

Autoregressive LM

Wei Xu

(some slides from Greg Durrett)

This Lecture

- ▶ Decoder-style LMs: GPT / GPT-2 / GPT-3
- ▶ In-context Learning
- ▶ Scaling Laws

(Recap) BERT / Masked Language Models

- ▶ Key factors:
 - ▶ Big enough model
 - ▶ Big enough data
 - ▶ Well-designed “self-supervised” objective (something like language modeling). Needs to be a hard enough problem!
- ▶ There are lots of ways to train these models!

LM Intuitions

Winograd Schema

- ▶ Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU 1968-1972)

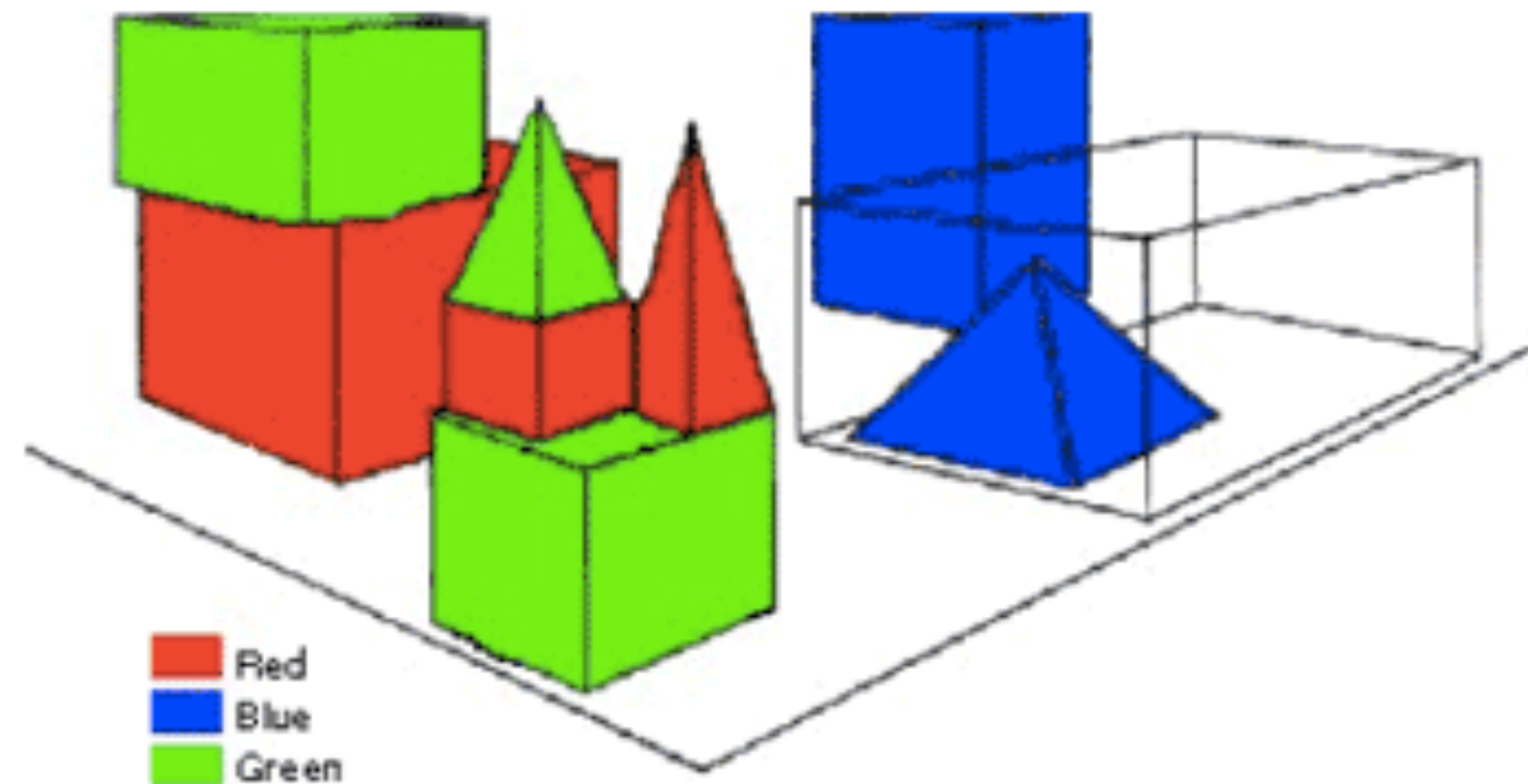
The city council refused the demonstrators a permit because they _____ violence

they advocated
they feared

- ▶ This is so complicated that it's an AI challenge problem! (AI-complete)
- ▶ Referential/semantic ambiguity

SHRDLU

- ▶ One of the first natural language understanding programs, developed by Terry Winograd at MIT in the late 1960s and early 1970s



Person: Pick up a big red block.

Computer: OK.

Person: Grasp the pyramid.

Computer: I don't understand which pyramid you mean.

Why is language modeling a good objective?

- ▶ “Impossible” problem but bigger models seem to do better and better at distributional modeling (no upper limit yet)
- ▶ Successfully predicting next words requires modeling lots of different effects in text

Context: My wife refused to allow me to come to Hong Kong when the plague was at its height and –” “Your wife, Johanne? You are married at last ?” Johanne grinned. “Well, when a man gets to my age, he starts to need a few home comforts.

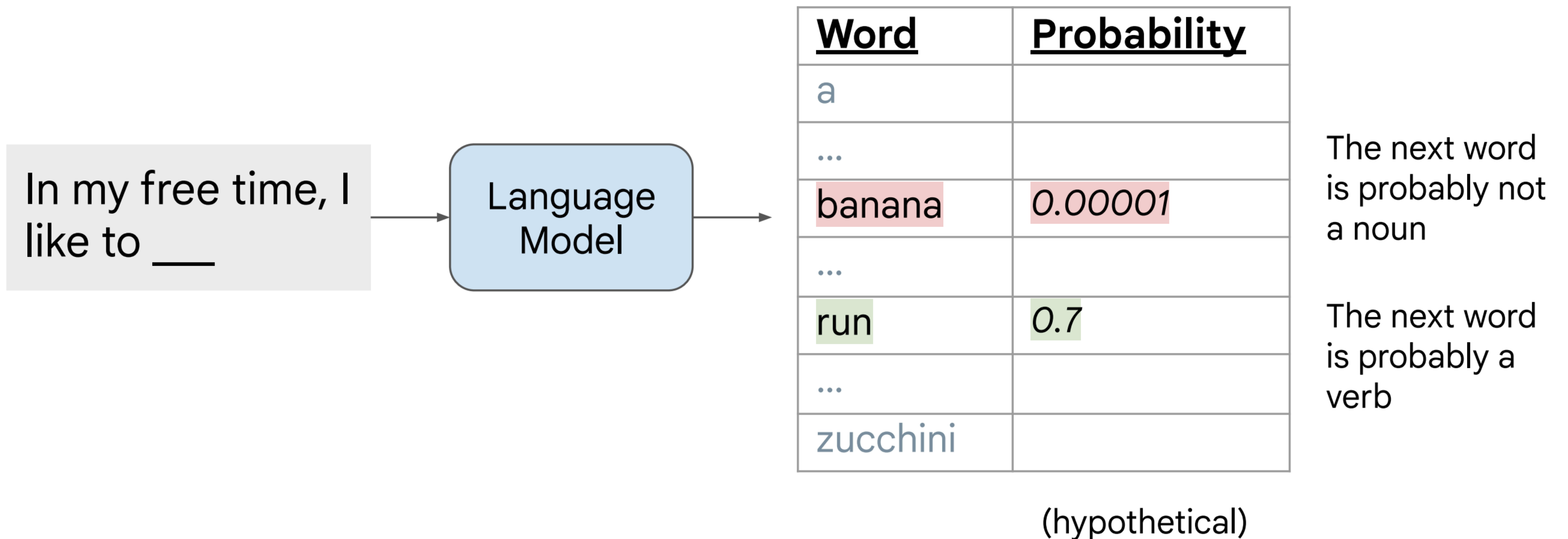
Target sentence: After my dear mother passed away ten years ago now, I became -----.

Target word: lonely

- ▶ LAMBADA dataset (Papernot et al., 2016): explicitly targets world knowledge and very challenging LM examples

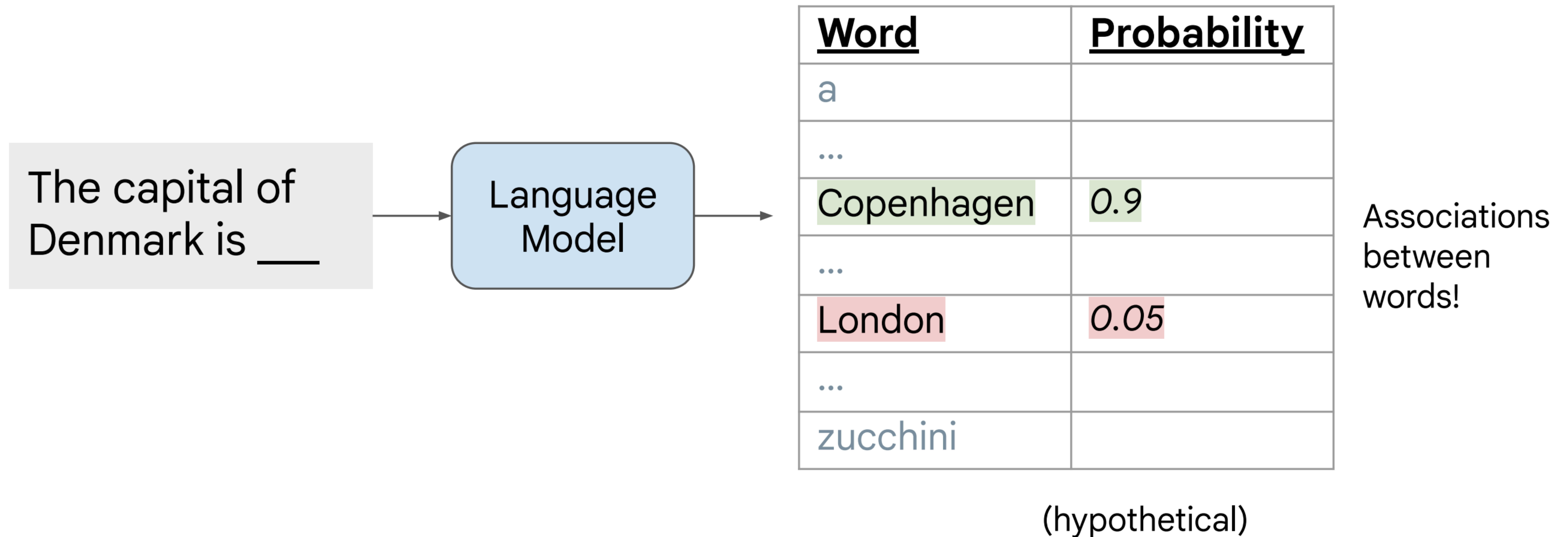
What do language model learn?

- ▶ Grammar



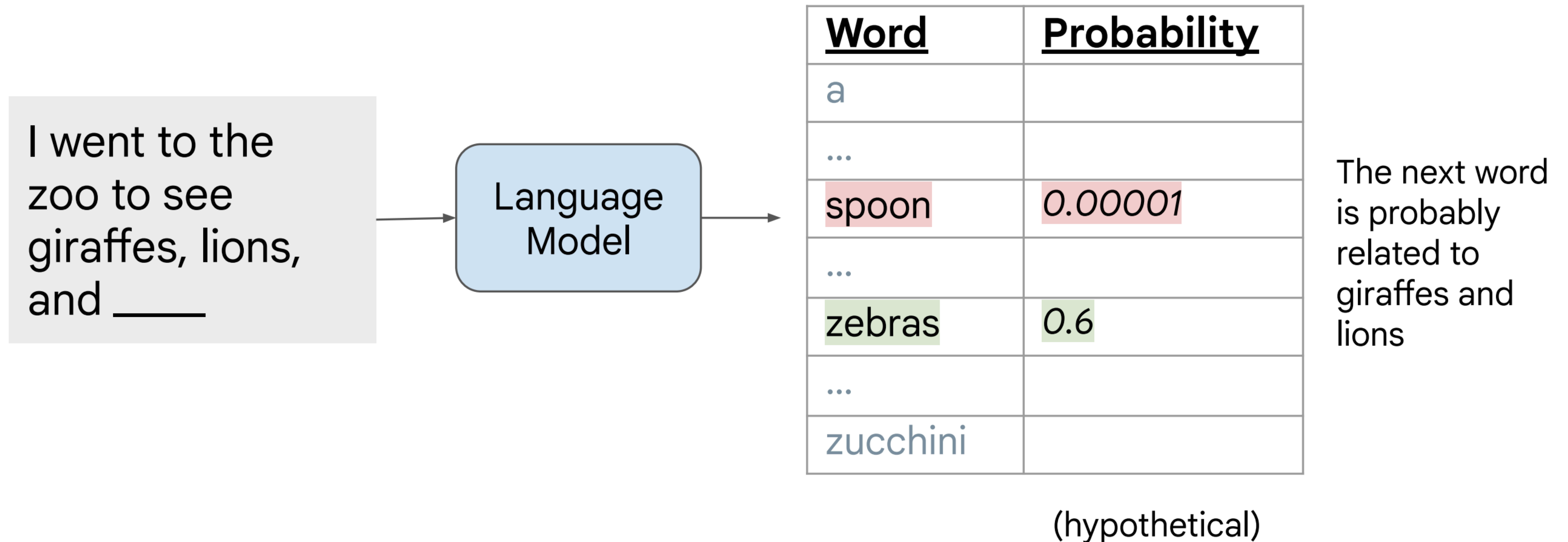
What do language model learn?

- Facts about the world



What do language model learn?

- Lexical semantics



What do language model learn?

- Sentiment analysis

I was engaged
and on the edge
of my seat the
whole time. The
movie was ____

Language
Model

<u>Word</u>	<u>Probability</u>
a	
...	
bad	0.1
...	
good	0.9
...	
zucchini	

Well, “engaged” is
pretty indicative
of a positive
sentiment

(hypothetical)

What do language model learn?

- ▶ Harder sentiment analysis

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was —

Language Model

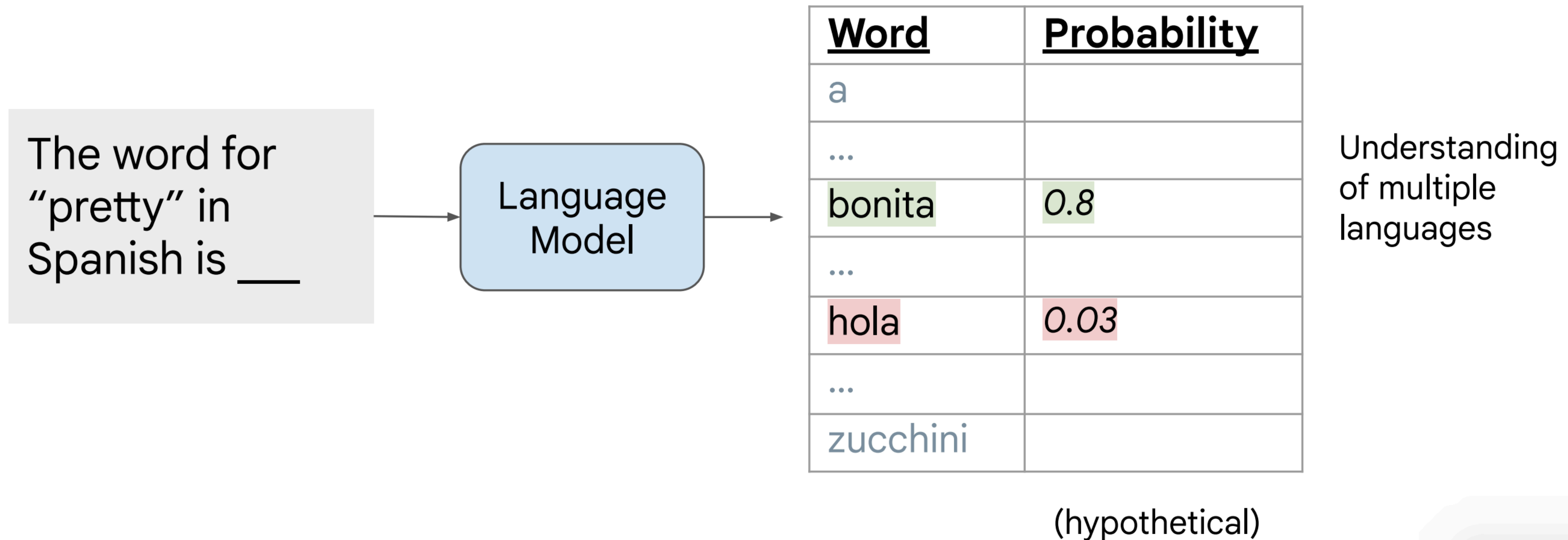
<u>Word</u>	<u>Probability</u>
a	
...	
bad	0.7
...	
good	0.3
...	
zucchini	

(hypothetical)

Some more-complex understanding needed

What do language model learn?

- Translation



What do language model learn?

- Spatial reasoning

Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the ____

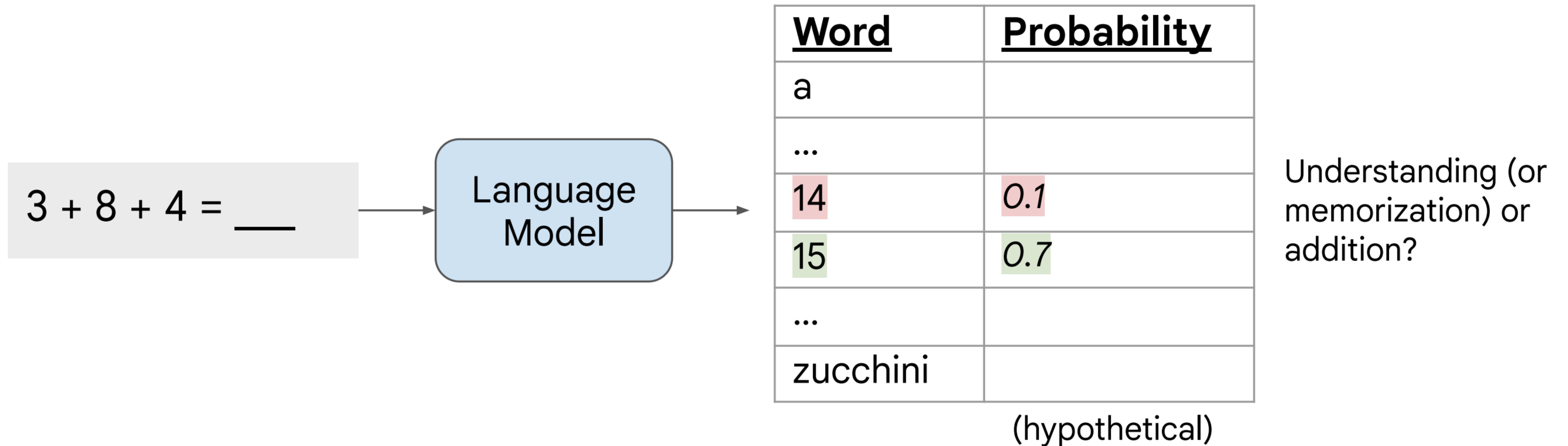
Language Model

<u>Word</u>	<u>Probability</u>
a	
...	
...	
kitchen	0.8
...	
zucchini	

(hypothetical)

What do language model learn?

