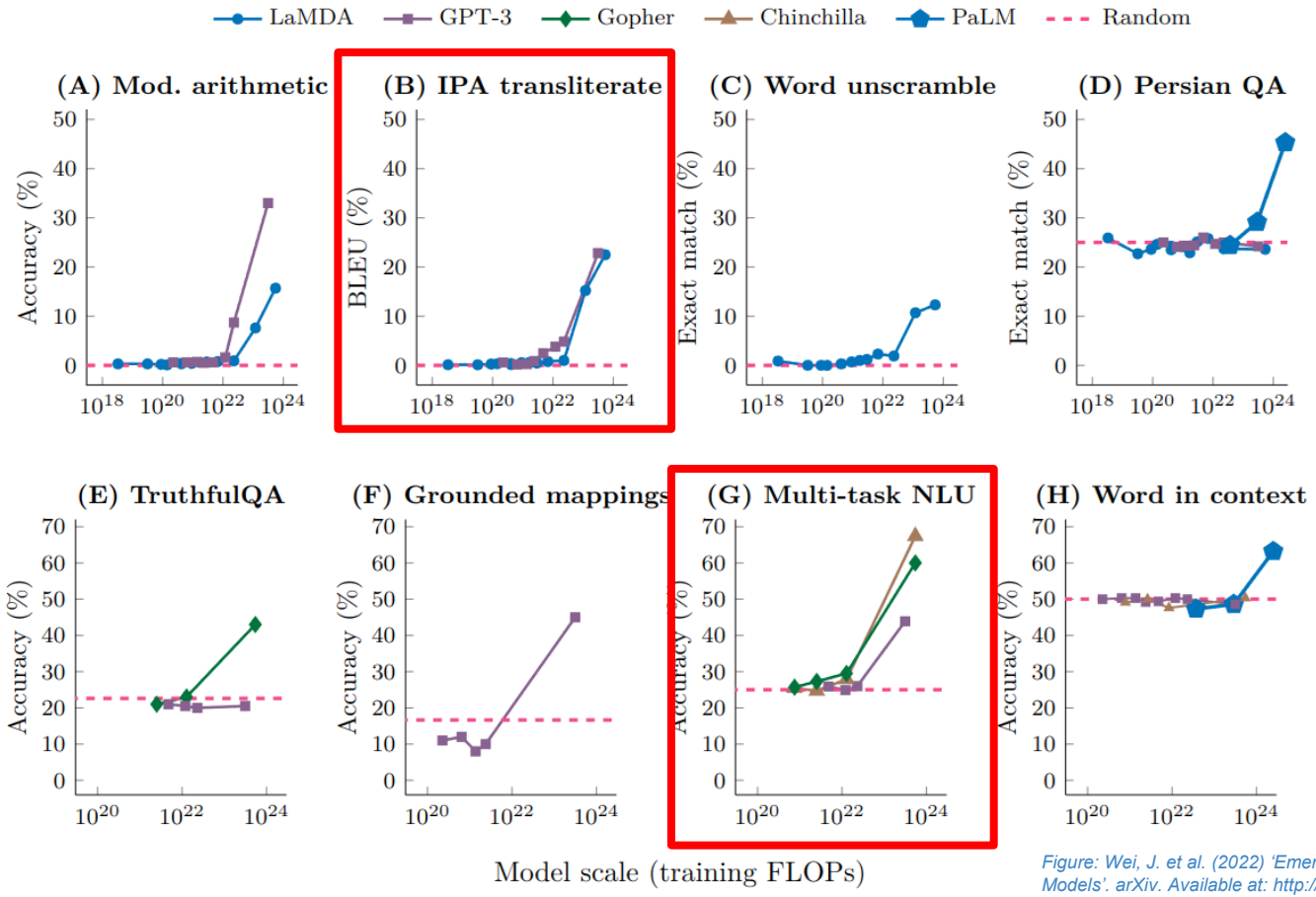


Figure: Wei, J. et al. (2022) 'Emergent Abilities of Large Language Models'. arXiv. Available at: http://arxiv.org/abs/2206.07682 (Accessed: 22 February 2024).

