PAUL G. ALLEN SCHOOL

UNIVERSITY of WASHINGTON

Natural Language Processing

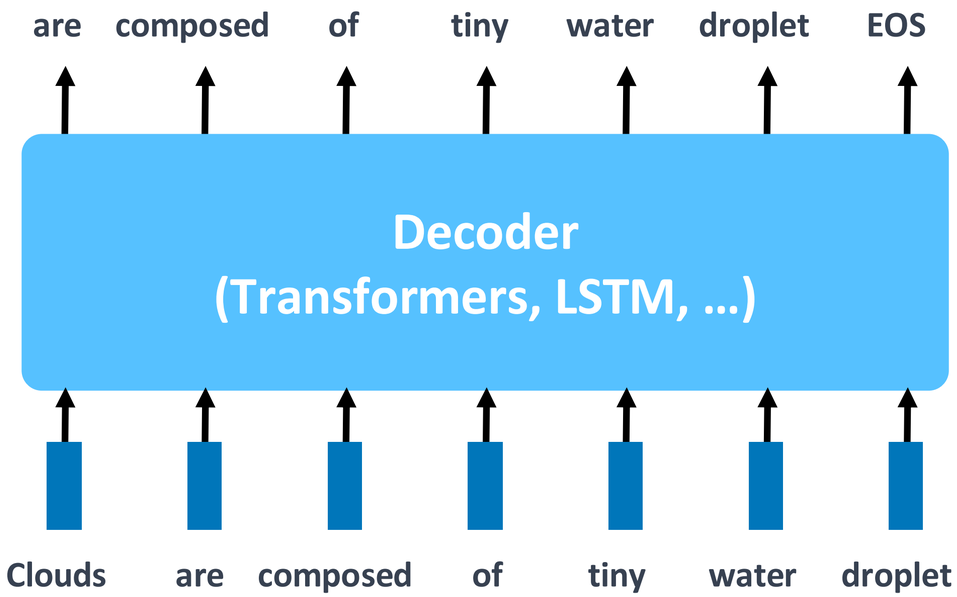
Lecture: In-context Learning, Scaling Laws, Emergent Capabilities

Lecturer: Luke Zettlemoyer

Slides by: Yejin Choi, Taylor Sorensen

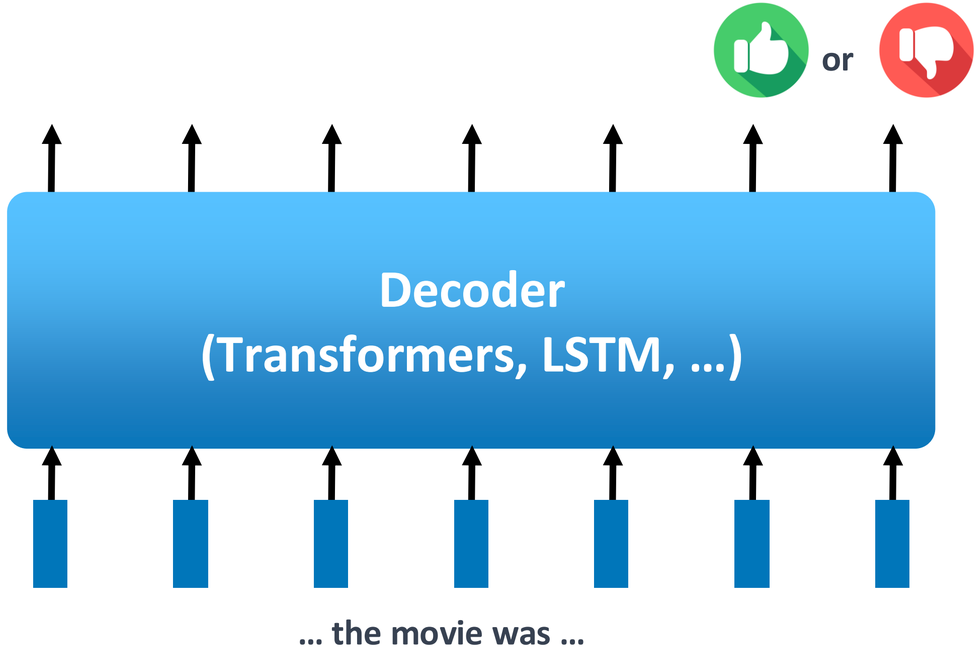
Review: Pretraining and Finetuning

Step 1:
Pre-training



Abundant data; learn general language

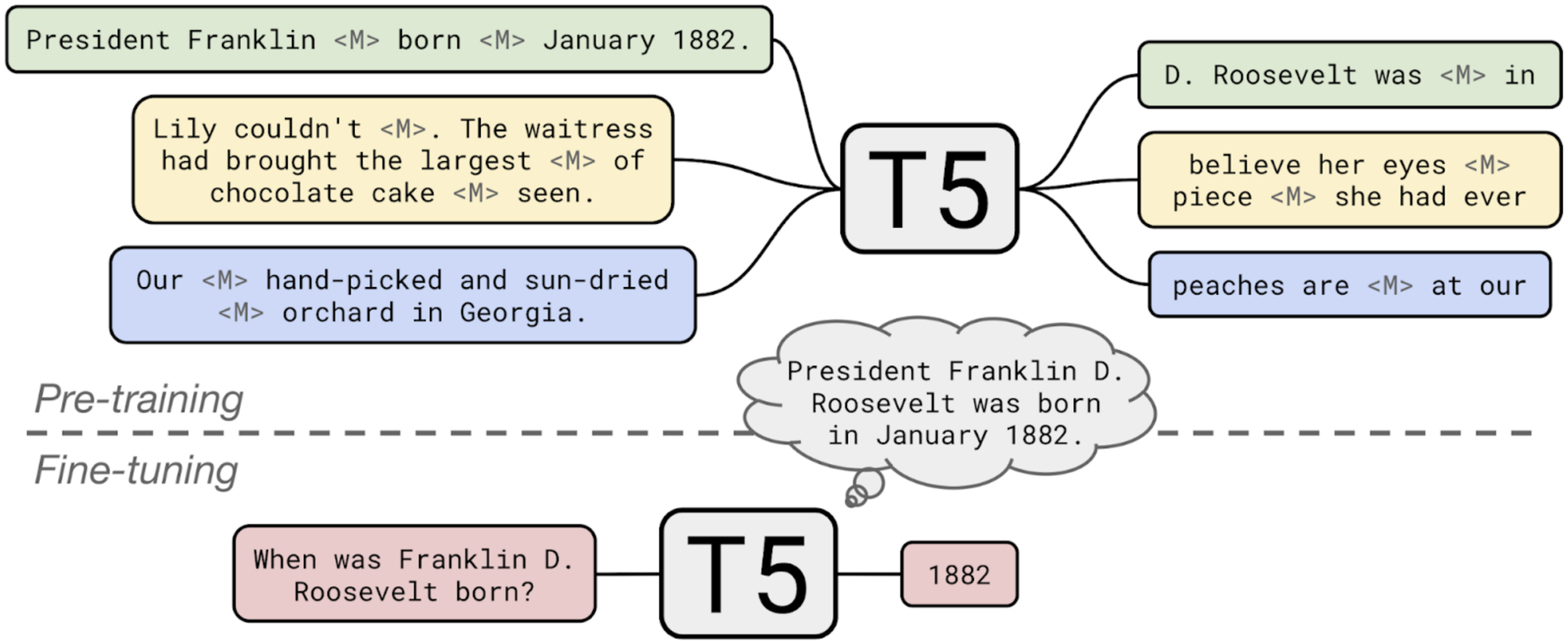
Step 2:
Fine-tuning



Limited data; adapt to the task

Pretrain/Finetune Paradigm

Example: T5 (Raffel et. al, 2019)