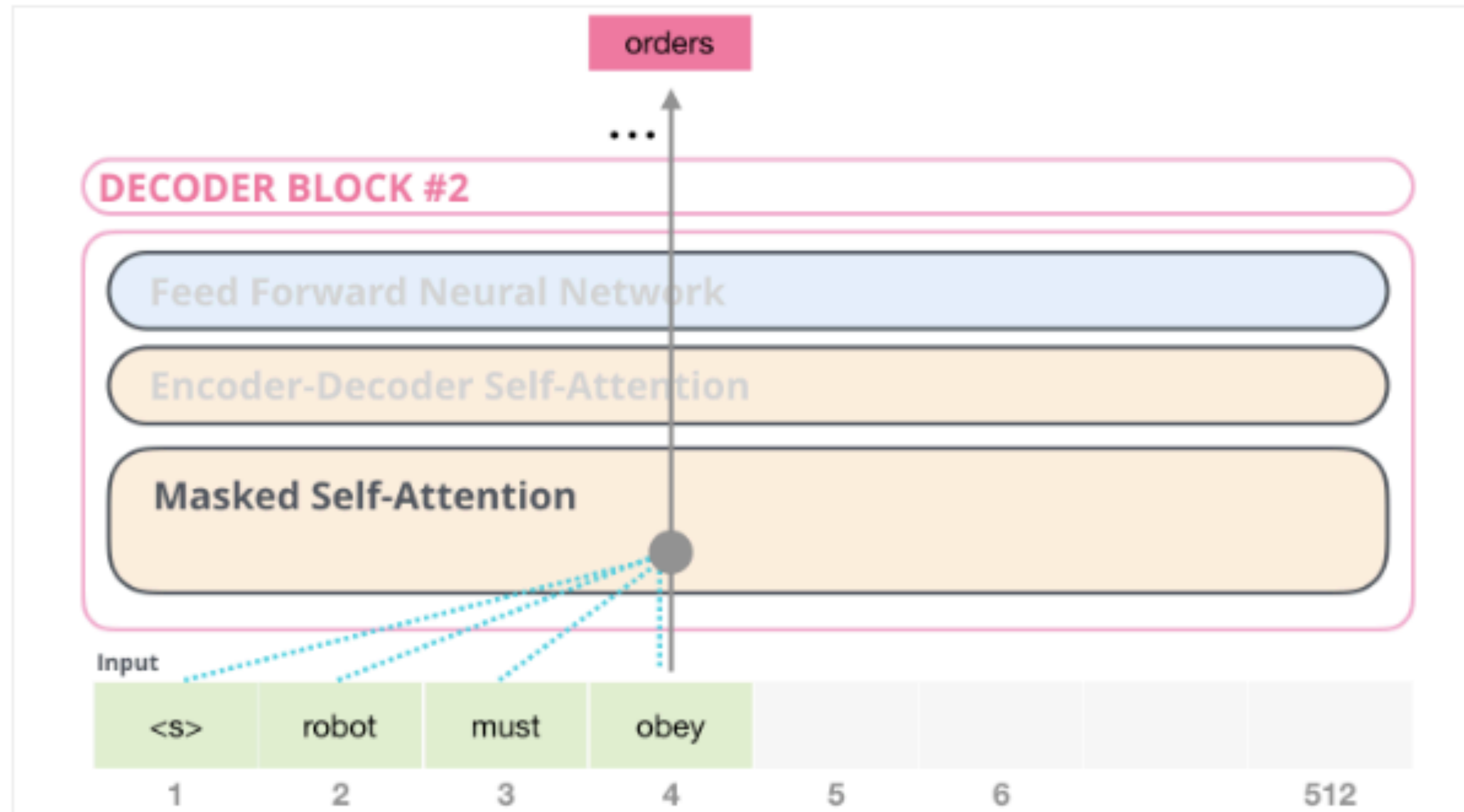
- Easy arithmetic



GPT/GPT2/GPT3

OpenAI GPT/GPT2

- ▶ Train a single unidirectional transformer LM on long contexts
- ▶ Masked self-attention: each token can only attend to past tokens



Radford et al. (2019)

OpenAI GPT/GPT2

- ▶ Train a single unidirectional transformer LM on long contexts
- ▶ Masked self-attention: each token can only attend to past tokens

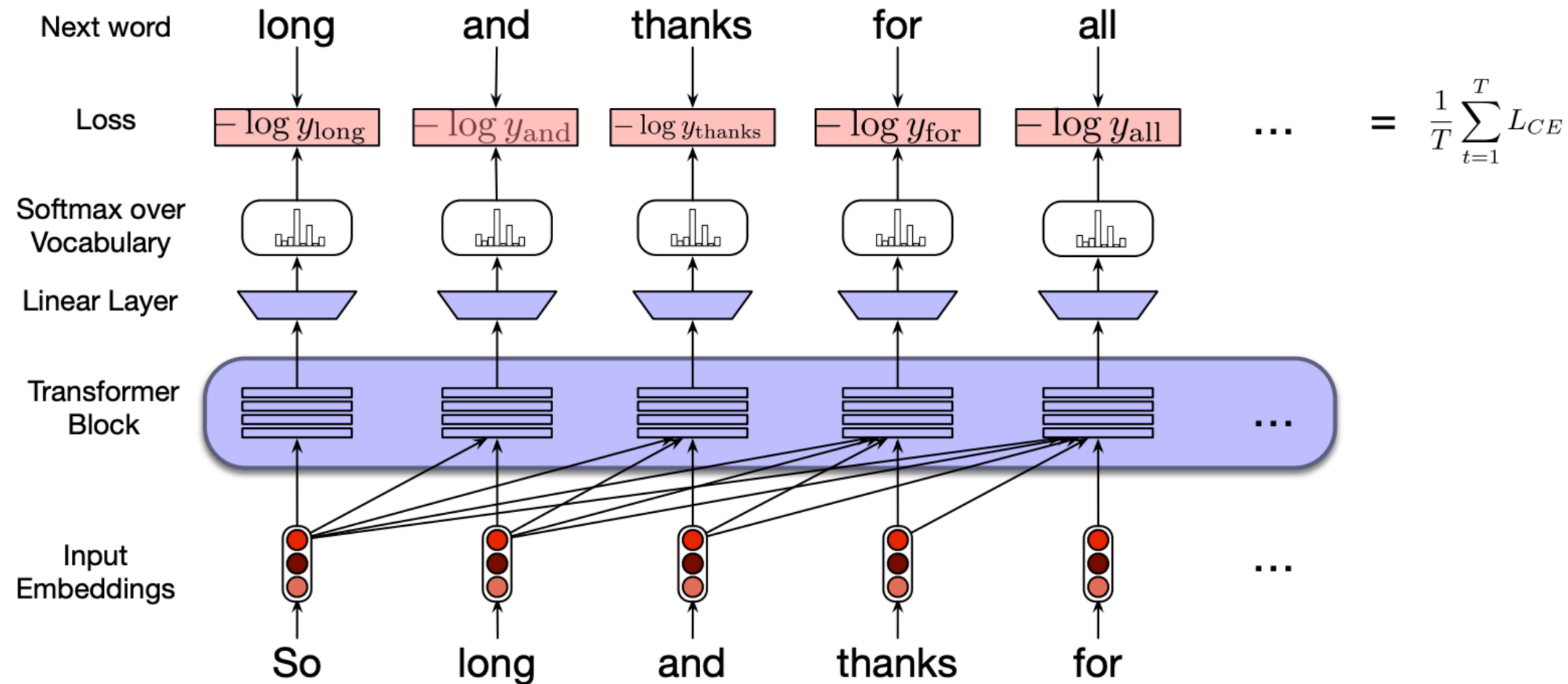


Image Credit: Jurafsky & Martin

OpenAI GPT/GPT2

- ▶ GPT2: trained on 40GB of text collected from upvoted links from reddit
- ▶ 1.5B parameters — the largest of these models trained as of March 2019

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

- ▶ Because it's a decoder-style language model, we can **generate** from it

GPT-2

- Some results from GPT-2 foreshadowed the benefits of scaling up language models (shown here for Winograd and LM Perplexity)

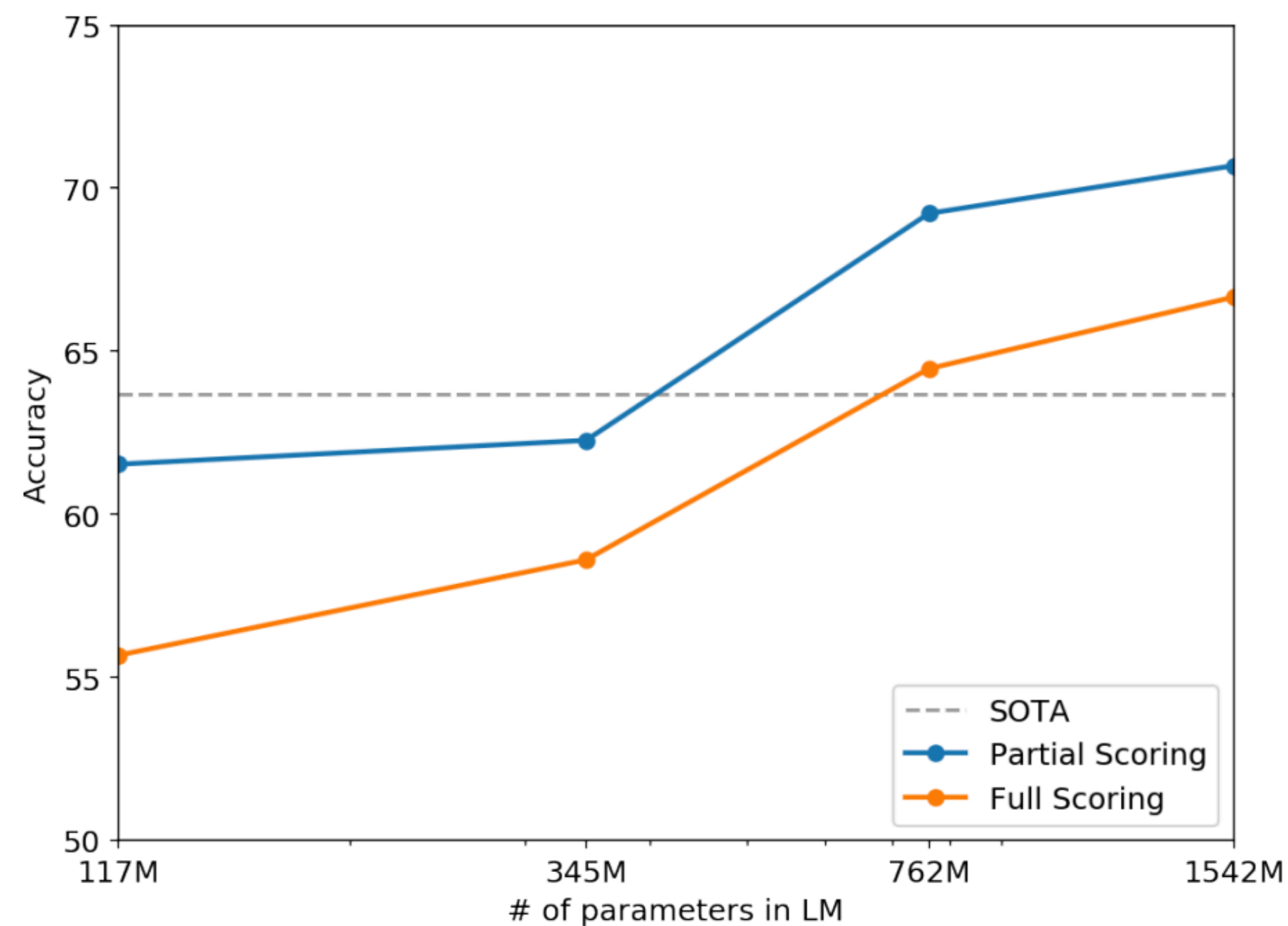


Figure 3. Performance on the Winograd Schema Challenge as a function of model capacity.

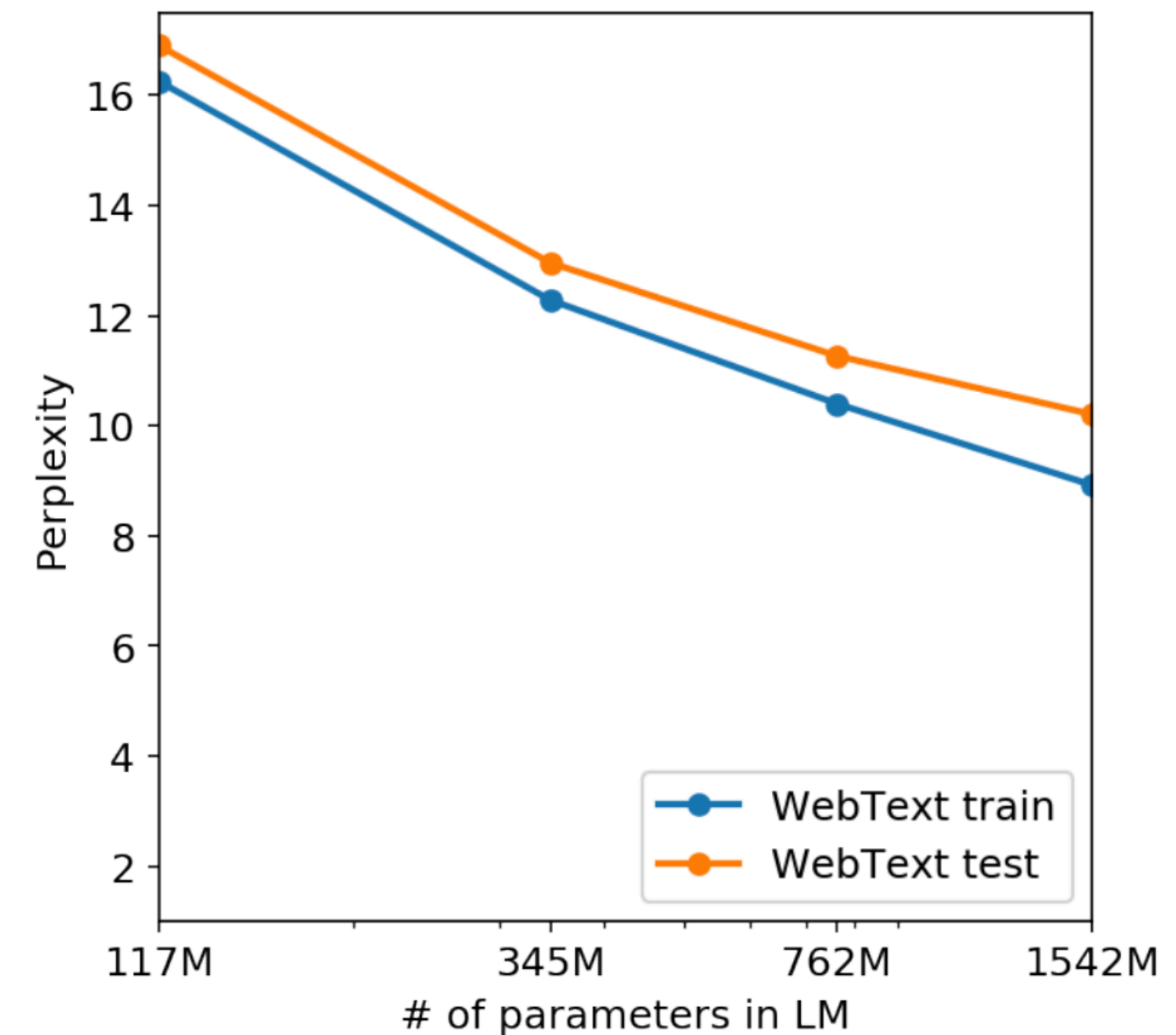
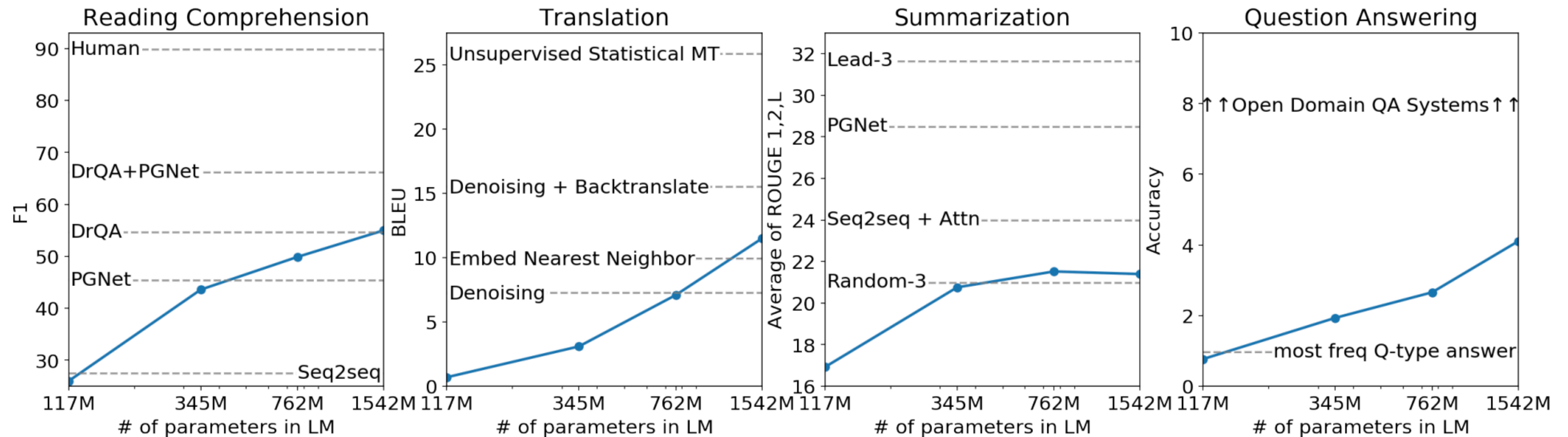


Figure 4. The performance of LMs trained on WebText as a function of model size.

GPT-2

- Started to show strong zero-shot performance, even though “in terms of practical applications, the zero-shot performance of GPT-2 is still far from use-able.”



Radford et al. (2019)

OpenAI GPT2

System Prompt
(human-written)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Model Completion
(machine-written, 10
tries)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move

slide credit:
OpenAI

OpenAI GPT2

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.

She was carrying a pair of black and white striped gloves and a small black bag.

slide credit:
OpenAI

Open Questions

- 1) How novel is the stuff being generated? (Is it just doing nearest neighbors on a large corpus?)
- 2) How do we understand and distill what is learned in this model?
- 3) How do we harness these priors for conditional generation tasks (summarization, generate a report of a basketball game, etc.)
- 4) Is this technology dangerous? (OpenAI pursued a “staged release”)

Ethical Considerations

Grover

- Sample from a large language model conditioned on a domain, date, authors, and headline

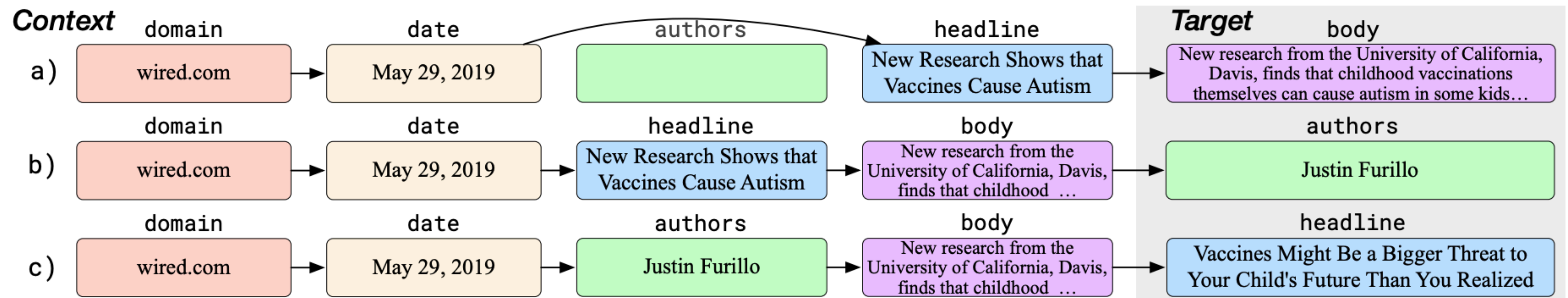


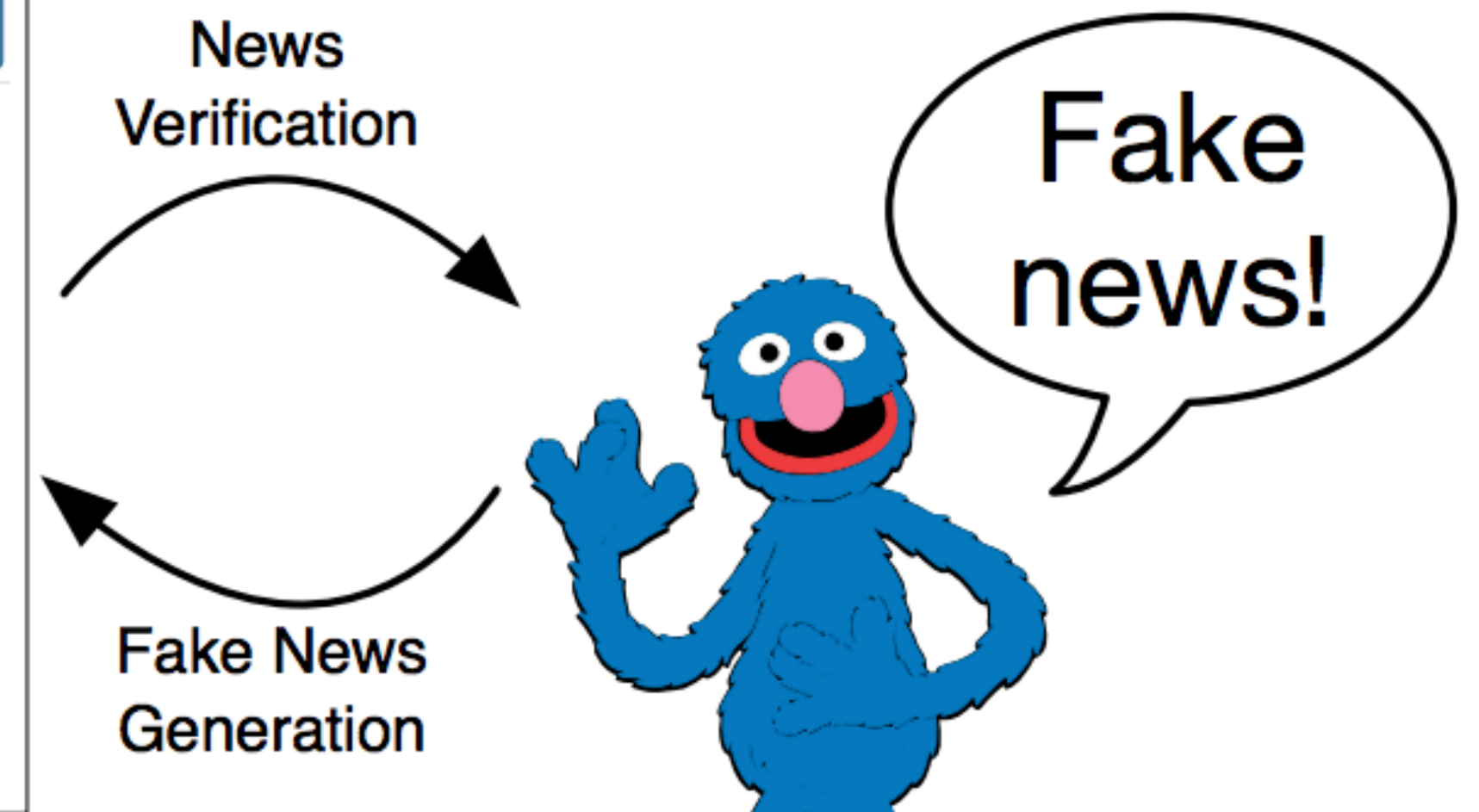
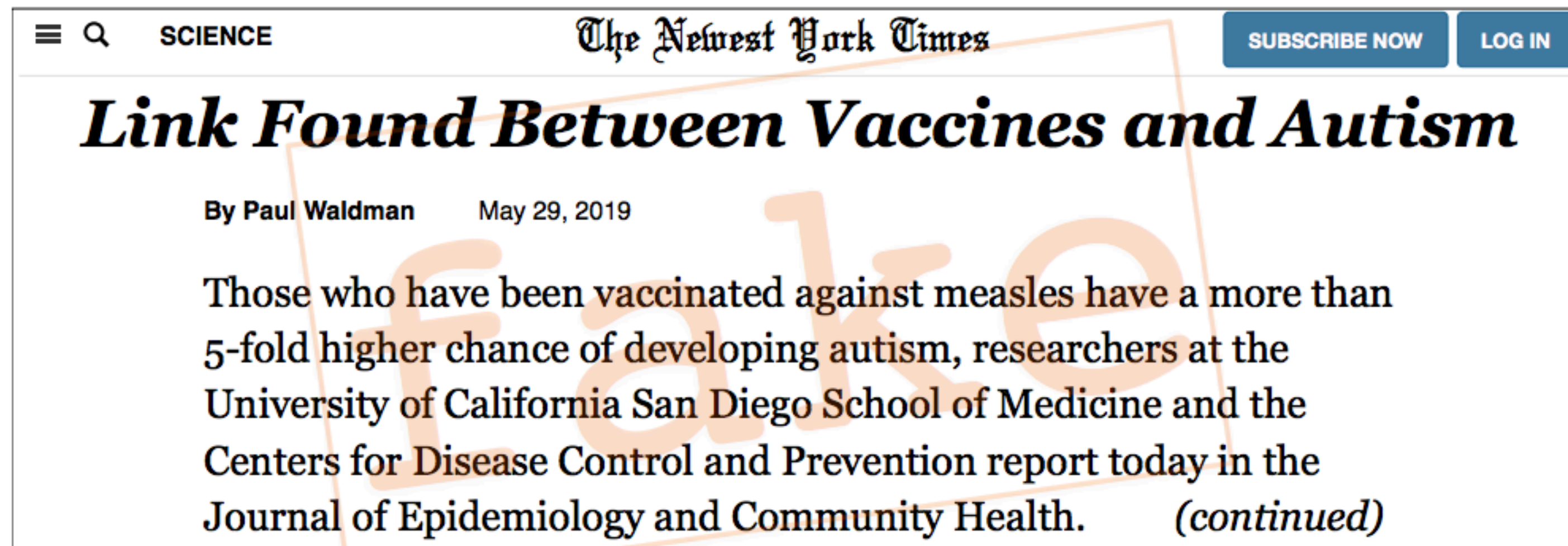
Figure 2: A diagram of three GROVER examples for article generation. In row a), the **body** is generated from partial context (the **authors** field is missing). In b), the model generates the **authors**. In c), the model uses the new generations to regenerate the provided **headline** to one that is more realistic.