Measuring Massive Multitask Language Understanding

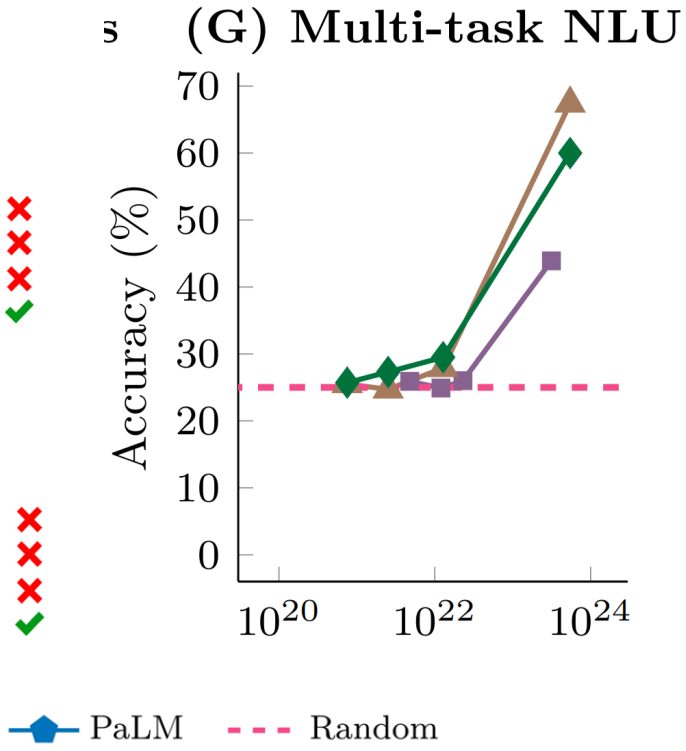
Microeconomics

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained.
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency.
(C) monopoly firms do not engage in significant research and development.
(D) consumer surplus is lost with higher prices and lower levels of output.

Figure 3: Examples from the Microeconomics task.

College Mathematics

In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a
(A) pair of points
(B) circle
(C) half-line
(D) line



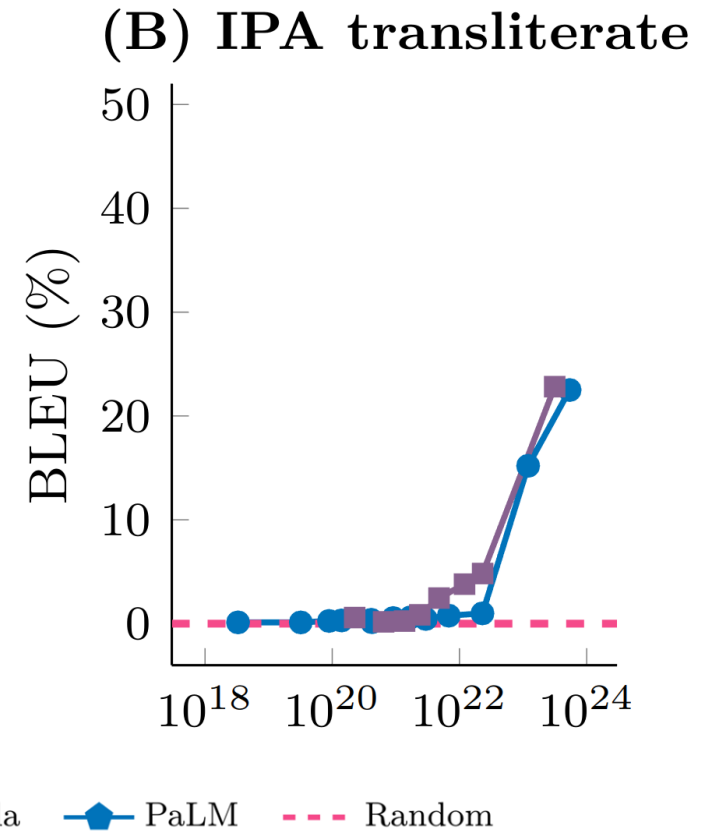
Questions (left): Hendrycks, D. et al. (2021) 'Measuring Massive Multitask Language Understanding'. arXiv. Available at: http://arxiv.org/abs/2009.03300 (Accessed: 8 April 2024).

Plot (right): Wei, J. et al. (2022) 'Emergent Abilities of Large Language Models'. arXiv. Available at: http://arxiv.org/abs/2206.07682 (Accessed: 22 February 2024).

International Phonetic Alphabet Transliteration

Input (English): The 1931 Malay census was an alarm bell.