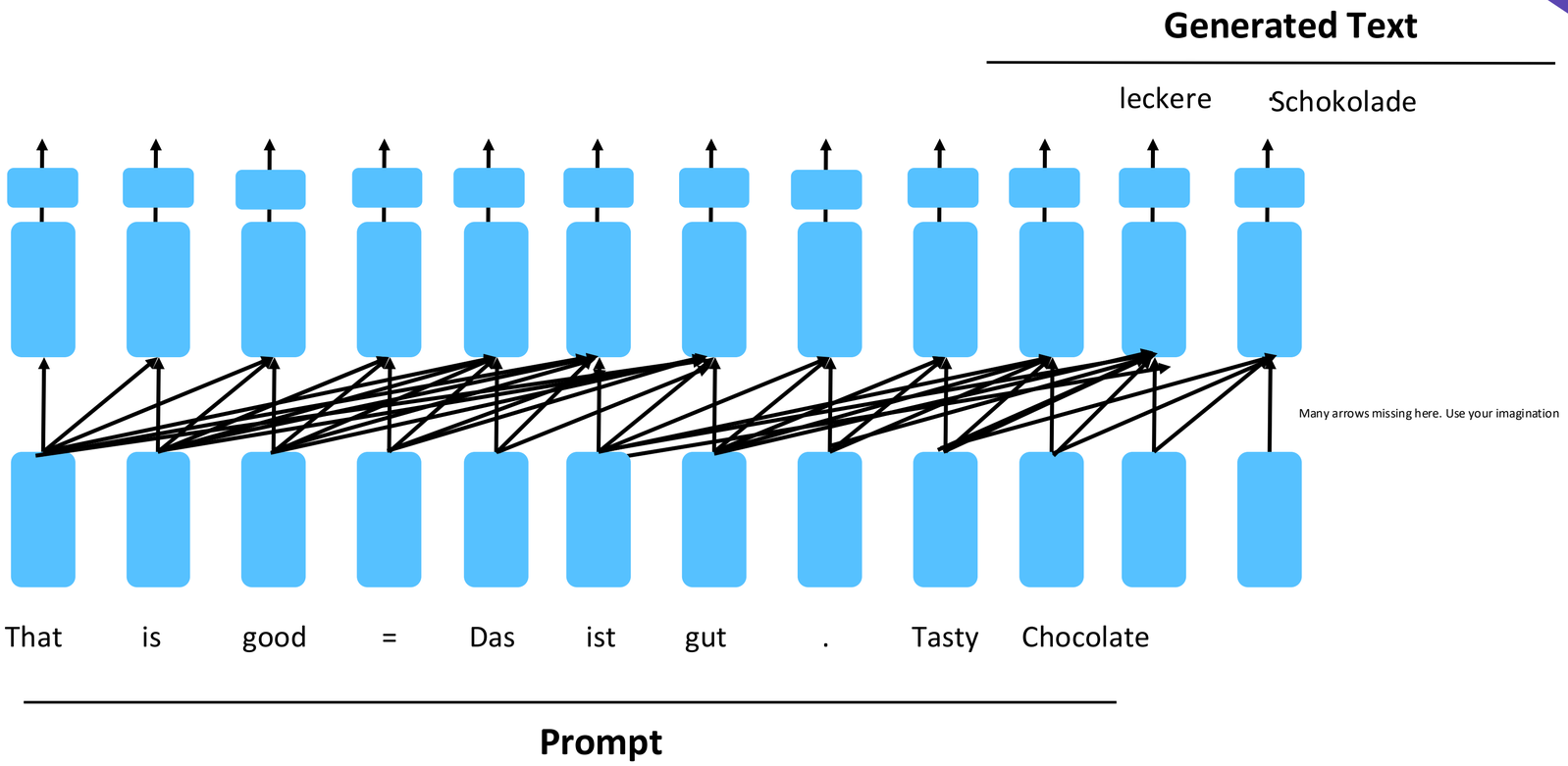
https://blog.research.google/2020/02/exploring-transfer-learning-with-t5.html

Enter GPT-2…

Decoder: GPT-2

Language Models are **Unsupervised Multitask Learners** [Radford et al., 2019]

- One of the most impressive things about GPT-2 was that it could obtain great performance on many NLP datasets zero-shot!



i.e. no fine-tuning and simply prompting the pre-trained model and generating the output

Much higher quality text

System Prompt (human-written)

Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today.

Model Completion (machine-written, 10 tries)

The 19-year-old singer was caught on camera being escorted out of the store by security guards.

The singer was wearing a black hoodie with the label 'Blurred Lines' on the front and 'Fashion Police' on the back.

Scroll down for video

Shoplifting: Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today (pictured)

The singer was wearing a black hoodie with the label 'Blurred Lines' on the front and 'Fashion Police' on the back

The singer was also wearing a pair of black-rimmed glasses, a black jacket, black jeans and black sandals.

Can zero-shot many tasks!
At SOTA levels!

Dataset	Metric	Our result	Previous record	Human
Winograd Schema Challenge	accuracy (+)	70.70%	63.7%	92%+
LAMBADA	accuracy (+)	63.24%	59.23%	95%+
LAMBADA	perplexity (–)	8.6	99	~1–2
Children's Book Test Common Nouns (validation accuracy)	accuracy (+)	93.30%	85.7%	96%
Children's Book Test Named Entities (validation accuracy)	accuracy (+)	89.05%	82.3%	92%
Penn Tree Bank	perplexity (–)	35.76	46.54	unknown
WikiText-2	perplexity (–)	18.34	39.14	unknown
enwik8	bits per character (–)	0.93	0.99	unknown
text8	bits per character (–)	0.98	1.08	unknown
WikiText-103	perplexity (–)	17.48	18.3	unknown

GPT-2 achieves state-of-the-art on Winograd Schema, LAMBADA, and other language modeling tasks.

Why does this work?