- NOTE: Not a GAN, discriminator trained separately from the generator

Zellers et al. (2019)

Grover

- ▶ Humans rank Grover-generated propaganda as more realistic than real “fake news”



- ▶ Fine-tuned Grover can detect Grover propaganda easily — authors argue for releasing it for this reason

Zellers et al. (2019)

Bias and Toxicity

- ▶ “Toxic degeneration”: systems that generate toxic stuff

GENERATION OPTIONS:

Model:

Prompt:

Toxicity:

⚠ Toxic generations may be triggering.

I'm sick of all the politically correct talk and crying and looking down at your pathetic self and wishing you could just get outta there and...

- ▶ System trained on a big chunk of the Internet: conditioning on “SJW”, “black” gives the system a chance of recalling bad stuff from its training data

Pre-Training Cost (with Google/AWS)

- ▶ BERT: Base \$500, Large (340M parameters) \$7000
- ▶ Grover-MEGA (1.5B parameters): \$25,000
- ▶ XLNet (BERT variant): \$30,000 — \$60,000 (unclear)
- ▶ This is for a single pre-training run...developing new pre-training techniques may require many runs
- ▶ *Fine-tuning* these models can typically be done with a single GPU

Pre-Training Cost (with Google/AWS)

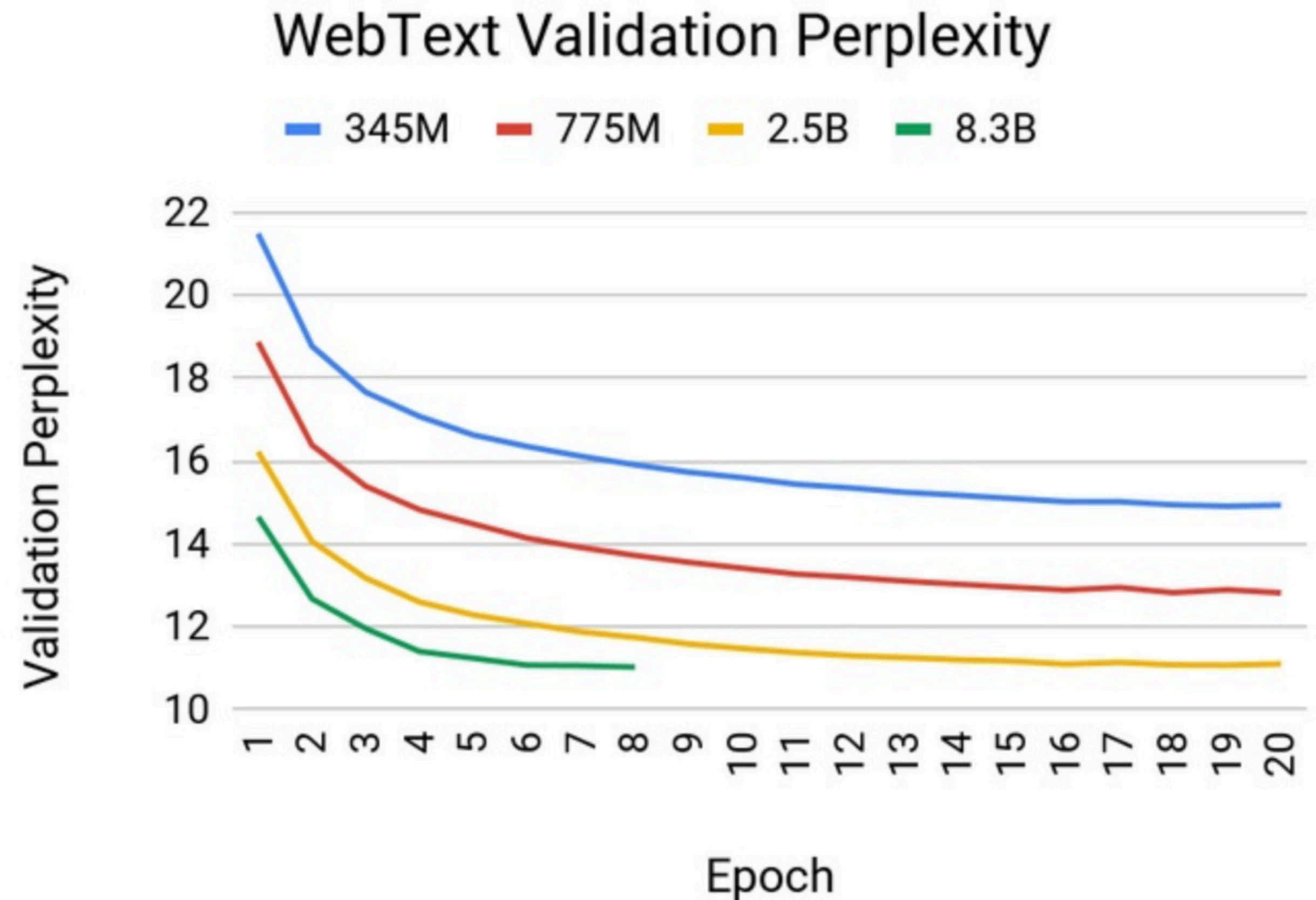
- ▶ GPT-3: estimated to be \$4~10M. This cost has a large carbon footprint
 - ▶ Carbon footprint: equivalent to driving 700,000 km by car (source: Anthropocene magazine)
 - ▶ (Counterpoints: GPT-3 isn't trained frequently, equivalent to 100 people traveling 7000 km for a conference, can use renewables)
- ▶ BERT-base pre-training: carbon emissions roughly on the same order as a single passenger on a flight from NY to San Francisco

Strubell et al. (2019)

<https://lambdalabs.com/blog/demystifying-gpt-3/>
<https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/>

GPT-3: Scaling Up

- ▶ Question: what are the scaling limits of large language models?
- ▶ NVIDIA: trained 8.3B parameter GPT model (5.6x the size of GPT-2), showed lower perplexity from this



GPT-3

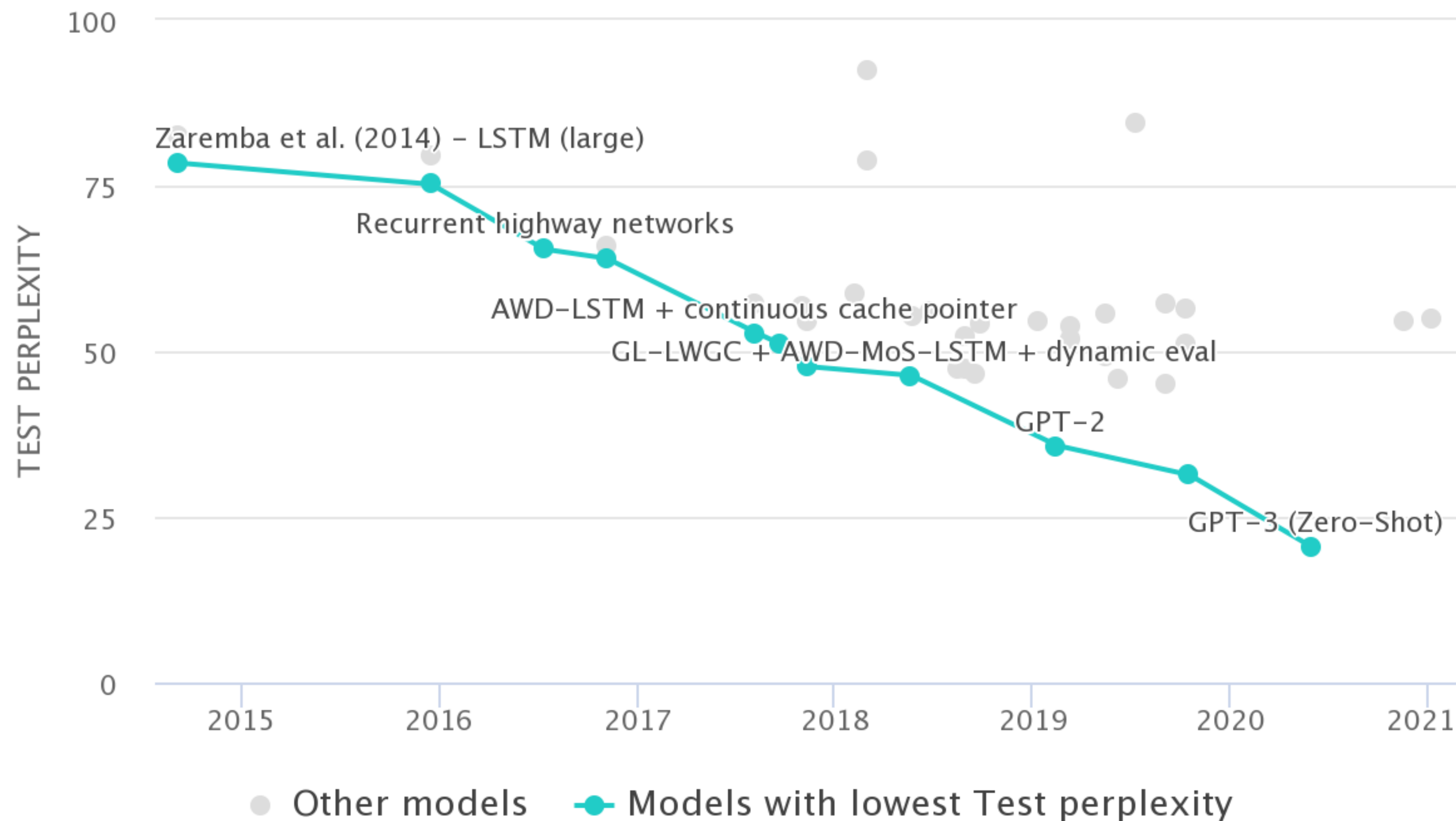
- ▶ GPT-2 but even larger: 1.3B (GPT2) —> 175B (GPT3) parameter models
- ▶ Sparse factorizations of the attention matrix to reduce computing time and memory use. context window is set to 2048 tokens.
- ▶ Data: filtered Common Crawl (410B tokens downsampled x0.44) + WebText dataset (19B x2.9) + two Internet-based book corpora (12Bx1.9, 55Bx0.43) + English Wiki (3B upsampled x3.4)

<https://twitter.com/cocoweixu/status/1285727605568811011>

Brown et al. (2020)

LM Evaluation - Perplexity

- ▶ The perplexity of modern language models have consistently been going down.



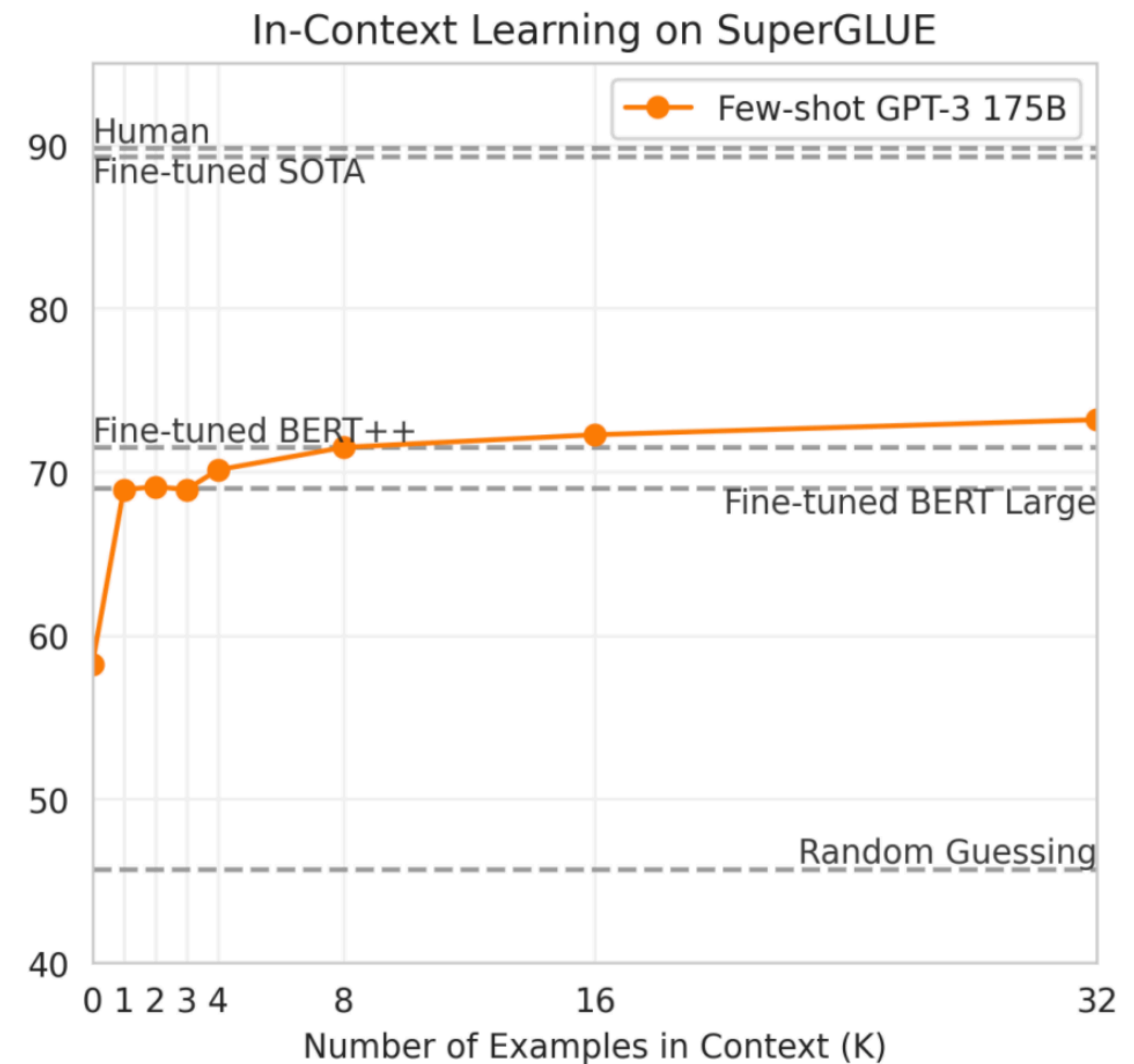
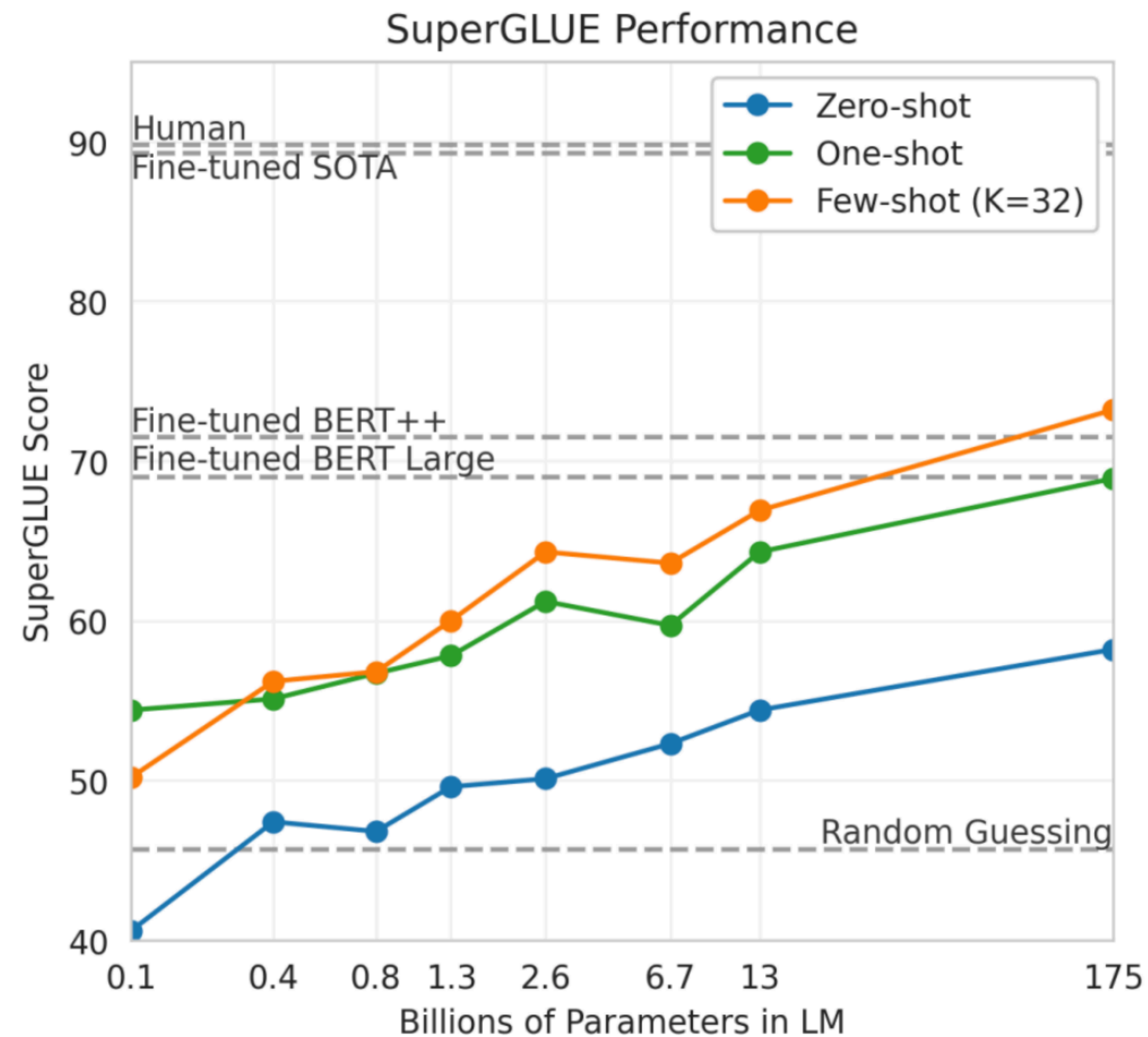
LM Evaluation - Perplexity

- ▶ Accuracy doesn't make sense — predicting the next word is generally impossible so accuracy values would be very low
- ▶ Evaluate LMs on the likelihood of held-out data (averaged to normalize for length)

$$\frac{1}{n} \sum_{i=1}^n \log P(w_i | w_1, \dots, w_{i-1})$$

- ▶ Perplexity: $\exp(\text{average negative log likelihood})$. Lower is better
 - ▶ Suppose we have probs $1/4, 1/3, 1/4, 1/3$ for 4 predictions
 - ▶ Avg NLL (base e) = 1.242 Perplexity = 3.464  geometric mean of denominators

GPT-3 Performance



SuperGLUE benchmark - Wang et al. (2019)

SuperGLUE

Table 2: Development set examples from the tasks in SuperGLUE. **Bold** text represents part of the example format for each task. Text in *italics* is part of the model input. Underlined text is specially marked in the input. Text in a monospaced font represents the expected model output.