Target (IPA): ðə 1931 ˈmeɪleɪ ˈsɛnsəs wɑz ən əˈlɑrm bɛl.



Example: Srivastava, A. et al. (2023) 'Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models'. arXiv. Available at: http://arxiv.org/abs/2206.04615 (Accessed: 21 March 2024).

Plot (right): Wei, J. et al. (2022) 'Emergent Abilities of Large Language Models'. arXiv. Available at: http://arxiv.org/abs/2206.07682 (Accessed: 22 February 2024).

ETH Zürich

Emergent Abilities

	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Few-shot prompting abilities				
• Addition/subtraction (3 digit)	2.3E+22	13B	GPT-3	Brown et al. (2020)
• Addition/subtraction (4-5 digit)	3.1E+23	175B		
• MMLU Benchmark (57 topic avg.)	3.1E+23	175B	GPT-3	Hendrycks et al. (2021a)
• Toxicity classification (CivilComments)	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Truthfulness (Truthful QA)	5.0E+23	280B		
• MMLU Benchmark (26 topics)	5.0E+23	280B		
• Grounded conceptual mappings	3.1E+23	175B	GPT-3	Patel & Pavlick (2022)
• MMLU Benchmark (30 topics)	5.0E+23	70B	Chinchilla	Hoffmann et al. (2022)
• Word in Context (WiC) benchmark	2.5E+24	540B	PaLM	Chowdhery et al. (2022)
• Many BIG-Bench tasks (see Appendix E)	Many	Many	Many	BIG-Bench (2022)
Augmented prompting abilities				
• Instruction following (finetuning)	1.3E+23	68B	FLAN	Wei et al. (2022a)
• Scratchpad: 8-digit addition (finetuning)	8.9E+19	40M	LaMDA	Nye et al. (2021)
• Using open-book knowledge for fact checking	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Chain-of-thought: Math word problems	1.3E+23	68B	LaMDA	Wei et al. (2022b)
• Chain-of-thought: StrategyQA	2.9E+23	62B	PaLM	Chowdhery et al. (2022)
• Differentiable search index	3.3E+22	11B	T5	Tay et al. (2022b)
• Self-consistency decoding	1.3E+23	68B	LaMDA	Wang et al. (2022b)
• Leveraging explanations in prompting	5.0E+23	280B	Gopher	Lampinen et al. (2022)
• Least-to-most prompting	3.1E+23	175B	GPT-3	Zhou et al. (2022)
• Zero-shot chain-of-thought reasoning	3.1E+23	175B	GPT-3	Kojima et al. (2022)
• Calibration via P(True)	2.6E+23	52B	Anthropic	Kadavath et al. (2022)
• Multilingual chain-of-thought reasoning	2.9E+23	62B	PaLM	Shi et al. (2022)
• Ask me anything prompting	1.4E+22	6B	EleutherAI	Arora et al. (2022)

Table: Wei, J. et al. (2022) 'Emergent Abilities of Large Language Models'. arXiv. Available at: http://arxiv.org/abs/2206.07682.

The Effect of Promoting Strategies?