"I'm not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I'm not a fool].**
In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "**Mentez mentez, il en restera toujours quelque chose**," which translates as, "**Lie lie and something will always remain.**"
"I hate the word '**perfume**,"' Burr says. 'It's somewhat better in French: '**parfum**.'
If listened carefully at 29:55, a conversation can be heard between two guys in French: "**-Comment on fait pour aller de l'autre coté? -Quel autre coté?**", which means "**- How do you get to the other side? - What side?**".
If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?
"**Brevet Sans Garantie Du Gouvernement**", translated to English: "**Patented without government warranty**".

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

Enter GPT-3…

GPT-3 Paper (Brown et al., 2020)

Method: "What if we made an autoregressive language model 10x bigger??"

Result: LMs can do in-context learning!!

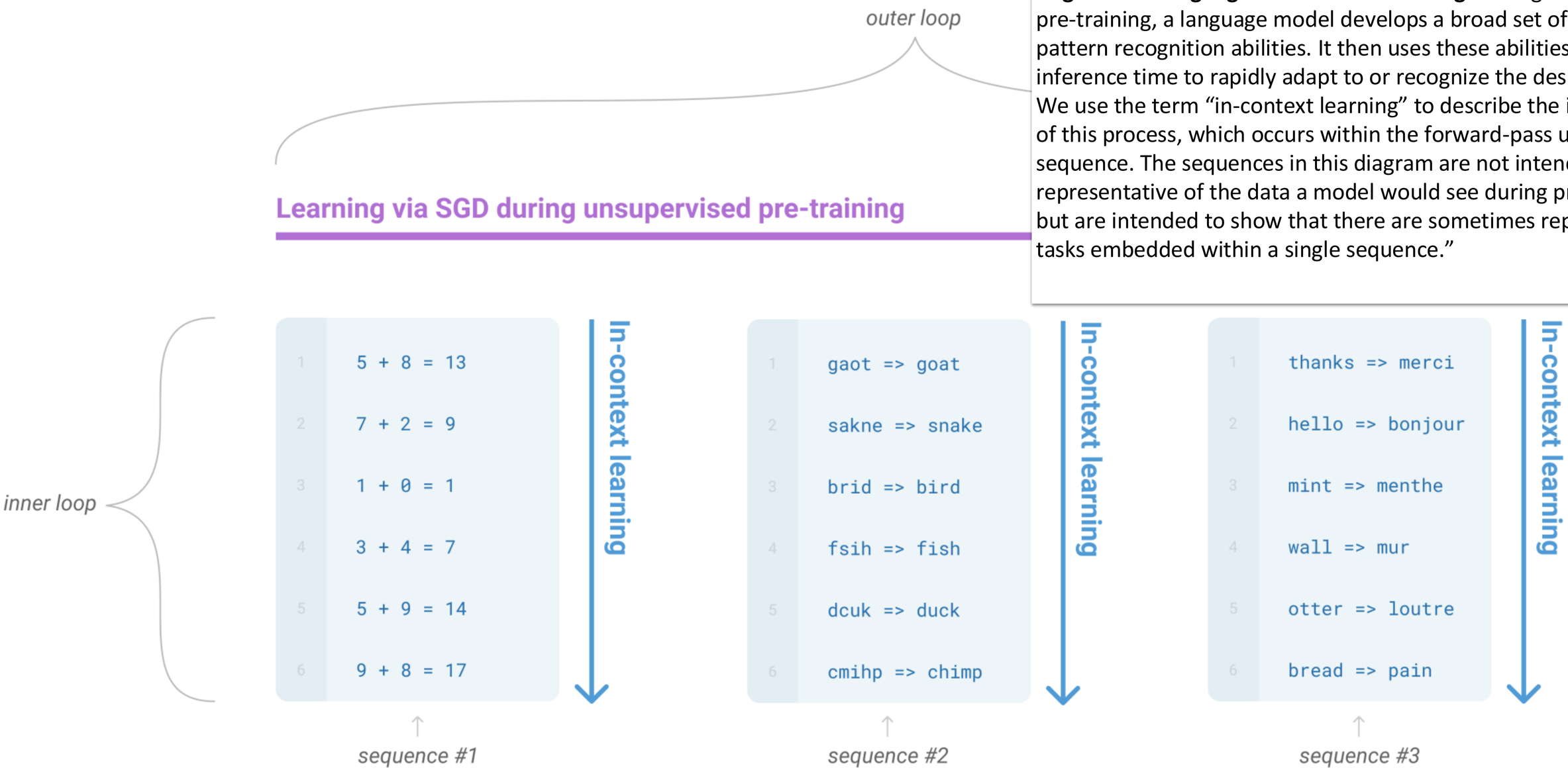
Language Models are Few-Shot Learners

Abstract

Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions – something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-3

tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning,

we find that GPT-3 can generate samples of news articles which human evaluators have difficulty



"**Figure 1.1: Language model meta-learning.** During unsupervised pre-training, a language model develops a broad set of skills and pattern recognition abilities. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. We use the term "in-context learning" to describe the inner loop of this process, which occurs within the forward-pass upon each sequence. The sequences in this diagram are not intended to be representative of the data a model would see during pre-training, but are intended to show that there are sometimes repeated sub-tasks embedded within a single sequence."

GPT-3 (Brown et. Al, 2020)

In-Context Learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

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

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
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
```

GPT-3 (Brown et. Al, 2020)

In-Context Learning

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
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
```