BoolQ	Passage: <i>Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.</i> Question: <i>is barq's root beer a pepsi product</i> Answer: No
CB	Text: <i>B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?</i> Hypothesis: <i>they are setting a trend</i> Entailment: Unknown
COPA	Premise: <i>My body cast a shadow over the grass.</i> Question: <i>What's the CAUSE for this?</i> Alternative 1: <i>The sun was rising.</i> Alternative 2: <i>The grass was cut.</i> Correct Alternative: 1
MultiRC	Paragraph: <i>Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week</i> Question: <i>Did Susan's sick friend recover?</i> Candidate answers: <i>Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)</i>
ReCoRD	Paragraph: <i>(CNN) <u>Puerto Rico</u> on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the <u>US</u> commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the <u>State Electoral Commission</u> show. It was the fifth such vote on statehood. "Today, we the people of <u>Puerto Rico</u> are sending a strong and clear message to the <u>US Congress</u> ... and to the world ... claiming our equal rights as <u>American</u> citizens, <u>Puerto Rico</u> Gov. <u>Ricardo Rossello</u> said in a news release. @highlight <u>Puerto Rico</u> voted Sunday in favor of <u>US</u> statehood</i> Query For one, they can truthfully say, "Don't blame me, I didn't vote for them," when discussing the <placeholder> presidency Correct Entities: US

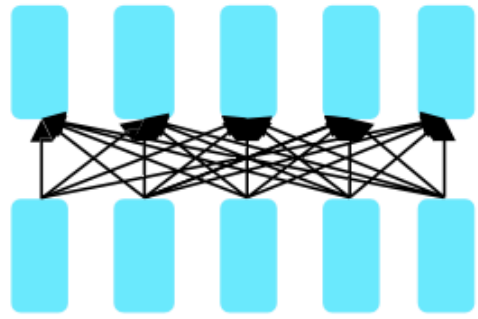
Autoregressive LM

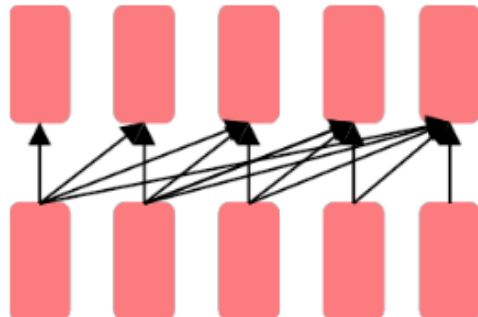
Wei Xu

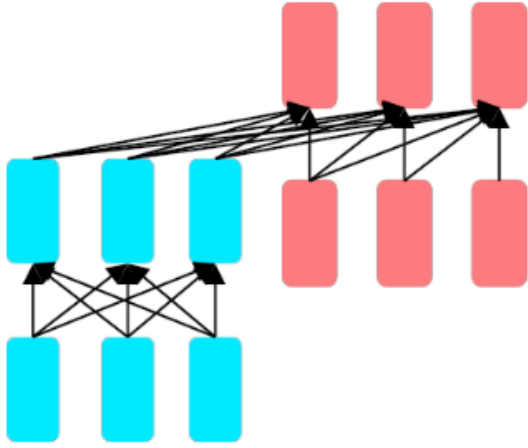
(some slides from Greg Durrett)

LM Pre-training

- ▶ Three types of neural architectures for LLM pre-training

- ▶ Encoder  bidirectional → masked token prediction
(e.g., ELMo, BERT)

- ▶ Decoder  unidirectional (autoregressive)
→ next-token prediction
(e.g. GPT, Llama)

- ▶ Encoder-Decoder  encode bidirectionally
→ generate autoregressively
(e.g., BART, T5)

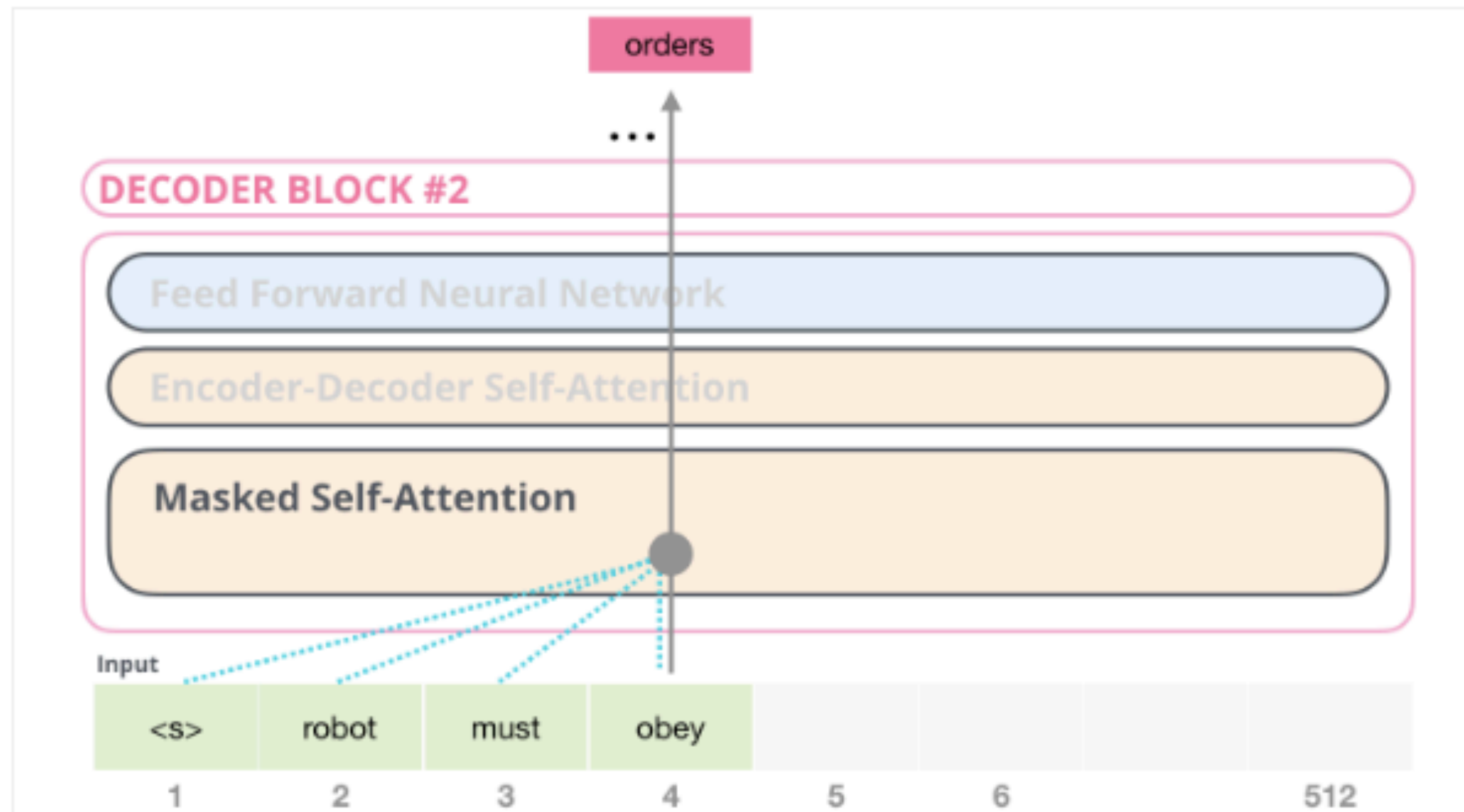
This Lecture

- ▶ Decoder-style LMs: GPT / GPT-2 / GPT-3
- ▶ In-context Learning
- ▶ Scaling Laws

GPT/GPT2/GPT3

OpenAI GPT/GPT2

- ▶ Train a single unidirectional transformer LM on long contexts
- ▶ Masked self-attention: each token can only attend to past tokens



Radford et al. (2019)

OpenAI GPT/GPT2

- ▶ Train a single unidirectional transformer LM on long contexts
- ▶ Masked self-attention: each token can only attend to past tokens

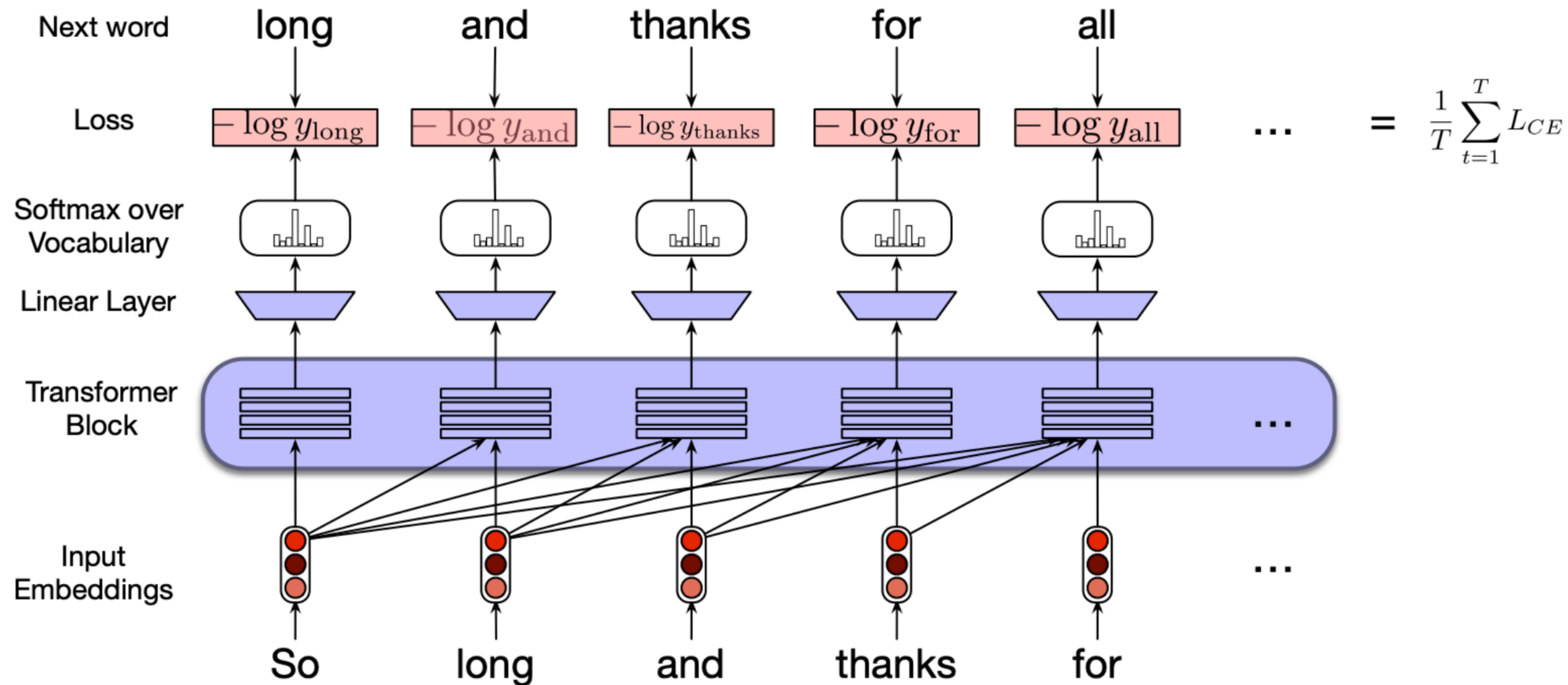


Image Credit: Jurafsky & Martin

OpenAI GPT/GPT2

- ▶ GPT2: trained on 40GB of text collected from upvoted links from reddit
- ▶ 1.5B parameters — the largest of these models trained as of March 2019

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

- ▶ Because it's a decoder-style language model, we can **generate** from it

GPT-2

- Some results from GPT-2 foreshadowed the benefits of scaling up language models (shown here for Winograd and LM Perplexity)

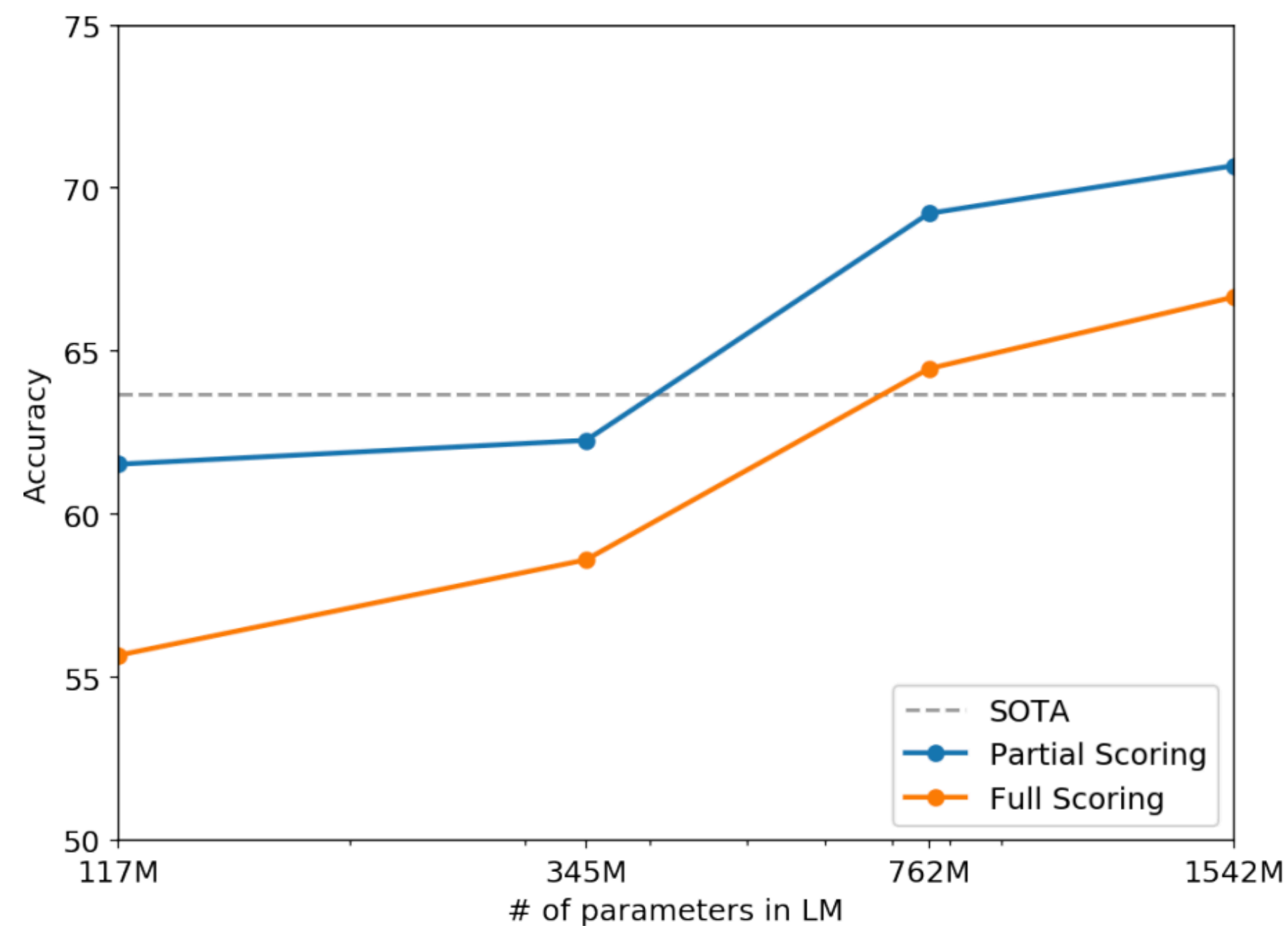


Figure 3. Performance on the Winograd Schema Challenge as a function of model capacity.

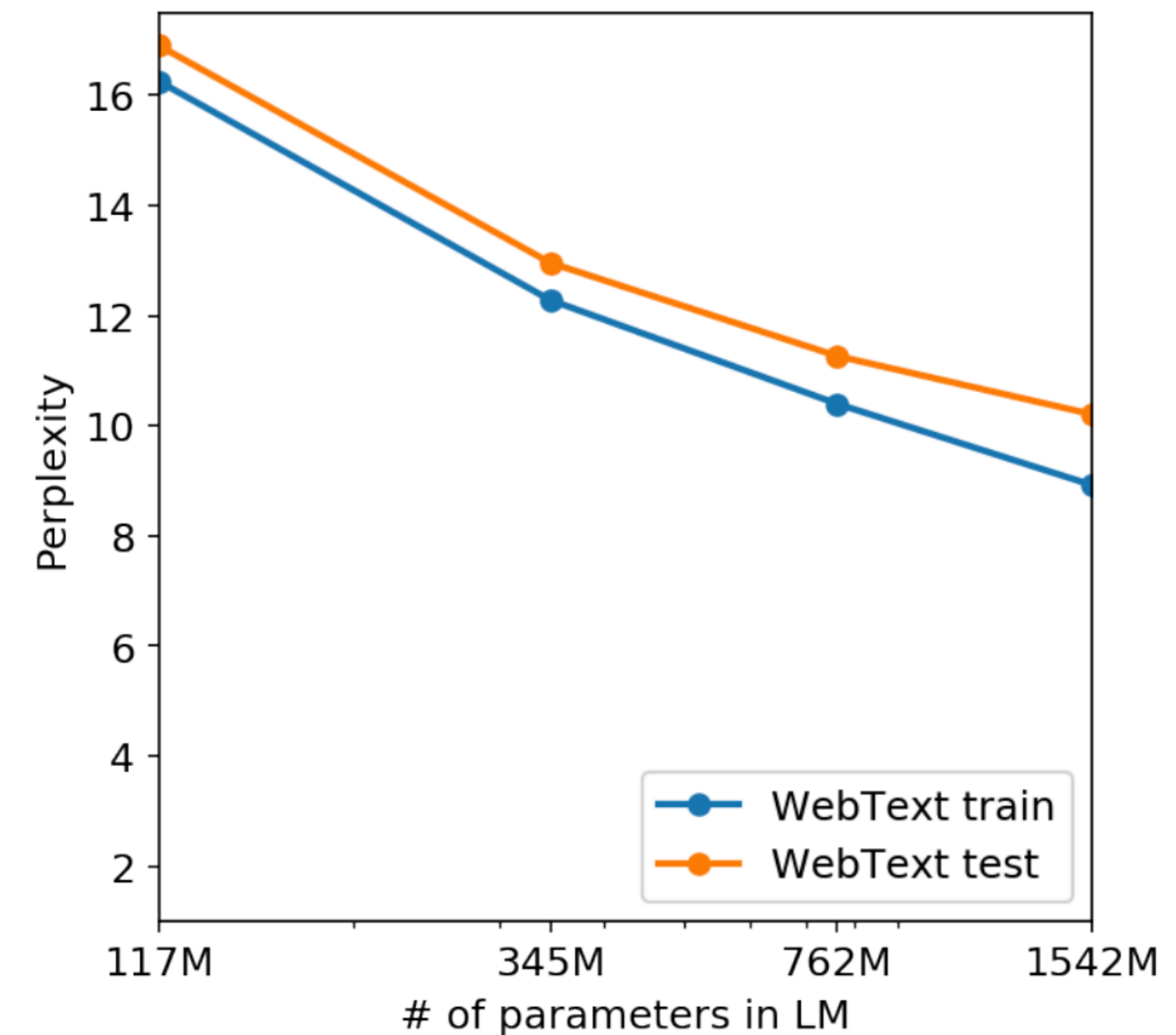
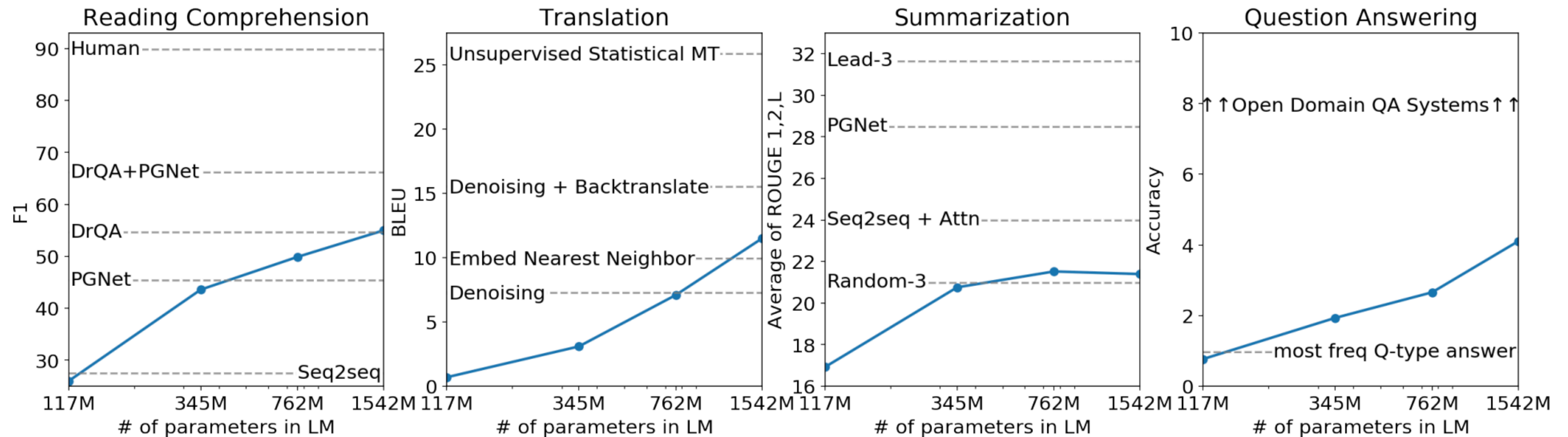


Figure 4. The performance of LMs trained on WebText as a function of model size.

GPT-2

- Started to show strong zero-shot performance, even though “in terms of practical applications, the zero-shot performance of GPT-2 is still far from use-able.”



Radford et al. (2019)

OpenAI GPT2

System Prompt
(human-written)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Model Completion
(machine-written, 10
tries)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move

slide credit:
OpenAI

OpenAI GPT2

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

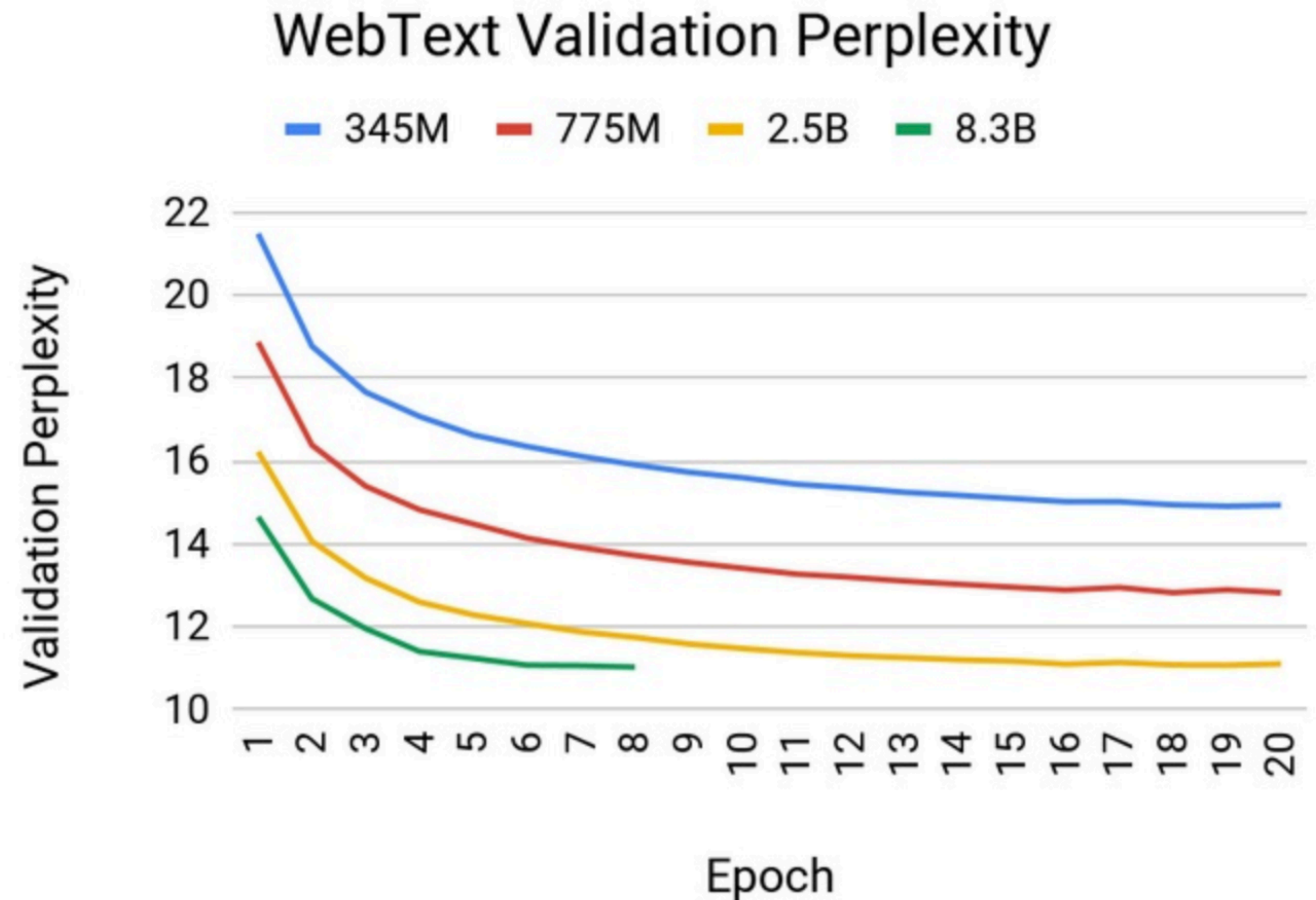
The singer was also wearing a pair of black-rimmed glasses, a black jacket, black jeans and black sandals.