Prompting Strategies

Zero-shot (Direct prompting):
No previous data or examples (no shots) are given before completing the request.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot (Prompt by one example):
One piece of data or example (one shot) is given before completing the request.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot (Prompt by multiple examples):
Multiple examples or data (few shots) are given before completing the request.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée
4  plush girafe => girafe peluche
5  cheese =>                           ← prompt
```

Figures: Brown, T. et al. (2020) 'Language Models are Few-Shot Learners', in Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 1877–1901.

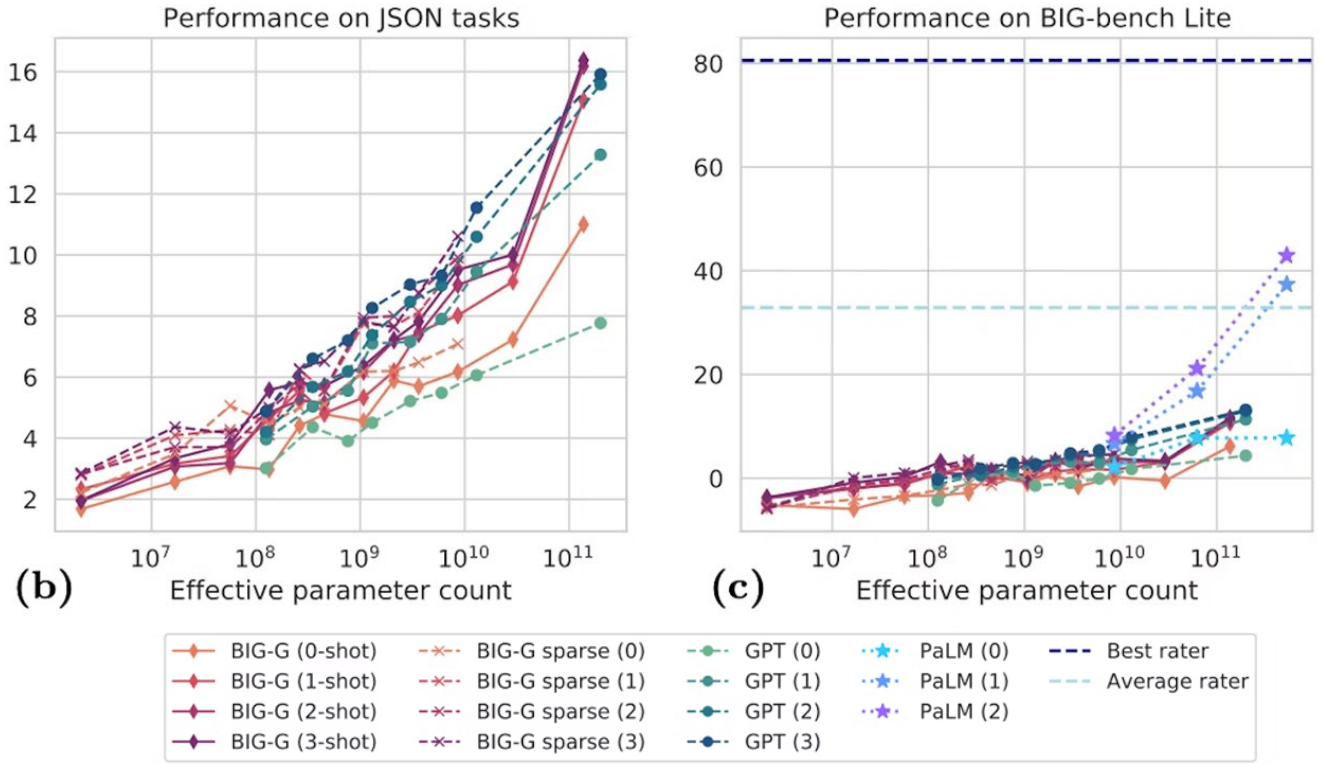
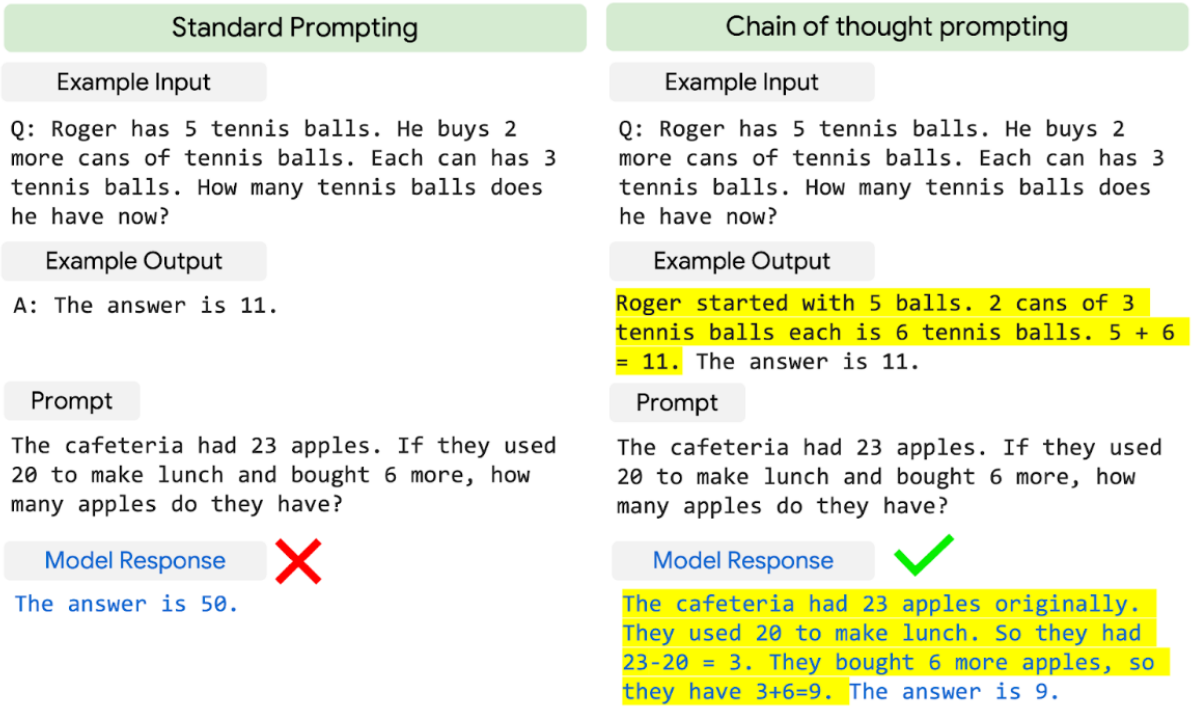


Figure: Srivastava, A. et al. (2023) 'Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models'. arXiv. Available at: http://arxiv.org/abs/2206.04615.

Chain-of-Thought Prompting

➔ Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei, J. et al. (2022))



Source: Pathways language model (PaLM): scaling to 540 billion parameters for breakthrough performance (2022). Available at: https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to-html?ref=assemblyai.com (Accessed: 2 April 2024).