GPT-3 (Brown et. Al, 2020)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

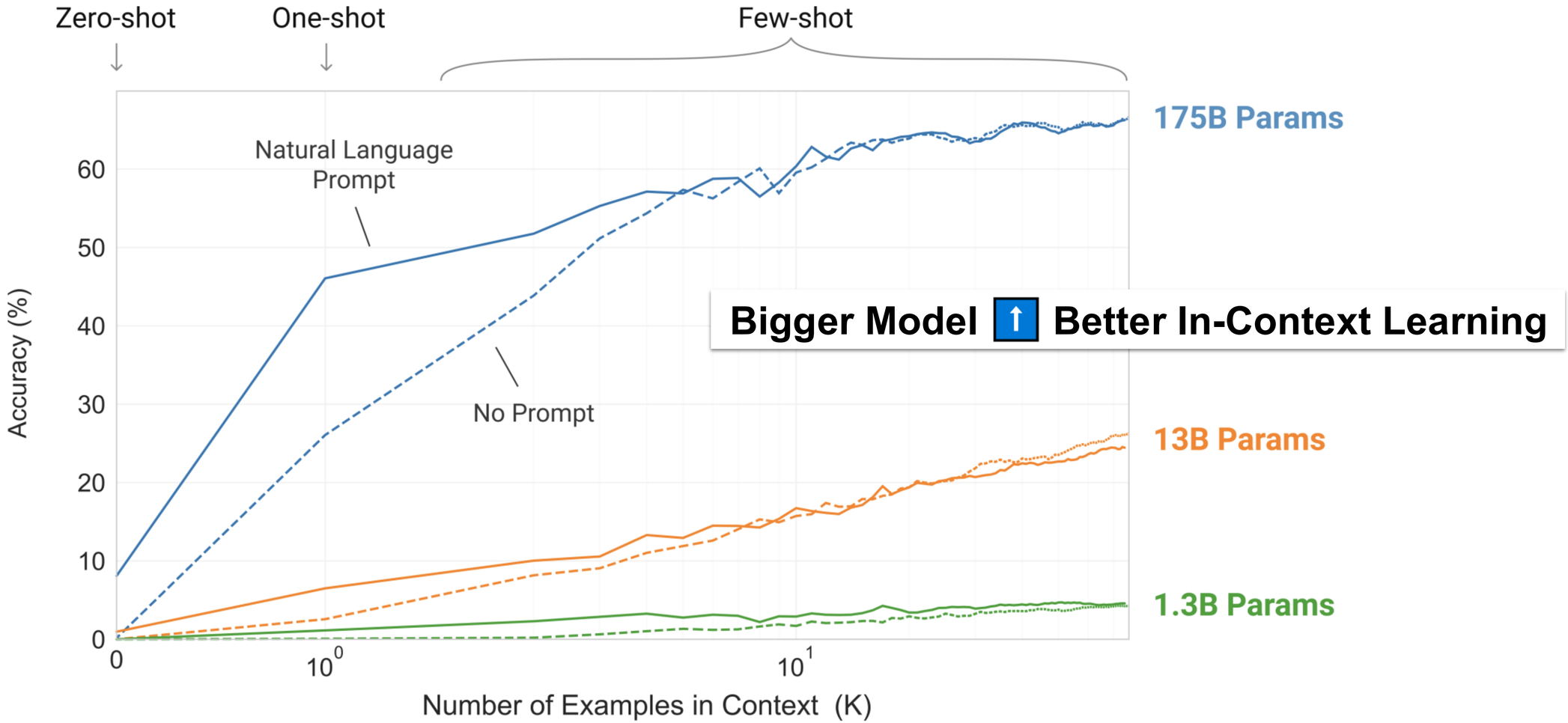


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

Questions

- Why does in-context learning work?
- What are benefits to in-context learning vs. other approaches?
- What are disadvantages to in-context learning?

Classic vs. Finetuning vs. In-Context

	Classic DL Approach	Pretrain/ Finetune	In-Context Learning
# Task-Specific Training Data	>1M (ideally)	10k	0(!)-20
Pretrain on NL Data	No	Yes	Yes
Gradient updates on training data	Yes	Yes	No
Where does "learning" come from	**Statistics of training data**	**Language representations from pretraining, modified to finetune data statistics**	**Language rep. from pretraining, mimic description of task + examples**

Difficulties with in-context learning

- The way you phrase a prompt can drastically affect performance (Sorensen et al., 2022, Sclar et al., 2023)
- Depends somewhat on the existence of a similar task in training data
- Not easy to tell *ex ante* if a model can reliably perform a certain task

An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels

Taylor Sorensen*, Joshua Robinson*, Christopher Michael Rytting*, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, David Wingate
Computer Science Department, Brigham Young University
{tsor13,joshua_robinson,chrisrytting}@byu.edu
{nfulda,wingated}@cs.byu.edu

QUANTIFYING LANGUAGE MODELS' SENSITIVITY TO SPURIOUS FEATURES IN PROMPT DESIGN *or: How I learned to start worrying about prompt formatting*

Melanie Sclar[1] Yejin Choi[1,2] Yulia Tsvetkov[1] Alane Suhr[3]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
[3]University of California, Berkeley
msclar@cs.washington.edu

Scrambling the labels doesn't hurt much…

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Sewon Min[1,2] **Xinxi Lyu**[1] **Ari Holtzman**[1] **Mikel Artetxe**[2]
Mike Lewis[2] **Hannaneh Hajishirzi**[1,3] **Luke Zettlemoyer**[1,2]
[1]University of Washington [2]Meta AI [3]Allen Institute for AI
{sewon,alrope,ahai,hannaneh,lsz}@cs.washington.edu
{artetxe,mikelewis}@meta.com

Figure 1: Results in classification (top) and multi-choice tasks (bottom), using three LMs with varying size. Reported on six datasets on which GPT-3 is evaluated; the channel method is used. See Section 4 for the full results. In-context learning performance drops only marginally

Okay, so bigger is better? Can you be more specific?

Scaling Laws

Scaling Laws (Kaplan et al., 2020)

- Kaplan et al., 2020 (OpenAI) explore how performance scales w.r.t. several parameters
- Vary:
  - Scale: - # Model Params, - Dataset size (tokens)
  - Other hyperparameters: Hidden layer sizes, context length, batch size
- Goal: Can we reliably predict test loss based on training scale (parameters and dataset size)?

Scaling Laws (Kaplan et al., 2020)

Result: Test loss very closely follows a *power law:*

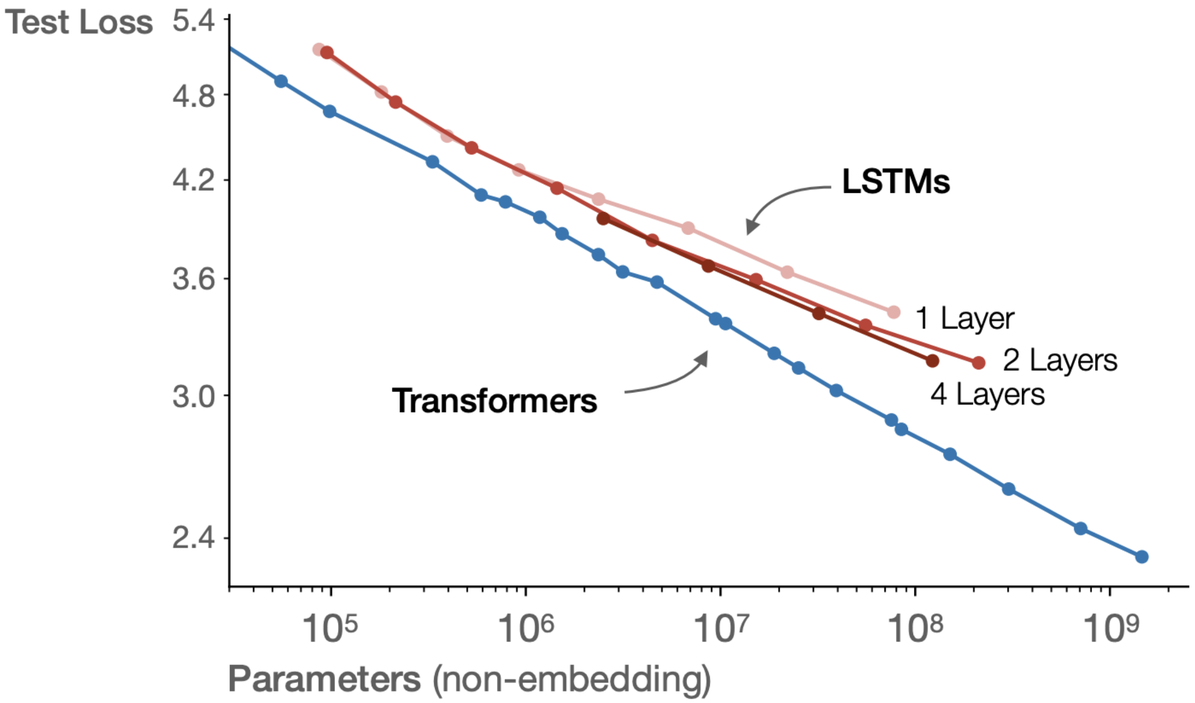
- Given constant dataset size ,

$$L(N) \approx \left(\frac{N_c}{N}\right)^{\alpha N}$$

- Given constant model size ,

$$L(D) \approx \left(\frac{D_c}{D}\right)^{\alpha D}$$

To linearly decrease test loss , you need to exponentially increase dataset size or model size