She was carrying a pair of black and white striped gloves and a small black bag.

slide credit:
OpenAI

GPT-3: Scaling Up

- ▶ Question: what are the scaling limits of large language models?
- ▶ NVIDIA: trained 8.3B parameter GPT model (5.6x the size of GPT-2), showed lower perplexity from this



GPT-3

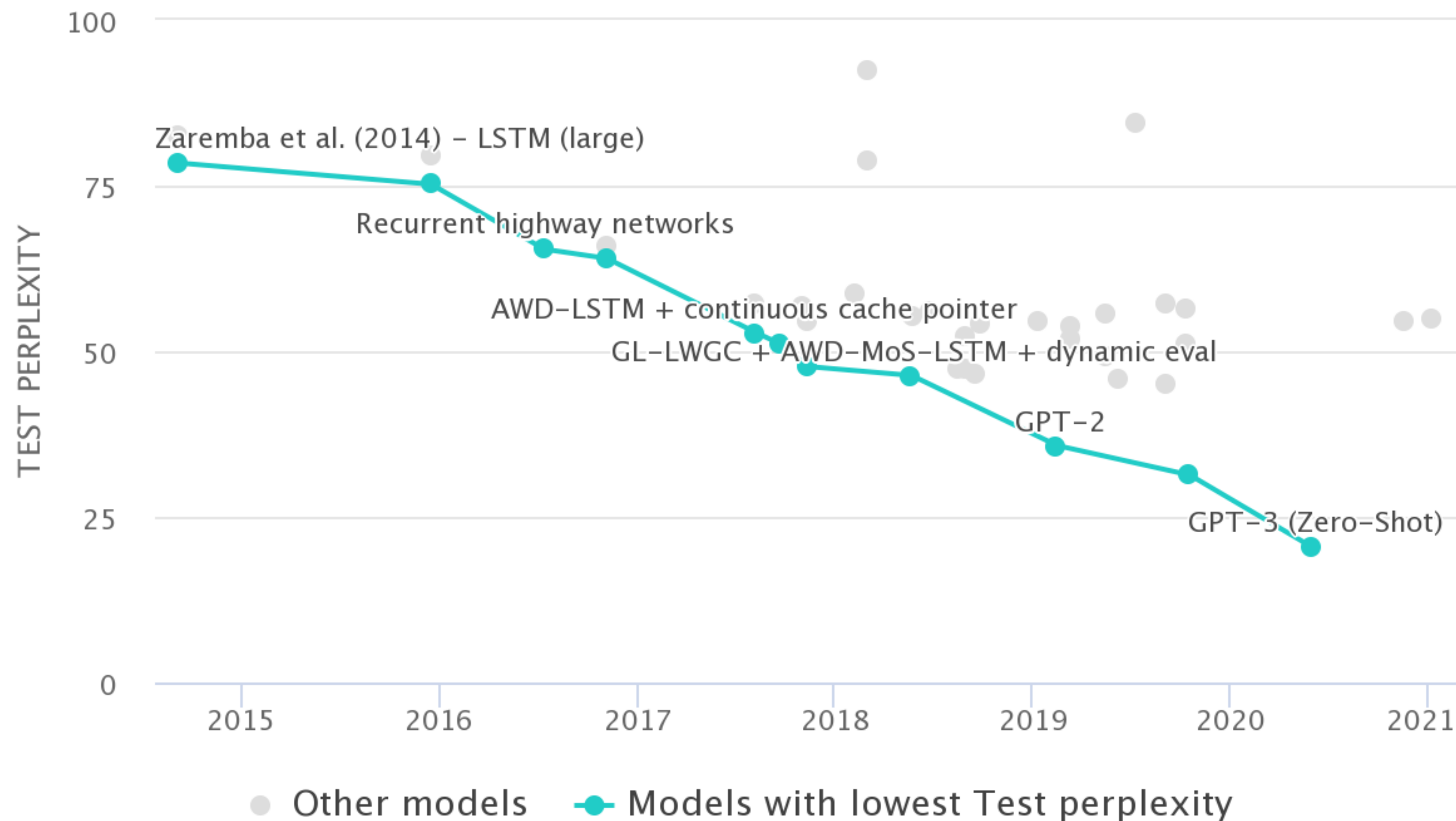
- ▶ GPT-2 but even larger: 1.3B (GPT2) —> 175B (GPT3) parameter models
- ▶ Sparse factorizations of the attention matrix to reduce computing time and memory use. context window is set to 2048 tokens.
- ▶ Data: filtered Common Crawl (410B tokens downsampled x0.44) + WebText dataset (19B x2.9) + two Internet-based book corpora (12Bx1.9, 55Bx0.43) + English Wiki (3B upsampled x3.4)

<https://twitter.com/cocoweixu/status/1285727605568811011>

Brown et al. (2020)

LM Evaluation - Perplexity

- ▶ The perplexity of modern language models have consistently been going down.



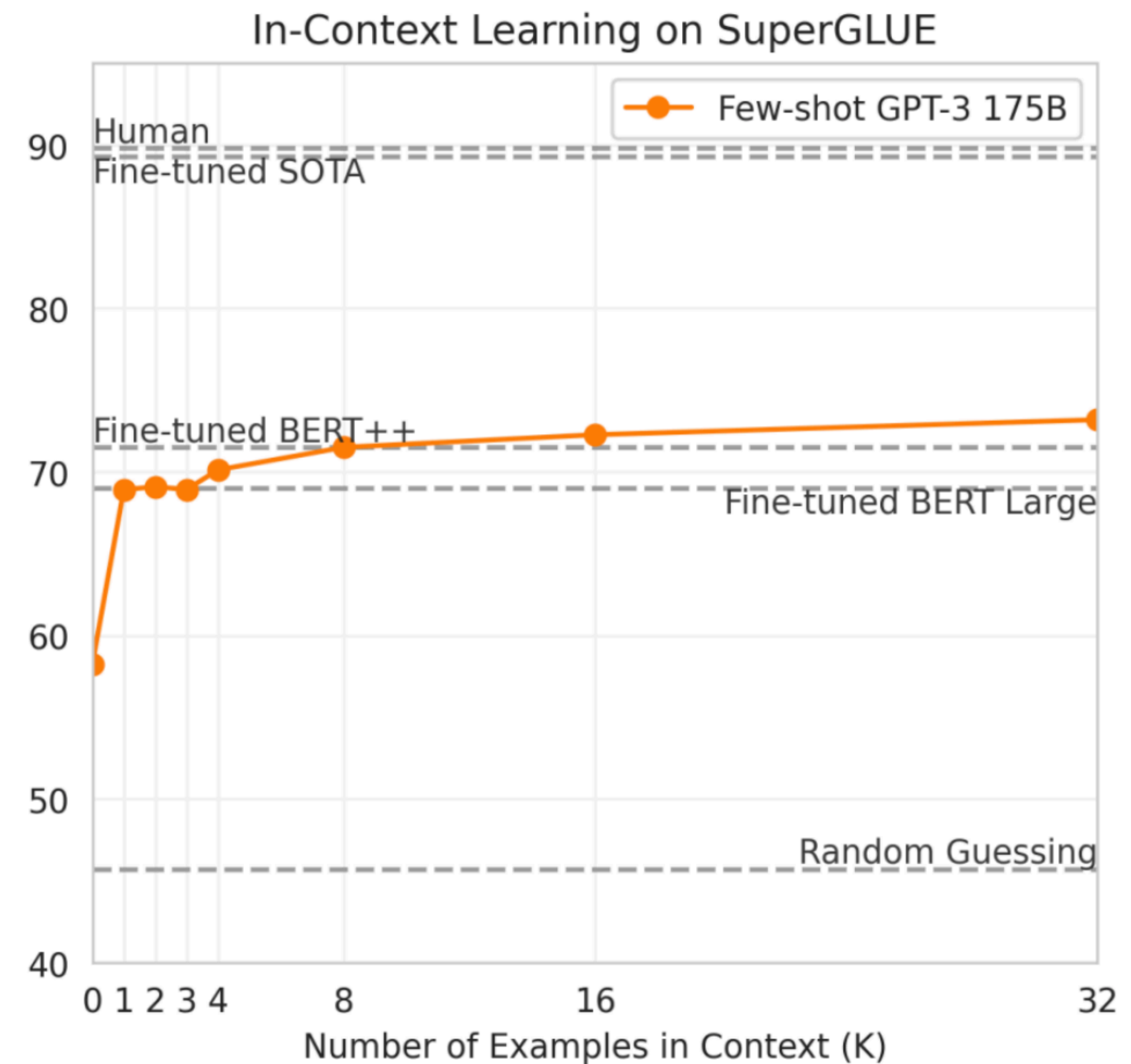
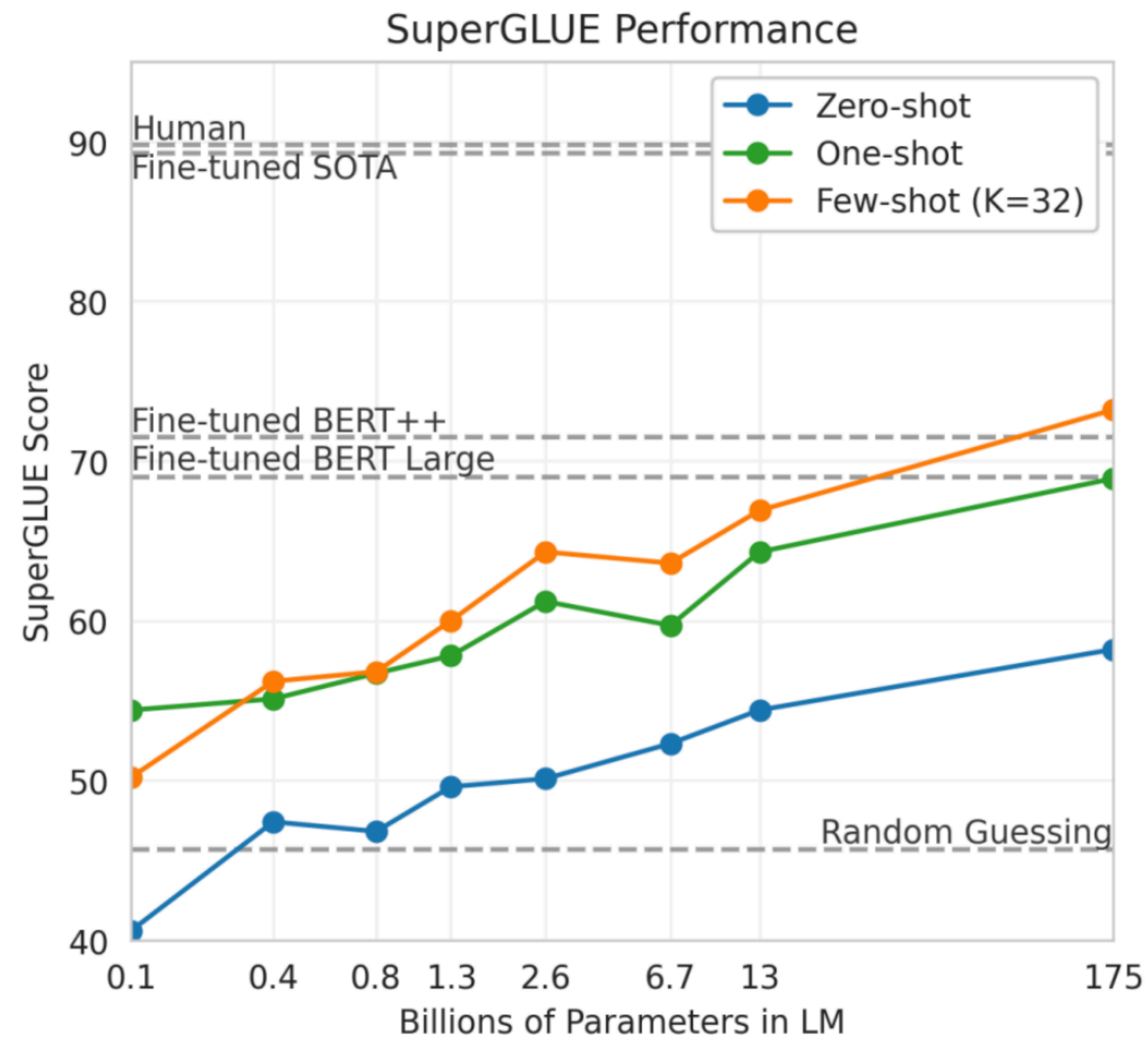
LM Evaluation - Perplexity

- ▶ Accuracy doesn't make sense — predicting the next word is generally impossible so accuracy values would be very low
- ▶ Evaluate LMs on the likelihood of held-out data (averaged to normalize for length)

$$\frac{1}{n} \sum_{i=1}^n \log P(w_i | w_1, \dots, w_{i-1})$$

- ▶ Perplexity: $\exp(\text{average negative log likelihood})$. Lower is better
 - ▶ Suppose we have probs $1/4, 1/3, 1/4, 1/3$ for 4 predictions
 - ▶ Avg NLL (base e) = 1.242 Perplexity = 3.464  geometric mean of denominators

GPT-3 Performance



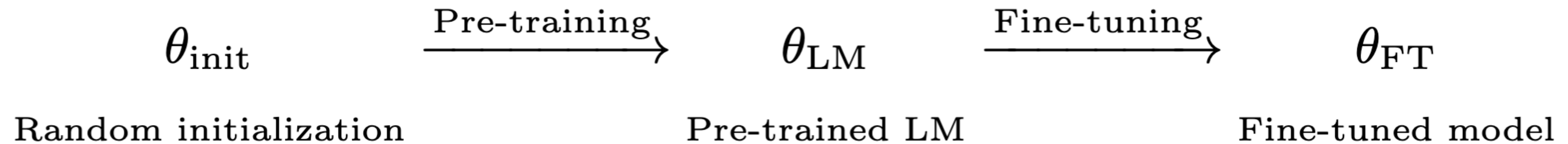
SuperGLUE benchmark - Wang et al. (2019)

SuperGLUE

Table 2: Development set examples from the tasks in SuperGLUE. **Bold** text represents part of the example format for each task. Text in *italics* is part of the model input. Underlined text is specially marked in the input. Text in a monospaced font represents the expected model output.

BoolQ	Passage: <i>Barq's – Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq's Famous Olde Tyme Root Beer until 2012.</i> Question: <i>is barq's root beer a pepsi product</i> Answer: No
CB	Text: <i>B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?</i> Hypothesis: <i>they are setting a trend</i> Entailment: Unknown
COPA	Premise: <i>My body cast a shadow over the grass.</i> Question: <i>What's the CAUSE for this?</i> Alternative 1: <i>The sun was rising.</i> Alternative 2: <i>The grass was cut.</i> Correct Alternative: 1
MultiRC	Paragraph: <i>Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week</i> Question: <i>Did Susan's sick friend recover?</i> Candidate answers: <i>Yes, she recovered (T), No (F), Yes (T), No, she didn't recover (F), Yes, she was at Susan's party (T)</i>
ReCoRD	Paragraph: <i>(CNN) <u>Puerto Rico</u> on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the <u>US</u> commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the <u>State Electoral Commission</u> show. It was the fifth such vote on statehood. "Today, we the people of <u>Puerto Rico</u> are sending a strong and clear message to the <u>US Congress</u> ... and to the world ... claiming our equal rights as <u>American</u> citizens, <u>Puerto Rico</u> Gov. <u>Ricardo Rossello</u> said in a news release. @highlight <u>Puerto Rico</u> voted Sunday in favor of <u>US</u> statehood</i> Query For one, they can truthfully say, "Don't blame me, I didn't vote for them," when discussing the <placeholder> presidency Correct Entities: US

(Recap) Pre-training —> Fine-tuning Pipeline

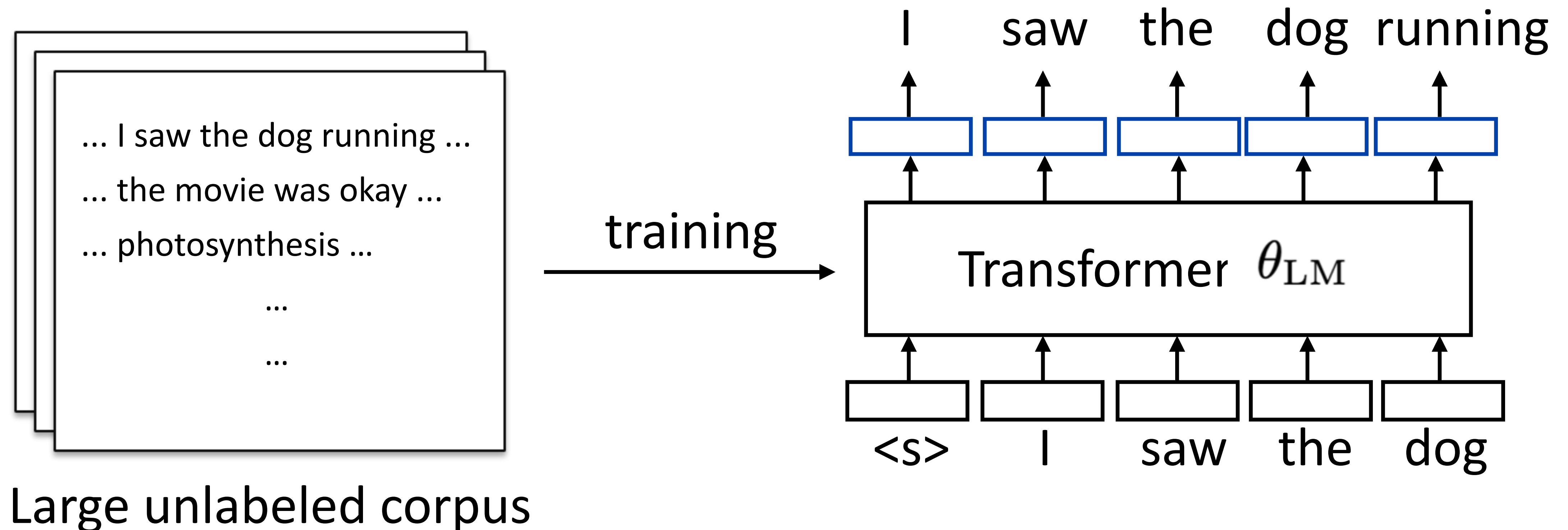


Pretraining: Next-Token Prediction

- Autoregressive Language Models, i.e., next-token prediction:

$$\mathcal{L}_{\text{NTP}}(\theta) = - \sum_{t=1}^T \log p_{\theta}(x_t \mid \mathbf{x}_{<t})$$

$x_1, x_2, \dots, x_{t-1}$

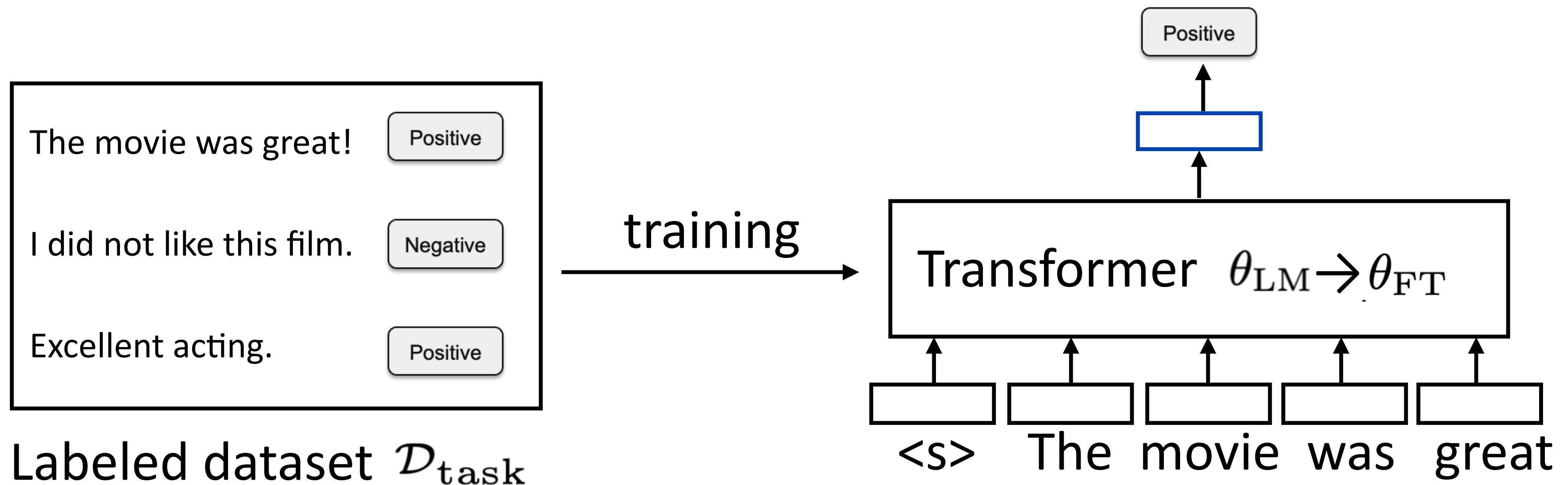


Fine-tuning: Supervised Learning

- Initialize with pretrained LM weights; continue training on a labeled dataset (Compute loss $\rightarrow$ Backpropagation $\rightarrow$ Update model weights).

$$\mathcal{L}_{\text{finetune}}(\theta) = - \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{\text{task}}} \log p_{\theta}(\mathbf{y} \mid \mathbf{x})$$

initialize from θ_{LM}



Prompting

GPT-3: A Paradigm Shift!