Specialized Promoting and Fine-Tuning

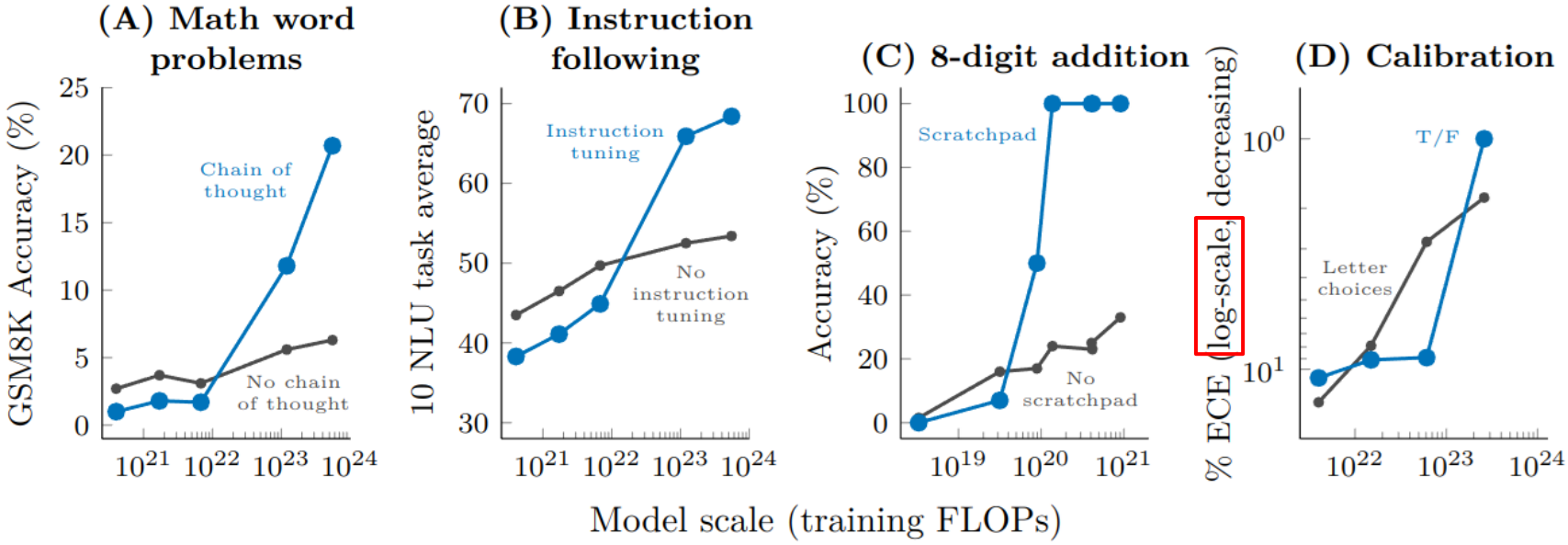


Figure: Wei, J. et al. (2022) 'Emergent Abilities of Large Language Models'. arXiv. Available at: http://arxiv.org/abs/2206.07682.

Prompting Strategies

- Multi-Step Reasoning
- Instruction Following
- Program Execution
- Model Calibration

Emergent Techniques
If a technique shows no improvement or is harmful when compared to the baseline of not using the technique until applied to a model of a large-enough scale, we also consider the **technique an emergent ability.**

Why Emergent Abilities exists?

ETH Zürich

Potential Explanation

Multi-Step Reasoning Tasks
A multi-step reasoning task with ℓ sequential steps,
may require a model with a depth of at least $O(\ell)$ layers

Closed-Book Question-Answering
Good performance on closed-book question-answering may require
a model with enough parameters to capture the knowledge base itself

Shift of Usage: Sociological
Increasing the scale has shifted *how* the community views and uses
language models. For example: from task specific NLP to *general purpose* LLMs

Potential Explanation

Explosion in Research