Scaling Laws (Kaplan et al., 2020)

Result: Test loss very closely follows a *power law:*

- Given constant dataset size ,

$$L(N) \approx \left(\frac{N_c}{N}\right)^{\alpha N}$$

- Given constant model size ,

$$L(D) \approx \left(\frac{D_c}{D}\right)^{\alpha D}$$

Parameter	α_N	α_D	N_c	D_c
Value	0.076	0.103	6.4×10^{13}	1.8×10^{13}

← Empirical estimates of parameters from experiments

Bringing it together:

$$L(N, D) \approx \left[\left(\frac{N_c}{N}\right)^{\frac{\alpha N}{\alpha D}} + \frac{D_c}{D}\right]^{\alpha D}$$

Scaling Laws (Kaplan et al., 2020)

- Bringing it together:

$$L(N, D) \approx \left[\left(\frac{N_c}{N} \right)^{\frac{\alpha N}{\alpha D}} + \frac{D_c}{D} \right]^{\alpha D}$$

- It requires a certain compute budget (FLOPs) to train a model of size on tokens
- Given a fixed compute budget, can calculate “optimal” (lowest test loss) choices using scaling law:

$$N_{\text{opt}}(C), D_{\text{opt}}(C) = \underset{N,D \text{ s.t. } \text{FLOPs}(N,D)=C}{\text{argmin}} L(N, D)$$

Parameter	α_N	α_D	N_c	D_c
Value	0.076	0.103	6.4×10^{13}	1.8×10^{13}

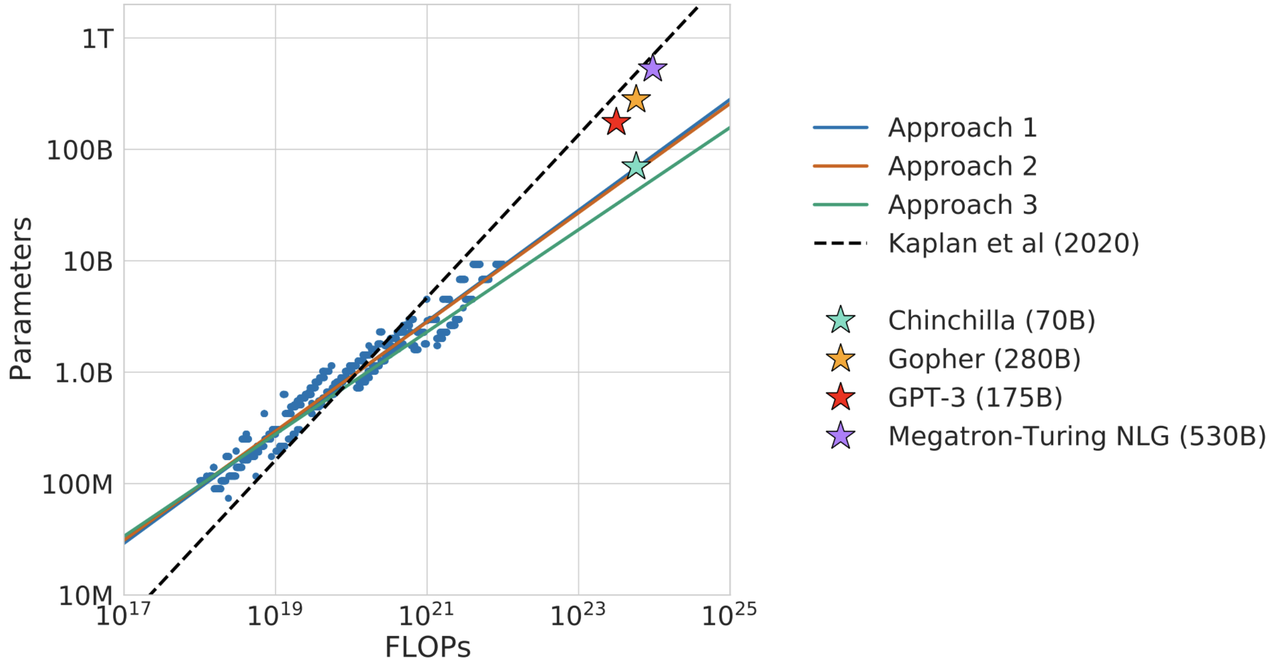
Table 2 Fits to $L(N, D)$

Note: Power laws in this paper (Jan 2020) inspired GPT-3 size (May 2020), and generalized to models 10x tested size!

Chinchilla (Hoffman et al., 2022)

DeepMind iterated on (Kaplan, 2020):

- Found better scaling law (lower test loss) by:
  - Changing learning rate scheduler
  - Testing larger models
- According to scaling laws, model should be smaller and tokens bigger for given compute budget
- Trained model based on new-scaling law: Chinchilla



$$\underset{N,D \text{ s.t. } \mathrm{FLOPs}(N,D)=C}{\operatorname{argmin}} L(N,D)$$

Chinchilla (Hoffman et al., 2022)

DeepMind iterated on (Kaplan, 2020):

- Found better scaling law (lower test loss) by:
  - Changing learning rate scheduler
  - Testing larger models
- According to scaling laws, model should be smaller and tokens bigger for given compute budget
- Trained model based on new-scaling law: Chinchilla

"Chinchilla optimal": Optimal model size/dataset size for a given compute budget, according to improved scaling laws

$$\underset{N,D \text{ s.t. } \mathrm{FLOPs}(N,D)=C}{\operatorname{argmin}} L(N, D)$$

LLaMA (Touvron et al., 2023)

- OpenAI/Deepmind only looked at the optimal size given a fixed *training* compute budget

$$\underset{N,D \text{ s.t. } \mathrm{FLOPs}(N,D)=C}{\operatorname{argmin}} L(N, D)$$

- What if you care more about *inference* time compute cost?
- Smaller model => Smaller inference cost
- To get best small model, should just train a small model on as much data as possible (beyond “Chinchilla-optimal”)
- “Overtrained” LLaMA-13B outperformed GPT-3 on many benchmarks

The Bitter Lesson

- Richard Sutton claimed: “The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.”
- Others claim “Scale is all you need”
- *What do you think?*

More Recently

- A lot of recent progress has been made from training bigger models on more data: LLaMA 2, GPT-4, Gemini, Mistral, etc.
  - Note: quality matters too! Need more *high-quality data,* low-quality data does not improve performance
- Limits of scale:
  - Limits on data: Modern LLMs are trained on bas
    10 new internets out of nowhere
  - Limits on compute: Big tech companies can't co
    much longer

TECHNOLOGY | ARTIFICIAL INTELLIGENCE

Sam Altman Seeks Trillions of Dollars to Reshape Business of Chips and AI

OpenAI chief pursues investors including the U.A.E. for a project possibly requiring up to $7 trillion

By Keach Hagey Follow and Asa Fitch Follow

Feb. 8, 2024 9:00 pm ET

(For context: $7T is more than GDP of all countries except US and China! Japan: $4.2T, Germany: $4T, …)

But that won't stop Sam Altman from trying!