- ▶ Until GPT-3, fine-tuning on task-specific labeled datasets was the standard way to use pre-trained models (e.g., BERT, T5, GPT-2) for downstream tasks (e.g., text classification, NER, etc.)
- ▶ Fine-tuning requires computing gradients and updating model parameters — computationally expensive for a 175B GPT-3 model

Fine-tuning

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



Brown et al. (2020)

GPT-3: Few-shot Prompting

- ▶ Model is frozen. No parameter updates.
(i.e., only a forward pass, no backward pass)
- ▶ Prompt: task instruction plus a few demonstrations (training examples)
- ▶ This is also called: **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:
2  sea otter => loutre de mer
3  peppermint => menthe poivrée
4  plush girafe => girafe peluche
5  cheese => .....
```

← task description

← examples

← prompt

Brown et al. (2020)

GPT-3: Few-shot Prompting

- ▶ Model is frozen and is given a few demonstrations.

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

Circulation revenue has increased by 5% in Finland. // Finance

They defeated ... in the NFC Championship Game. // Sports

Apple ... development of in-house chips. // Tech

The company anticipated its operating profit to improve. // _____

- ▶ “in-context learning” - unlike conventional machine learning in that there’s no optimization of any parameters.
- ▶ Model “learns” by conditioning on a few examples of the task.

Brown et al. (2020), Schick et al. (2020), Schick and Schütze (2021)

Prompting

- ▶ Zero-shot Prompt:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} p_{\theta_{\text{LM}}}(\mathbf{y} \mid \mathbf{i}, \mathbf{x})$$

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: ← i, task description
2 cheese => ← x
```

Prompting

- ▶ Few-shot Prompt:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} p_{\theta_{\text{LM}}} \left(\mathbf{y} \mid \mathbf{i}, \underbrace{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_K, \mathbf{y}_K)}_{\text{Few-shot examples}}, \mathbf{x} \right)$$

- ▶ Zero-shot Prompt:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} p_{\theta_{\text{LM}}} (\mathbf{y} \mid \mathbf{i}, \mathbf{x})$$

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: ← i, task description
2 cheese =>                     ← x
```

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: ← i, task description
2 sea otter => loutre de mer    ← examples
3 peppermint => menthe poivrée ← (x1, y1)
4 plush girafe => girafe peluche ← ⋮
5 cheese => .....             ← x
```

Prompting: Inference Without Training

► Few-shot Prompt:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} p_{\theta_{\text{LM}}} \left(\mathbf{y} \mid \mathbf{i}, \underbrace{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_K, \mathbf{y}_K)}_{\text{Few-shot examples}}, \mathbf{x} \right)$$

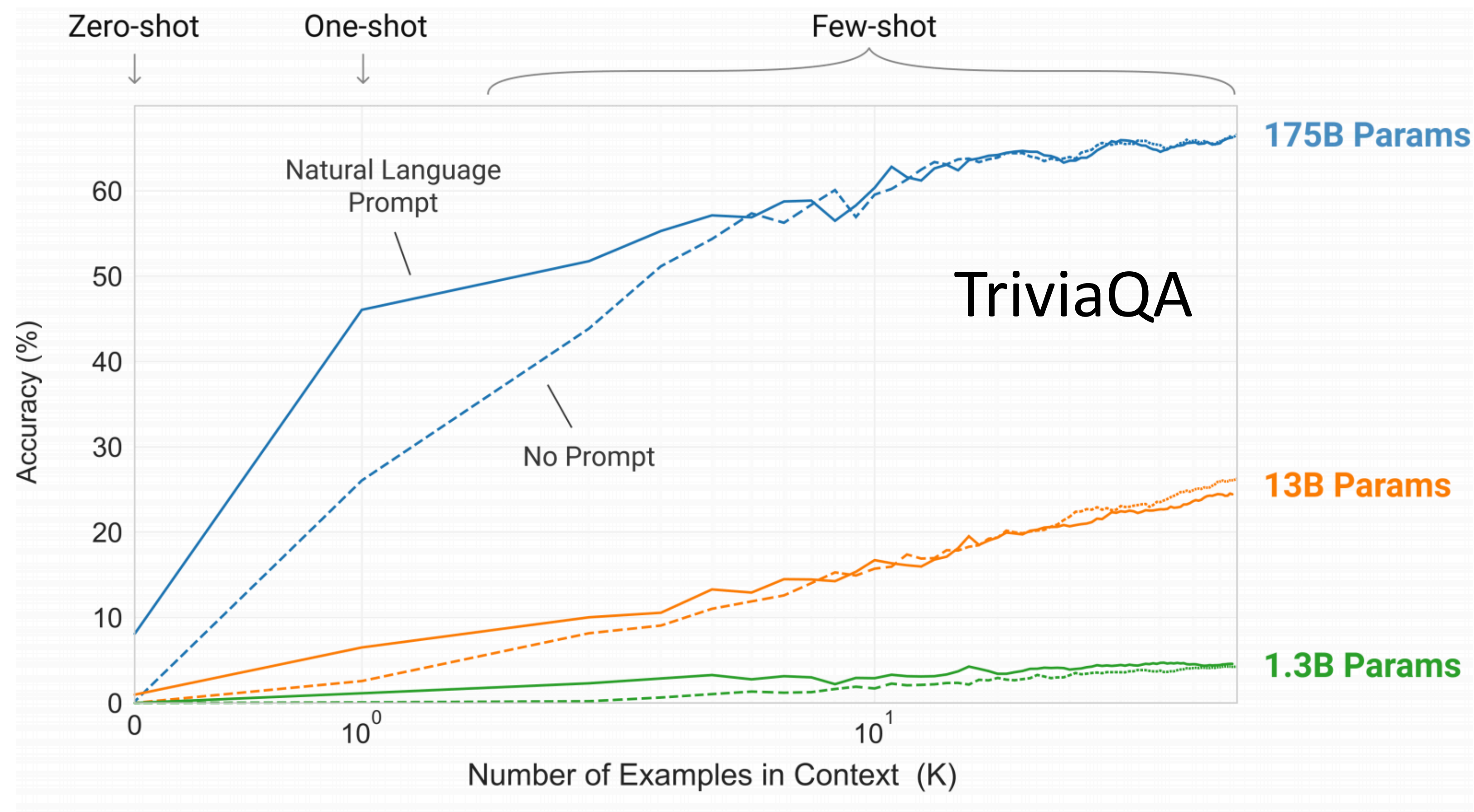
► Zero-shot Prompt:

$$\hat{\mathbf{y}} = \arg \max_{\mathbf{y}} p_{\theta_{\text{LM}}} (\mathbf{y} \mid \mathbf{i}, \mathbf{x})$$

- A pretrained LM performs new tasks by conditioning on instructions, (optional) examples, input text.
- No backward pass. No gradients. No parameter updates (parameters of LM are frozen).

GPT-3: Few-shot Learning

- ▶ **Key observation:** few-shot learning only works with the very largest models!
- ▶ **Later (2022-present):** smaller models also can do this now, by training longer on larger data, instruction-tuning, and post-training,



Brown et al. (2020)

TriviaQA

Context	→	Q: 'Nude Descending A Staircase' is perhaps the most famous painting by which 20th century artist?
		A:
Target Completion	→	MARCEL DUCHAMP
Target Completion	→	r mutt
Target Completion	→	Duchamp
Target Completion	→	marcel duchamp
Target Completion	→	R.Mutt
Target Completion	→	Marcel duChamp
Target Completion	→	Henri-Robert-Marcel Duchamp
Target Completion	→	Marcel du Champ
Target Completion	→	henri robert marcel duchamp
Target Completion	→	Duchampian
Target Completion	→	Duchamp
Target Completion	→	duchampian
Target Completion	→	marcel du champ
Target Completion	→	Marcel Duchamp
Target Completion	→	MARCEL DUCHAMP

Figure G.34: Formatted dataset example for TriviaQA. TriviaQA allows for multiple valid completions.

GPT-3

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	89.0	91.0	96.9	93.9	94.8	92.5
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	76.1	93.8	62.3	88.2	92.5	93.3
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

- ▶ Sometimes very impressive, (MultiRC, ReCoRD), sometimes very bad
- ▶ Results on other datasets are equally mixed — but still strong for a few-shot model!

MultiRC (multi-sentence)

Sent 1: The hijackers attacked at 9:28.

Sent 2: While traveling 35,000 feet above eastern Ohio, United 93 suddenly dropped 700 feet.

Sent 3: Eleven seconds into the descent, the FAA's air traffic control center in Cleveland received the first of two radio transmissions from the aircraft.

Sent 4: During the first broadcast, the captain or first officer could be heard declaring "Mayday" amid the sounds of a physical struggle in the cockpit.

Sent 5: The second radio transmission, 35 seconds later, indicated that the fight was continuing.

Sent 6: The captain or first officer could be heard shouting: "Hey get out of here-get out of here-get out of here."

Sent 7: On the morning of 9/11, there were only 37 passengers on United 93-33 in addition to the 4 hijackers.

Sent 8: This was below the norm for Tuesday mornings during the summer of 2001.

Sent 9: But there is no evidence that the hijackers manipulated passenger levels or purchased additional seats to facilitate their operation.

Sent 10: The terrorists who hijacked three other commercial flights on 9/11 operated in five-man teams.

Sent 11: They initiated their cockpit takeover within 30 minutes of takeoff.

Sent 12: On Flight 93, however, the takeover took place 46 minutes after takeoff and there were only four hijackers.

Question: Which two factors were different between the three other hijacked planes and United 93?

the day of the takeover

A) The amount of time that passed before the takeover started

B)* United 93 took longer and had less hijackers

C) The airline operating the planes

D) The weather and fuel used by the airplane

E) The navigation system used by the planes

Reasoning needed: Discourse relation (contrast)

One needs to identify that the discourse marker *however* in Sent 12 indicates a contrast relation between Flight 93 and the flights mentioned in Sent 10. Also, *only* in Sent 12 indicates that the number of hijackers were fewer than in the contrasted other flights.

Question: What was below average for this particular day?

A) the number of passengers in the first class.

B)* the number of passengers on board.

C) the number of hijackers

D) the amount of air traffic in the skies

E) the temperature

Reasoning needed: Event coreference

One needs to identify that *This* in Sent 8 co-refers to (event of) number of passengers in Sent 7. Note that Sent 12 contains *only four hijackers* and understanding that *only* indicates a smaller number of entities than expected (as in previous question), might mislead a system into believing that (C) is the correct answer.

GPT-3

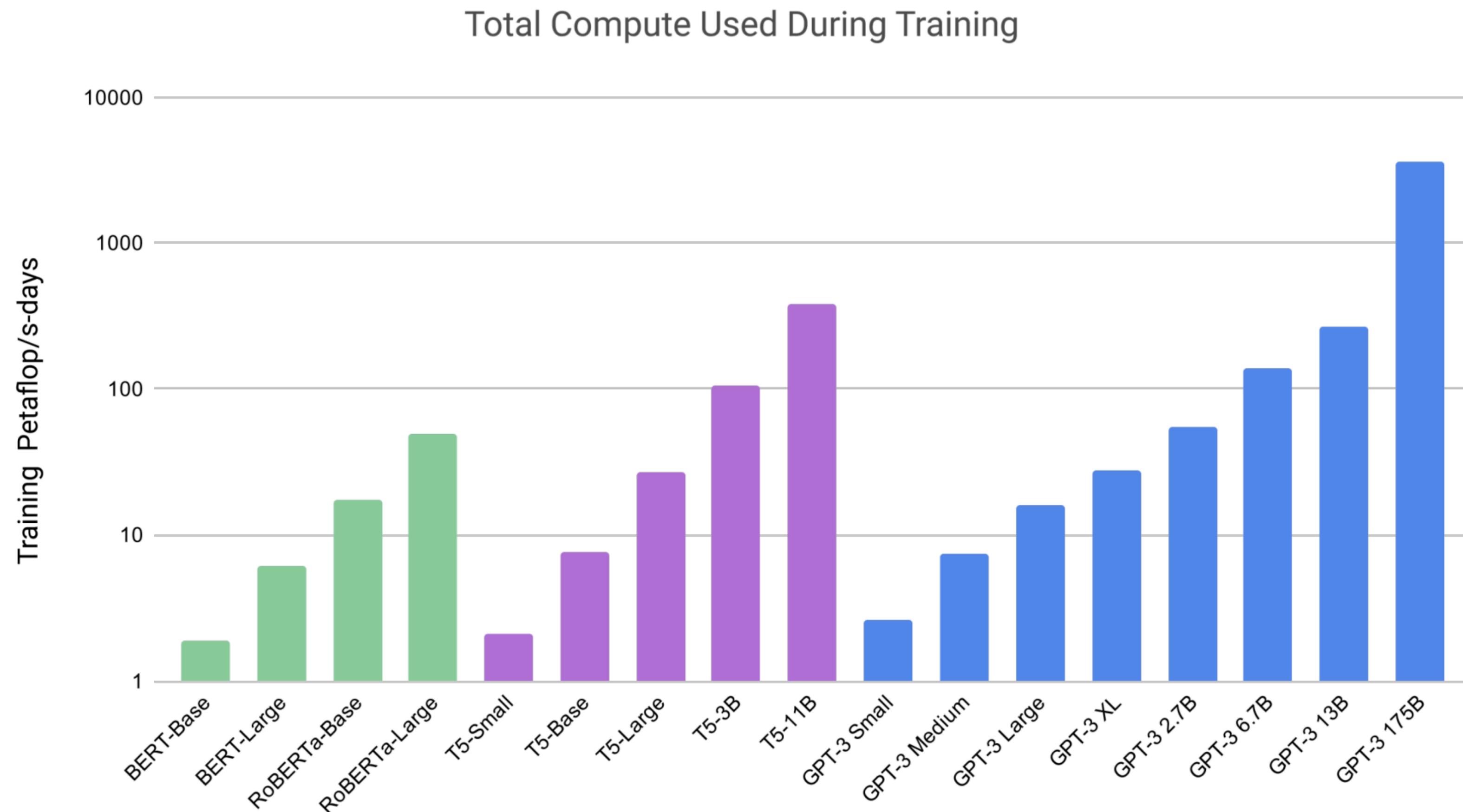
- ▶ GPT-2 but even larger: 1.3B → 175B parameter models

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

- ▶ Trained on 570GB of Common Crawl
 - ▶ 175B parameter model’s parameters alone take >400GB to store (4 bytes per param). Trained in parallel on a “high bandwidth cluster provided by Microsoft”
- Brown et al. (2020)

GPT-3 Pre-training Cost

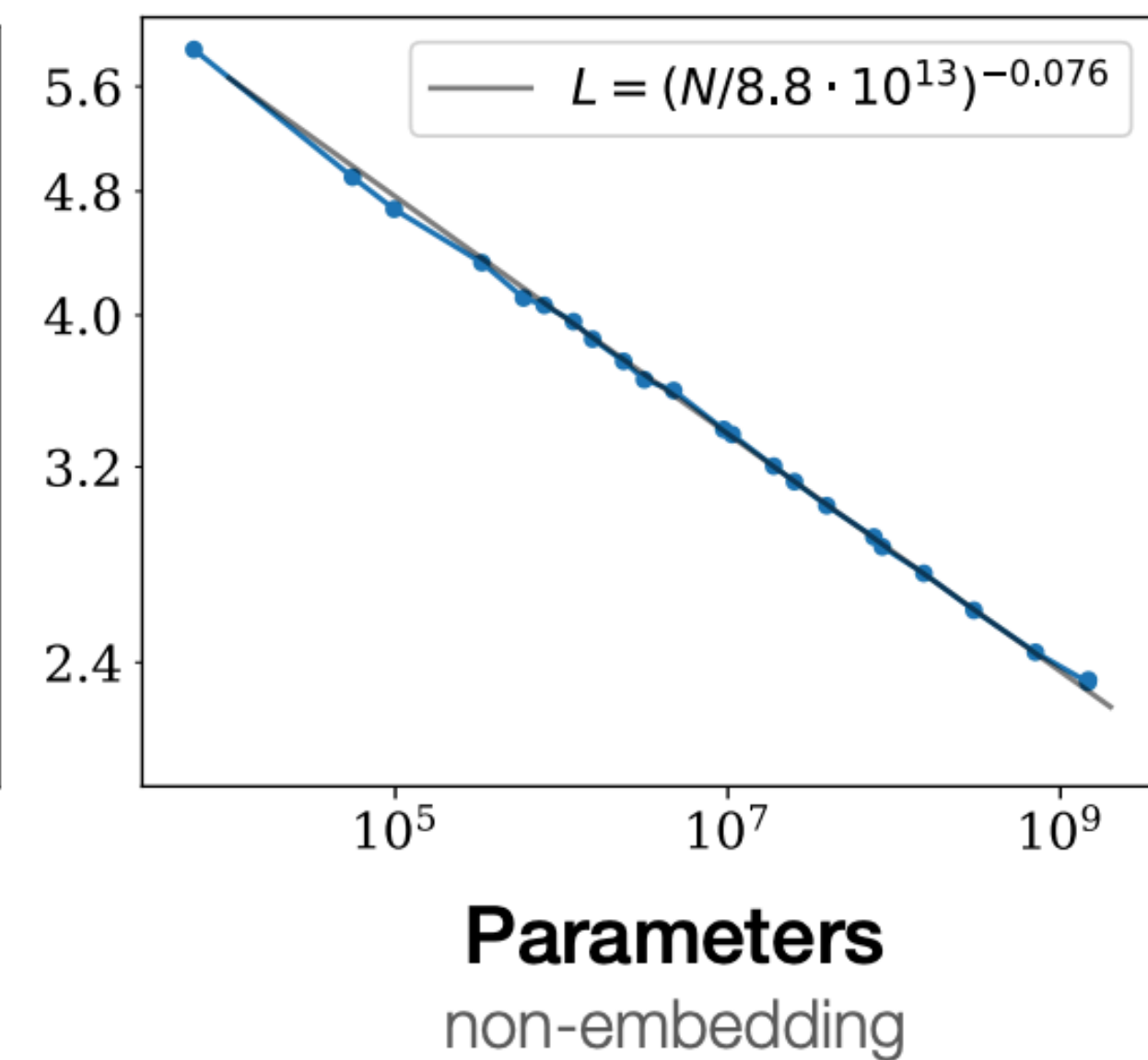
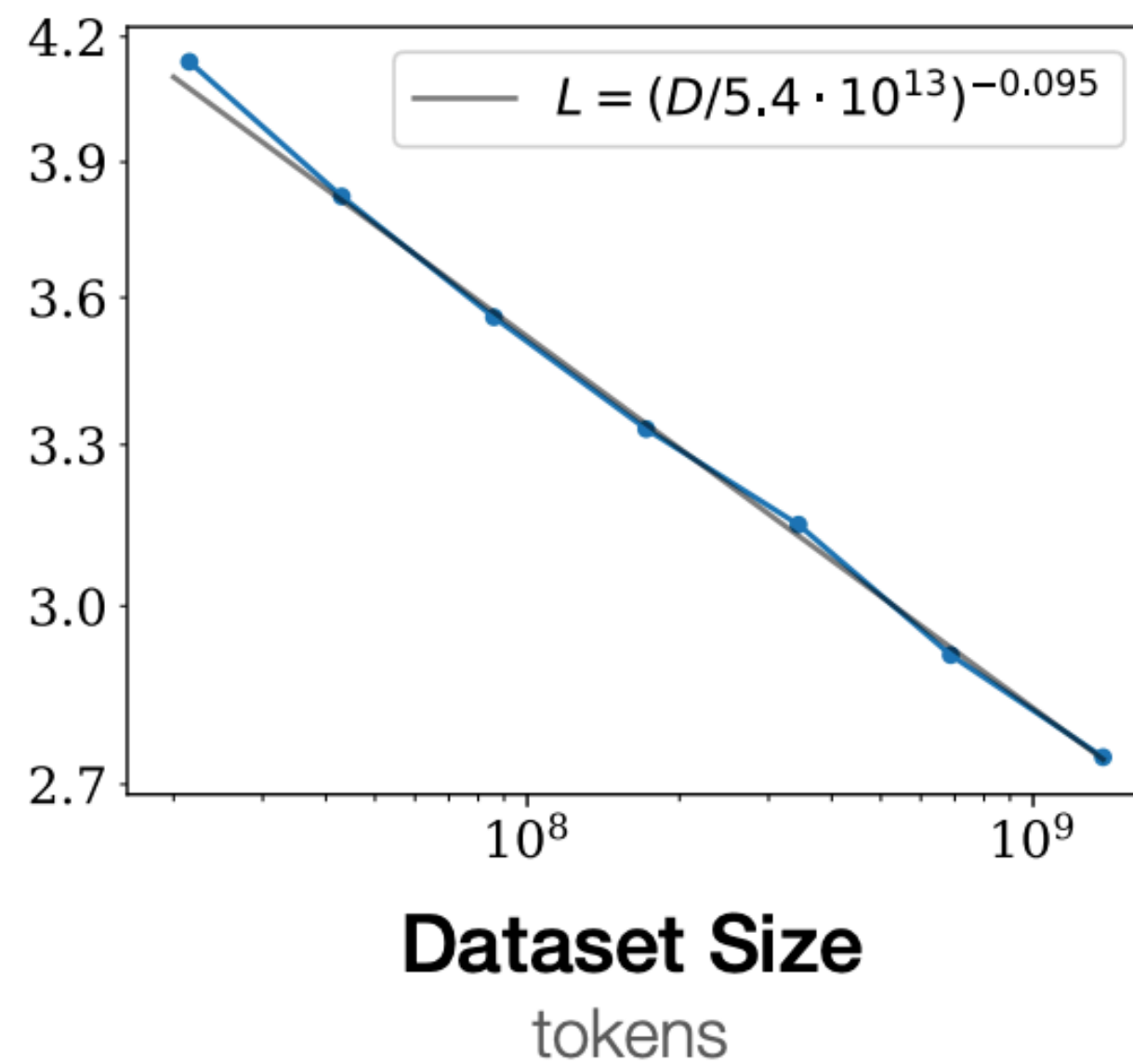
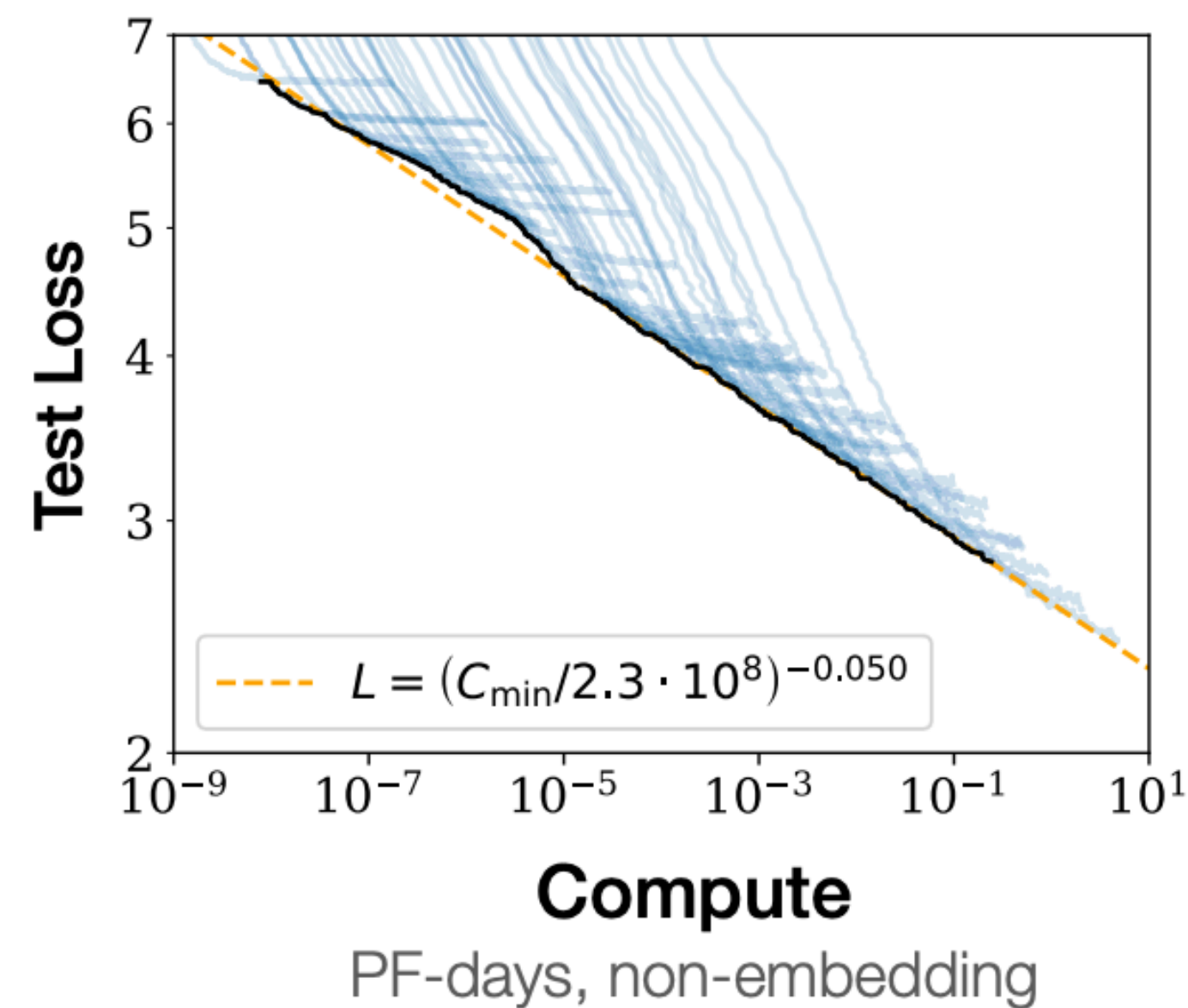
- ▶ Trained on Microsoft Azure, estimated to cost \$4~10M (1000x BERT-large)