Scaling has led to an explosion in research on and development of models that are *general purpose* in that they are single models that aim to perform a range of tasks not explicitly encoded in the training data

The explanations given are quite vague:
- emergent abilities may disappear using different metrics
- emergent abilities may be caused by the objective used to measure them

→ *topics of the second presentation today*

Beyond Scaling

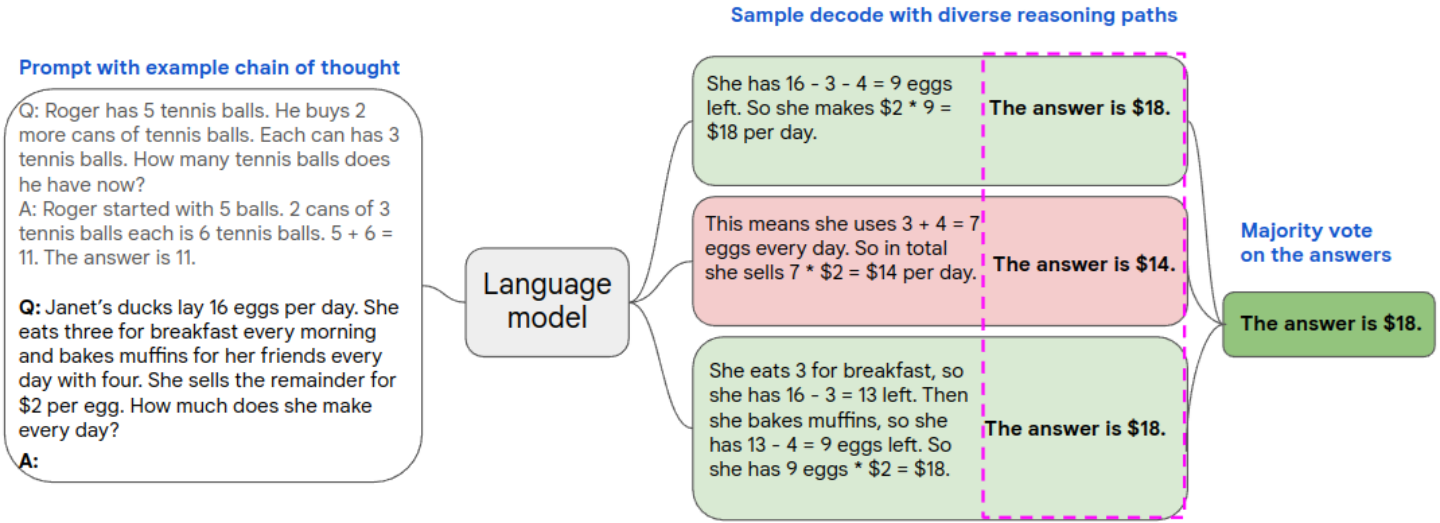
Beyond Scaling - Scaling is Limited

- Hardware Constraints
- Cost Constraints
- Quality / Quantity of Training Data
- Architecture and Training Strategies

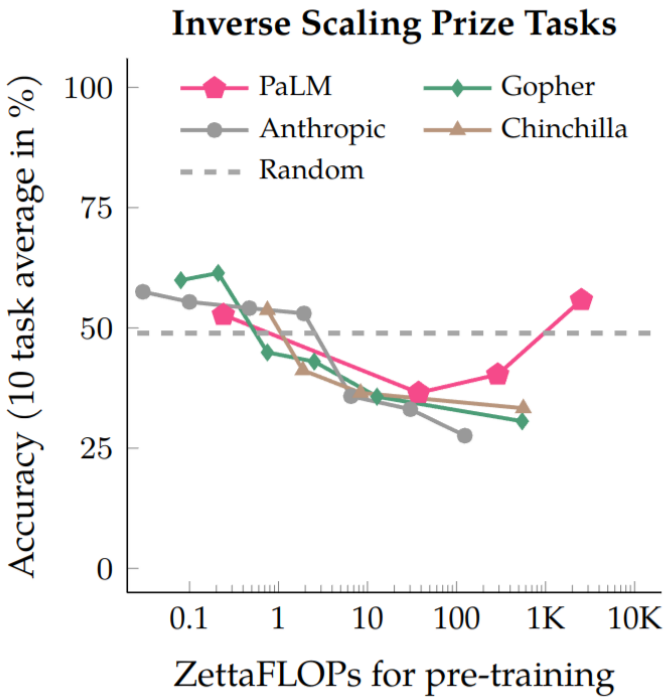
Beyond Scaling - Alternative Prompting Strategies

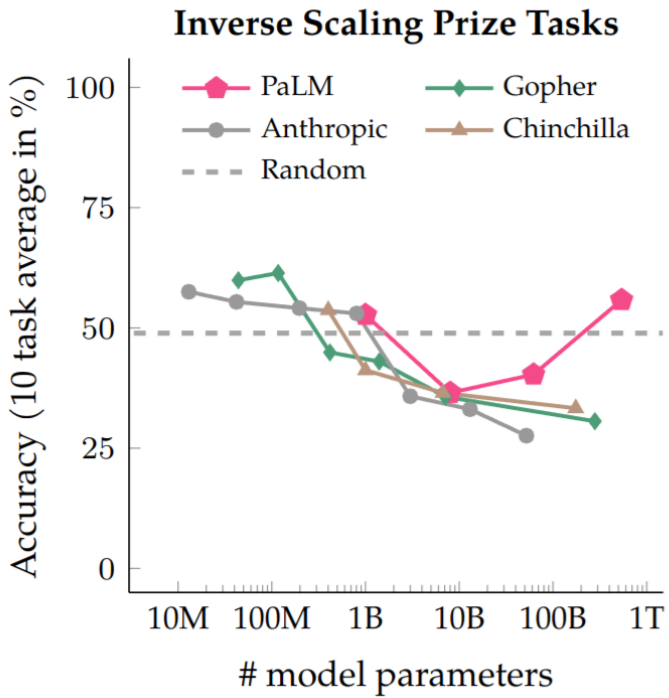
Self-consistency improves chain-of-thought reasoning in language models (ICLR '23).
X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, & D. Zhou.

Self-consistency: majority vote



Scaling not Always Improve: Inverse Scaling





Wei, J. et al. (2023) 'Inverse scaling can become U-shaped'. arXiv. Available at: https://doi.org/10.48550/arXiv.2211.02011.

ETH Zürich

Emergent Risks

- Truthfulness
- Bias of in the model output
- Toxicity

Emergent risks are hard to predict
Emergent risks also include phenomena that might only exist in future language models or that have not yet been characterized in current language models

Conclusion

Future Work

- What controls which abilities will emerge?
- What controls when abilities emerge?
- Who can we improve / augment the training of LLMs?
- How can we make desirable abilities emerge faster?
- How can we ensure undesirable abilities never emerge?

follow-up and related papers

- ➔ TruthfulQA: Measuring How Models Mimic Human Falsehoods (Lin, S., et al. (2021))
- ➔ Predicting Emergent Abilities with Infinite Resolution Evaluation (Hu, S., et al. (2023))
- ➔ Are Emergent Abilities of Large Language Models a Mirage? (Schaeffer, R., et al. (2023))
- ➔ Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models (Srivastava, A. et al. (2023))
- ➔ Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei, J. et al. (2022))

Conclusion

Strengths

- comprehensive overview of the state of the art
- concrete examples for future work
- easy to follow structure, they reintroduce core concepts

Limitations

- limited effort to explain the phenomenon of emergence
- partially unclear which experiments are carried out and which results are from prior papers

thank you

Any Questions?