Emergent Capabilities

Emergent Abilities of Large Language Models

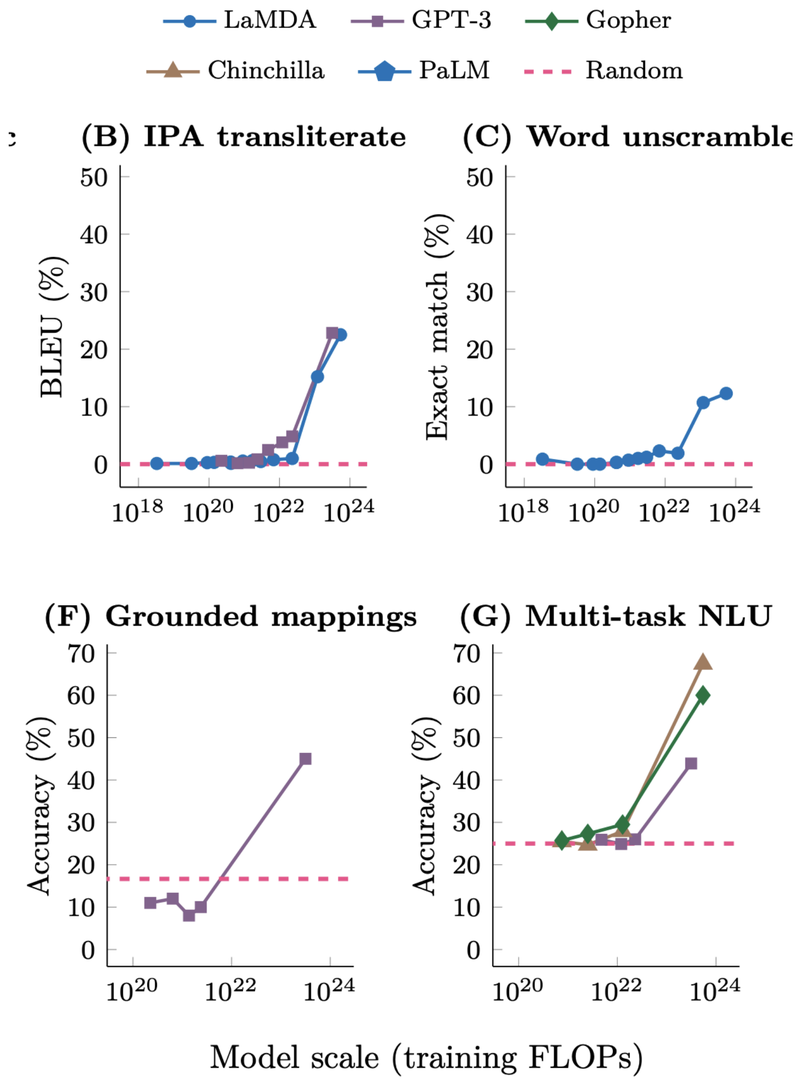
Jason Wei [1] *jasonwei@google.com*
Yi Tay [1] *yitay@google.com*
Rishi Bommasani [2] *nlprishi@stanford.edu*
Colin Raffel [3] *craffel@gmail.com*
Barret Zoph [1] *barretzoph@google.com*
Sebastian Borgeaud [4] *sborgeaud@deepmind.com*
Dani Yogatama [4] *dyogatama@deepmind.com*
Maarten Bosma [1] *bosma@google.com*
Denny Zhou [1] *dennyzhou@google.com*
Donald Metzler [1] *metzler@google.com*
Ed H. Chi [1] *edchi@google.com*
Tatsunori Hashimoto [2] *thashim@stanford.edu*
Oriol Vinyals [4] *vinyals@deepmind.com*
Percy Liang [2] *pliang@stanford.edu*
Jeff Dean [1] *jeff@google.com*
William Fedus [1] *liamfedus@google.com*

[1] *Google Research* [2] *Stanford University* [3] *UNC Chapel Hill* [4] *DeepMind*

Emergent Capabilities (Wei et al., 2022)

Test loss predictably improves with scale - but what about capabilities?

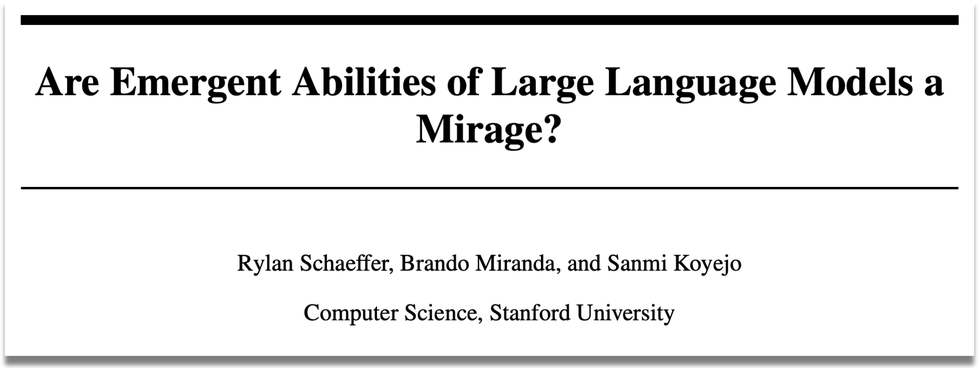
- "Emergence is when quantitative changes in a system result in qualitative changes in behavior." (Anderson, 1972)
- "An ability is emergent if it is not present in smaller models but is present in larger models." (Wei et al., 2022)
- Corollary: *It may fundamentally be hard to predict capabilities of future models*



Emergent capabilities a mirage?

(Schaeffer et al., 2023) take issue with the characterization of “emergent capabilities”

- Most metrics used in (Wei et al., 2022) were “hard” metrics which don’t give partial credit like accuracy



Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo

Computer Science, Stanford University

Hard Accuracy:

A) 123 + 456 = 579 ✅

B) 123 + 456 = 578 ❌

C) 123 + 456 = 42 ❌

In (Wei et al., 2022), B and C are both wrong, even though B is much closer to correct than C

Emergent capabilities a mirage?