1 petaflop/s-day is equivalent to 8 V100 GPUs at full efficiency of a day
Brown et al. (2020)

GPT Scaling Laws

- ▶ With more compute, larger models get further down the loss “frontier”
- ▶ Building a bigger model (increasing compute) will decrease test loss!



petaflop (10²⁰)/s-days

Kaplan et al. (2020)

Scaling Laws

Wei Xu

Compute Cost

- ▶ For a fixed compute budget C , there is a tradeoff between N and D
- ▶ For example, training compute for a Transformer LM is approximately:

$$C \approx 6ND$$

$$(\text{FLOPs} \approx 6 \cdot \text{params} \cdot \text{tokens})$$

- ▶ Model size **N**: number of parameters
- ▶ Training dataset size **D**: number of tokens/examples
- ▶ Training compute **C**: floating point operations (FLOPs)

Training LLMs under Resource Limits

- ▶ Typically, there is:
 - ▶ a limit on compute budget (number of GPUs $\times$ training time)
 - ▶ a limit on the amount of training data available
- ▶ To build a very large language model, we can only afford training once
- ▶ Q: What is the optimal model size (number of parameters) to train?

Scaling Laws

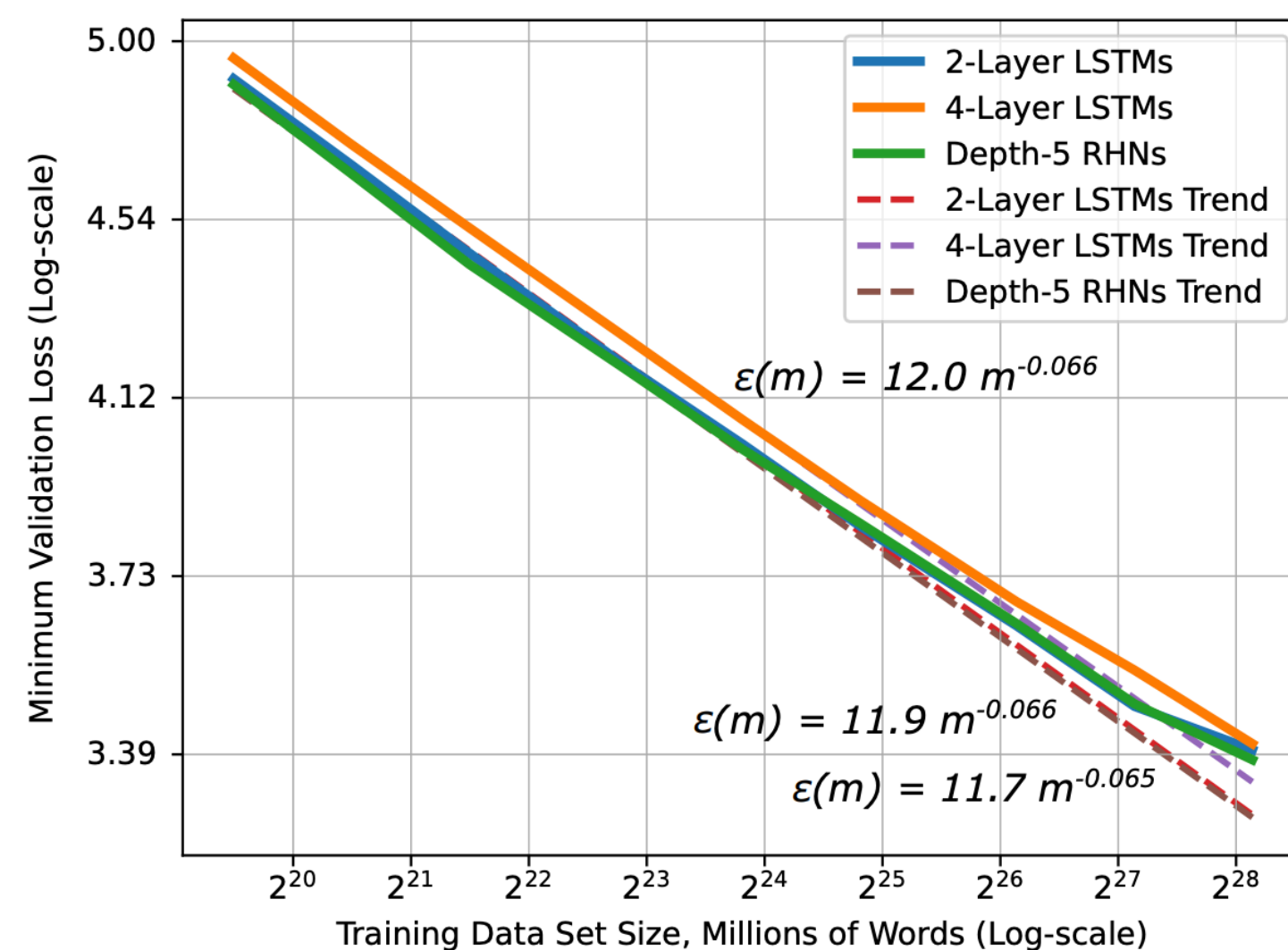
- ▶ How do we allocate the resources more optimally?
- ▶ → How model performance would change as resources increase?

Performance = function (N, D, C)

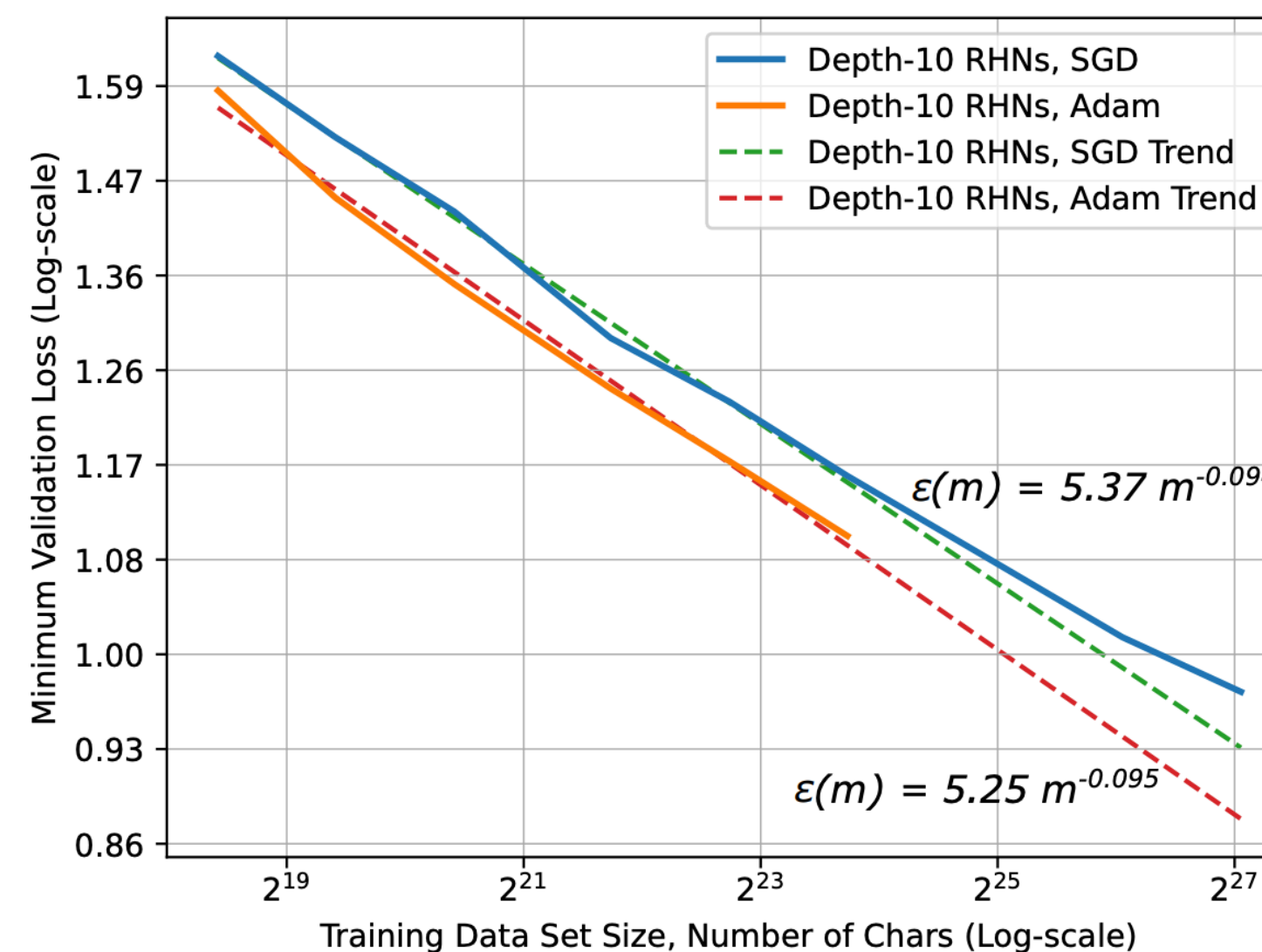
- ▶ Model size **N**: number of parameters
 - ▶ Training dataset size **D**: number of tokens/examples
 - ▶ Training compute **C**: floating point operations (FLOPs)
-
- ▶ Other factors (e.g., model architecture, hyperparameters) fixed/tuned

Baidu Scaling Laws

- ▶ Predates modern LLM scaling laws
- ▶ **Power-law** relationships with dataset and model size
- ▶ Observed across diverse tasks and neural network architectures



Machine Translation
(seq2seq - Luong et al. 2015)



Character Language Models
(RHN - Zilly et al. 2017)

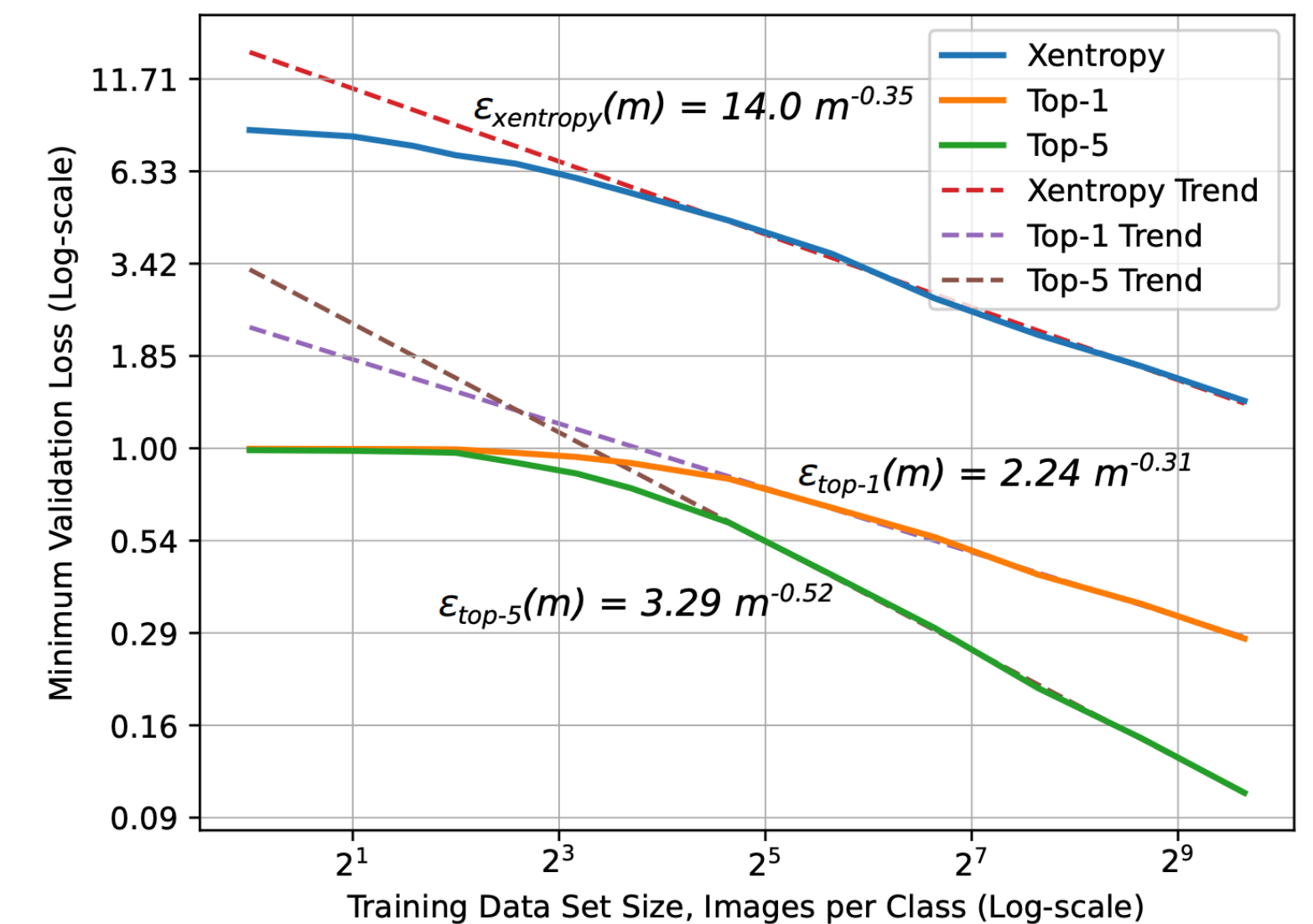
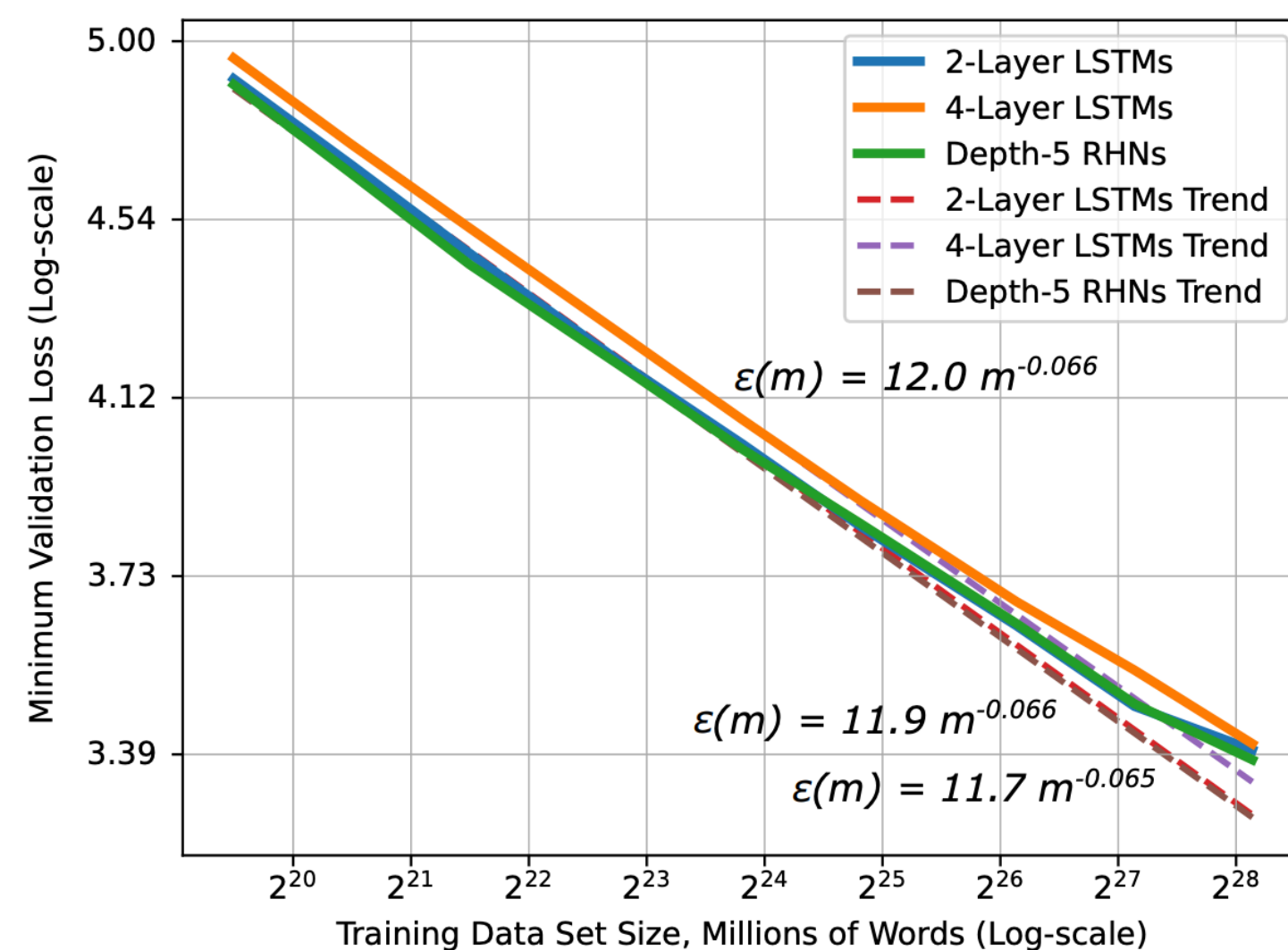


Image Classification
(ResNets - He et al. 2016)