(Schaeffer et al., 2023) measure soft metrics (e.g., how many digits are correct, probability of the right answer) for "emergent abilities"

- Find much more predictable scaling
- Different metric choices lead to different appearances of "emergent" or not emergent
- "Emergent abilities" are a mirage(?)

Hard Accuracy:

A) 123 + 456 = 579 ✅

B) 123 + 456 = 578 ❌

C) 123 + 456 = 42 ❌

Soft Accuracy (# correct digits):

A) 123 + 456 = 579 3/3 ✅

B) 123 + 456 = 578 2/3 🤷

C) 123 + 456 = 42 0/3 ❌

Other Phenomena

Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve

R. Thomas McCoy Shunyu Yao Dan Friedman Matthew Hardy Thomas L. Griffiths

Princeton University

Fundamentally, LLMs are doing next-word prediction on Internet text

Hypothesis: LLM performance will depend on:

- Probability of the task being performed
- Probability of the target output
- Probability of the target input

Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve

R. Thomas McCoy Shunyu Yao Dan Friedman Matthew Hardy Thomas L. Griffiths

Princeton University

Counting

Count the letters.

Input 1: iiiiiiiiiiiiiiiiiiiiiiiiiiiii
Correct: 30
✓ **GPT-4:** 30

Input 2: iiiiiiiiiiiiiiiiiiiiiiiiiiiii
Correct: 29
✗ **GPT-4:** 30

30 shows up much more often than 29 in pretraining text

Shift ciphers

Decode by shifting each letter 13 positions backward in the alphabet.

Input: Jryy, vg jnf abg rknpgyl cynaarq sebz gur ortvaavat.
Correct: Well, it was not exactly planned from the beginning.
✓ **GPT-4:** Well, it was not exactly planned from the beginning.

Decode by shifting each letter 12 positions backward in the alphabet.

Input: Iqxx, uf ime zaf qjmofxk bxmzzqp rday ftq nqsuzzuzs.
Correct: Well, it was not exactly planned from the beginning.
✗ **GPT-4:** Wait, we are not prepared for the apocalypse yet.

13-shifted ciphers are more common than 12-shifted ciphers online

Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve

R. Thomas McCoy Shunyu Yao Dan Friedman Matthew Hardy Thomas L. Griffiths

Princeton University

Article swapping

Swap each article (*a*, *an*, or *the*) with the word before it.

Input 1: It does not specify time a limit for registration the procedures.
Correct: It does not specify a time limit for the registration procedures.
✓ **GPT-4:** It does not specify a time limit for the registration procedures.

Input 2: It few with it to lying take the get just a hands would kinds.
Correct: It few with it to lying the take get a just hands would kinds.
✗ **GPT-4:** It flew with a few kinds to take the lying just to get the hands.

Grammatical text is more common than ungrammatical text

Linear functions

Multiply by **9/5** and add **32**.

Input: 328
Correct: 622.4
✓ **GPT-4:** 622.4

Multiply by **7/5** and add **31**.

Input: 328
Correct: 490.2
✗ **GPT-4:** 457.6

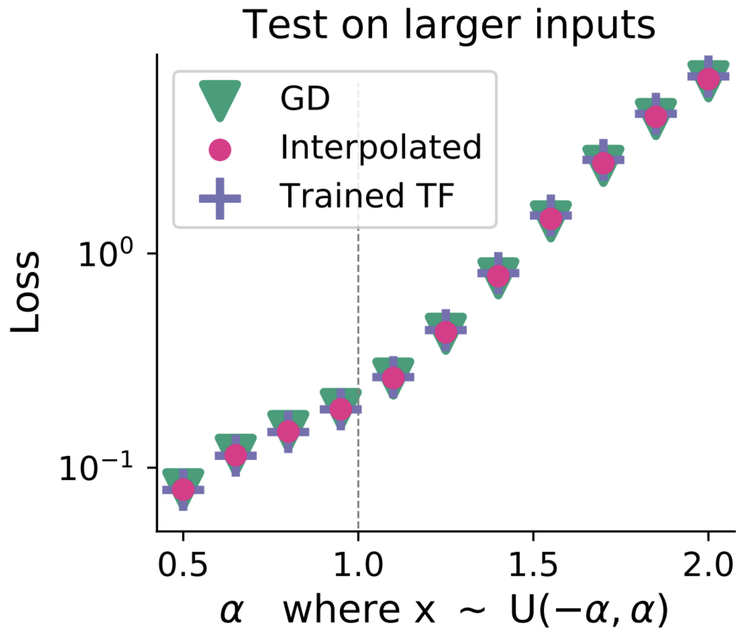
(9/5)x + 32 is common because it is the Celsius -> Fahrenheit conversion

Transformers Learn In-Context by Gradient Descent

Johannes von Oswald [1 2] Eyvind Niklasson [2] Ettore Randazzo [2] João Sacramento [1]
Alexander Mordvintsev [2] Andrey Zhmoginov [2] Max Vladymyrov [2]

How does in-context learning mechanistically work?

- Hypothesis: Maybe they actually can do gradient descent in context??
- Demonstrate that linear self-attention can emulate one step gradient descent on a linear regression task
- In-context predictions follow closely predictions one would arrive to with gradient descent

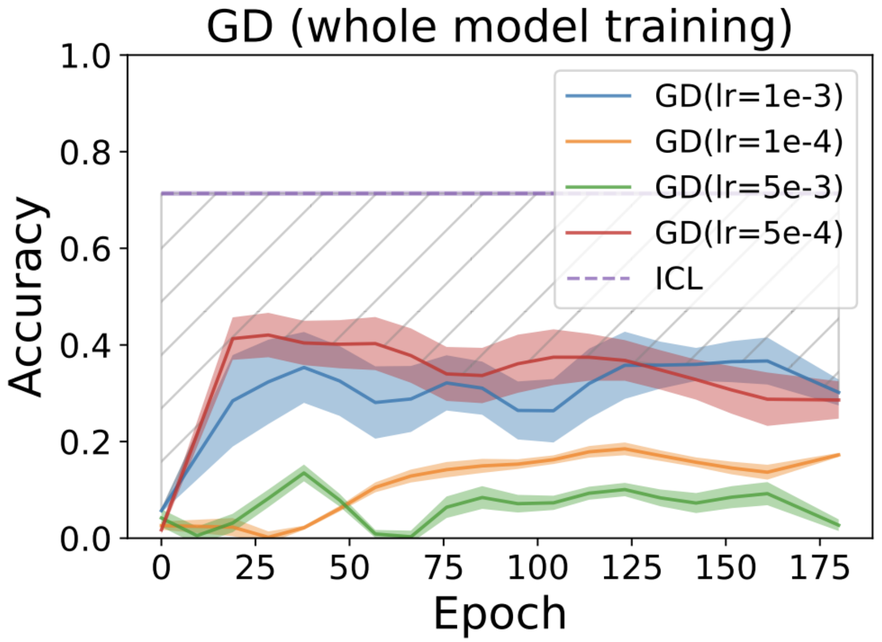


Do pretrained Transformers Really Learn In-context by Gradient Descent?

Lingfeng Shen[♡] **Aayush Mishra**[♡] **Daniel Khashabi**
Johns Hopkins University, Baltimore MD

Okay, so transformers *can* do gradient descent. But do LLMs *actually* do GD?

- Test on GPT-J and LLaMA
- Find a large gap between gradient descent and LLM in-context learning
- Open question whether LLMs do GD



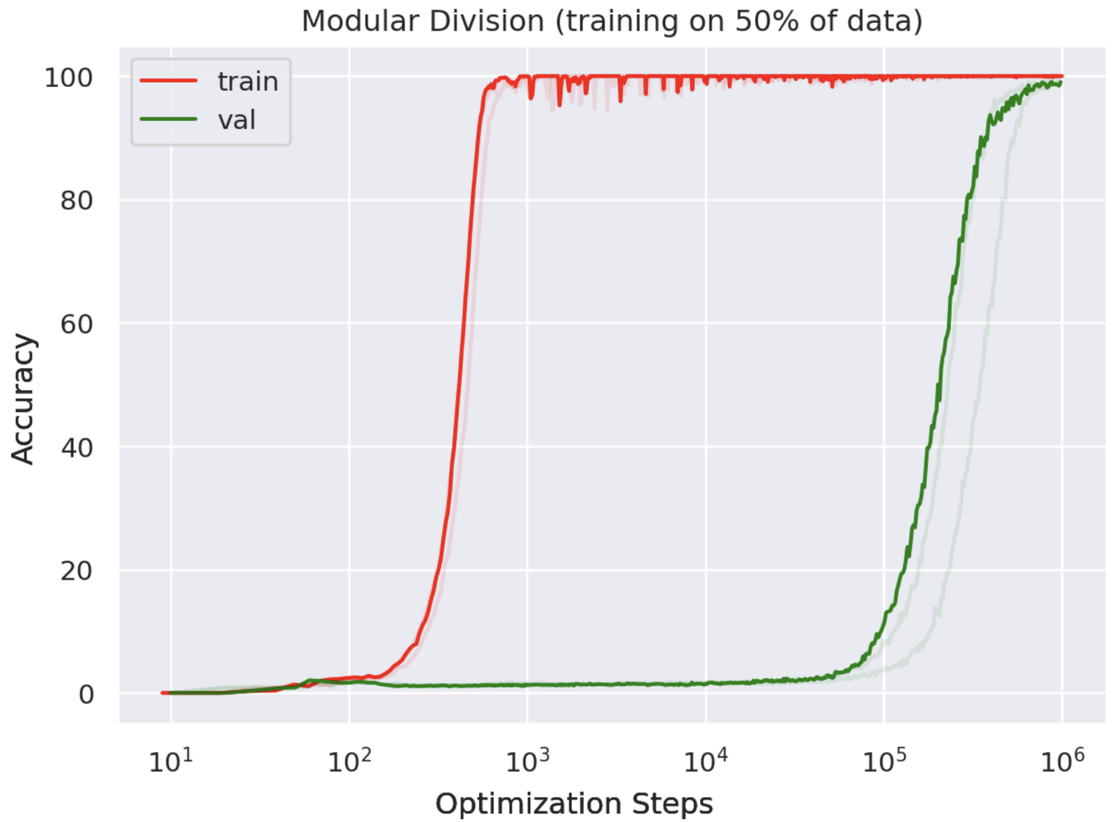
GROKKING: GENERALIZATION BEYOND OVERFITTING ON SMALL ALGORITHMIC DATASETS

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin
OpenAI

Vedant Misra*
Google

ML conventional wisdom: Training too long on the same data is *bad*, leads to overfitting

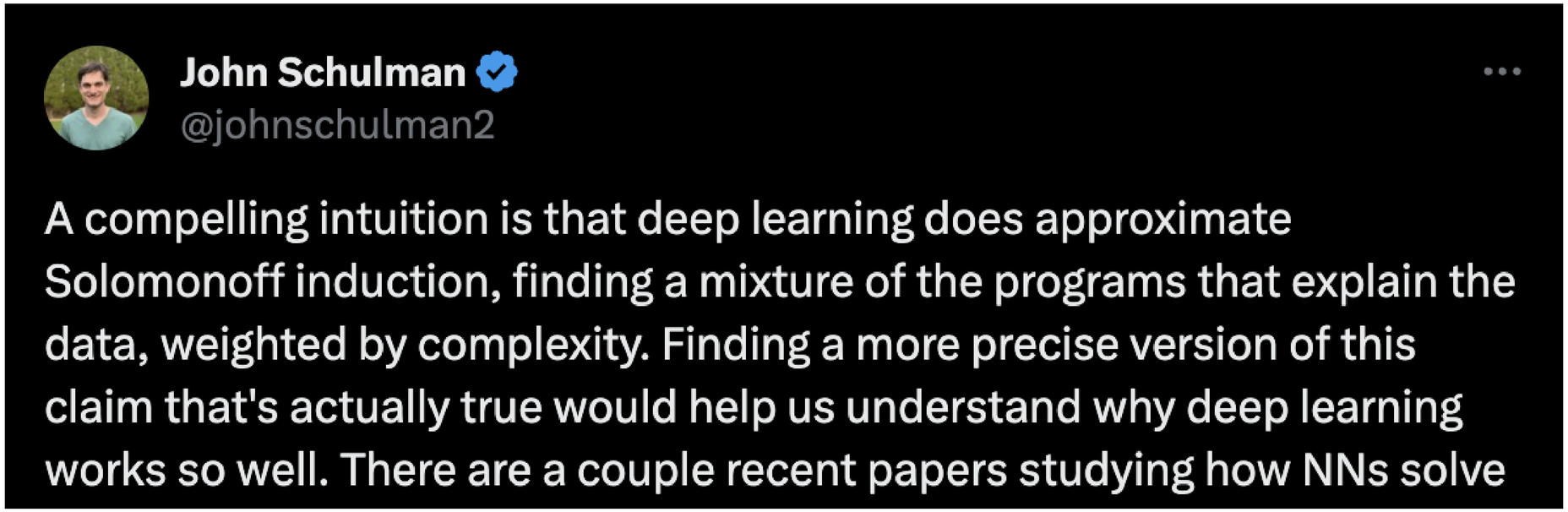
- "Grokking": models that have been overtrained on the same set of data somehow suddenly learn to generalize



GROKKING: GENERALIZATION BEYOND OVERFITTING ON SMALL ALGORITHMIC DATASETS

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin
OpenAI

Vedant Misra*
Google



https://x.com/johnschulman2/status/1741178475946602979?s=20

Thank you!