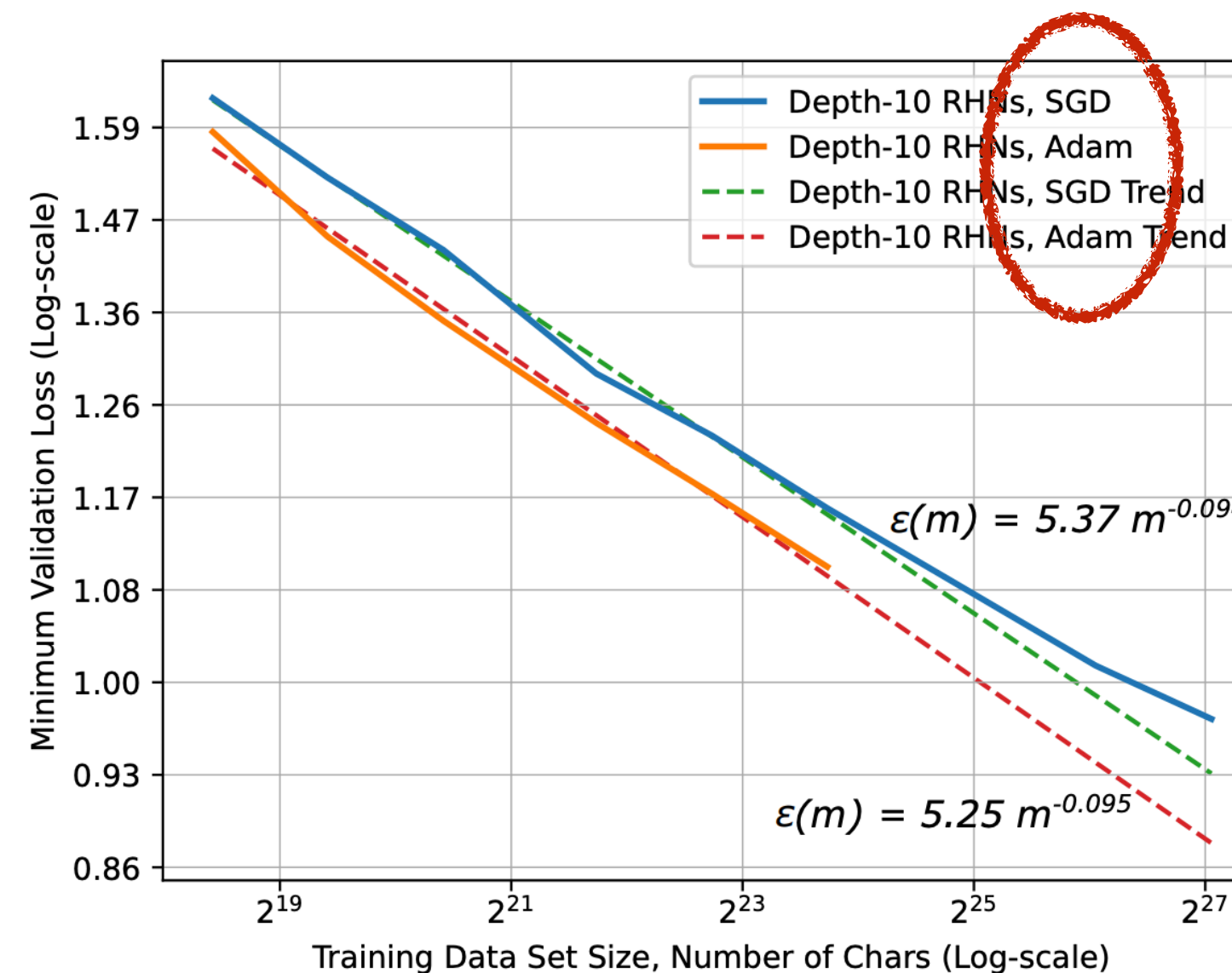
Hestness et al. (2017)

Baidu Scaling Laws

- ▶ Predates modern LLM scaling laws
- ▶ **Power-law** relationships with dataset and model size
- ▶ Observed across diverse tasks and neural network architectures



Machine Translation
(seq2seq - Luong et al. 2015)



Character Language Models
(RHN - Zilly et al. 2017)

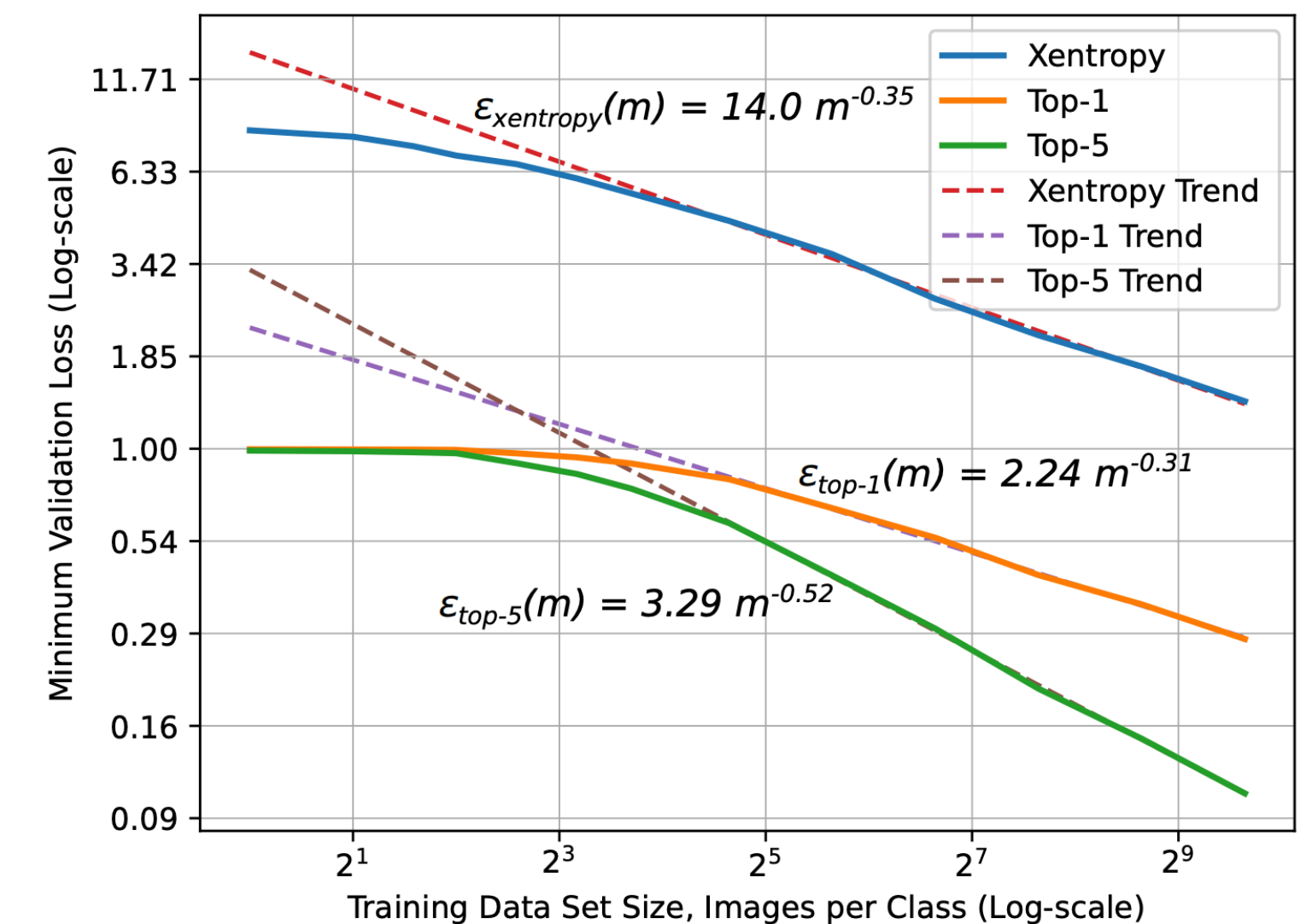
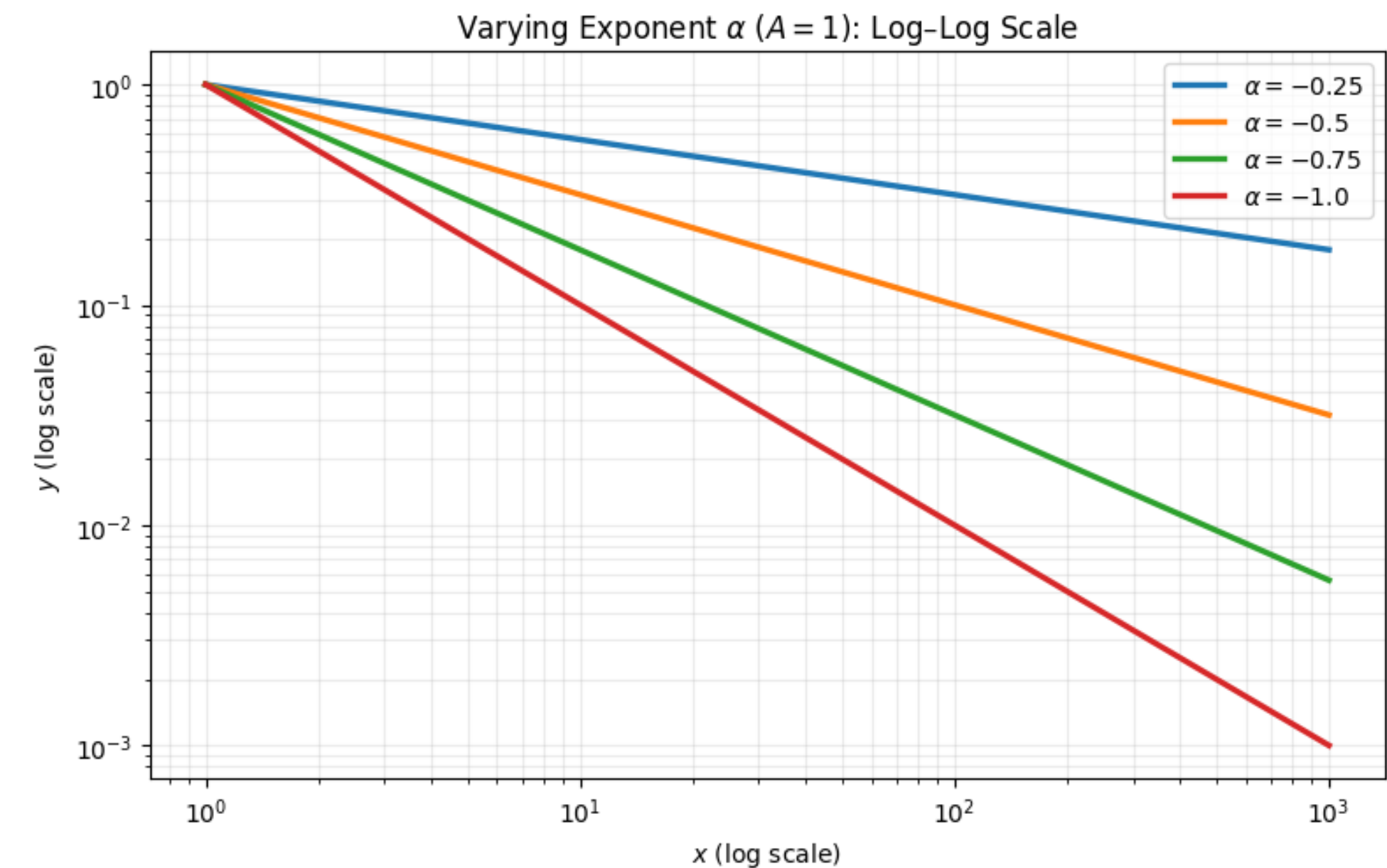
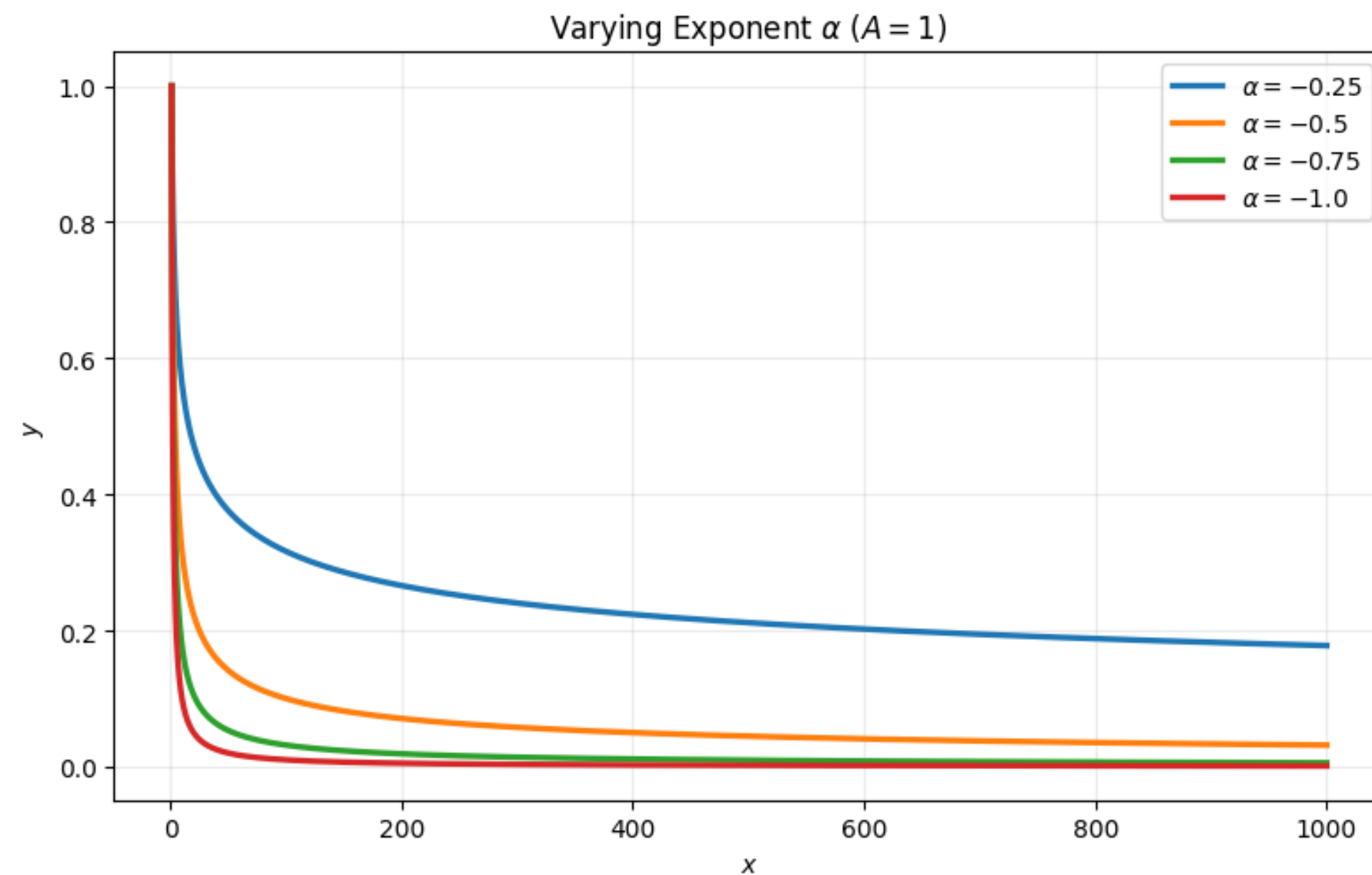


Image Classification
(ResNets - He et al. 2016)

Hestness et al. (2017)

Power Law

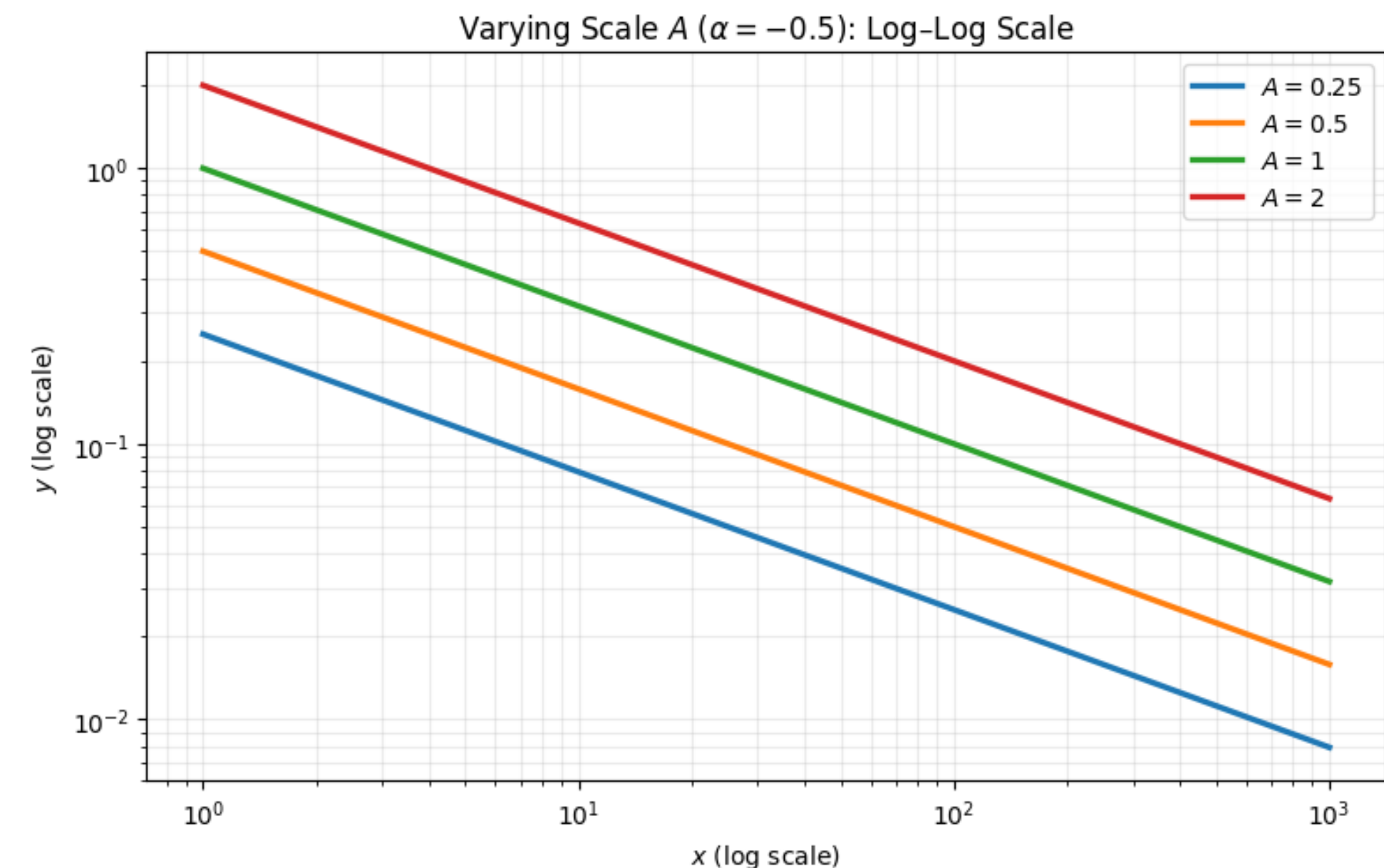
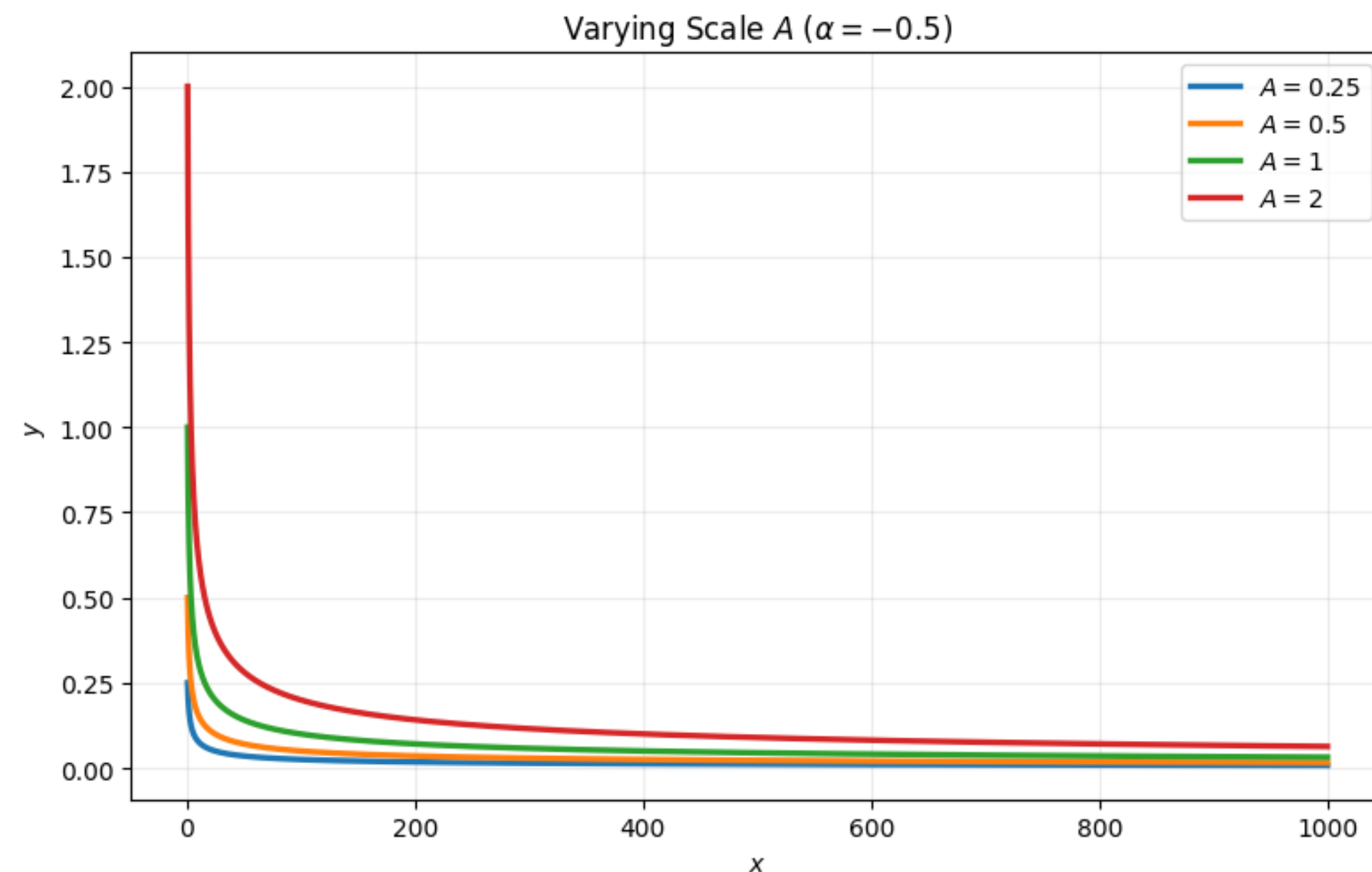
$$y = Ax^\alpha \implies \log y = \underbrace{\log A}_{\text{intercept}} + \underbrace{\alpha}_{\text{slope}} \log x$$



- More negative α means faster improvement (lower loss/error) with scale

Power Law

$$y = Ax^\alpha \implies \log y = \underbrace{\log A}_{\text{intercept}} + \underbrace{\alpha}_{\text{slope}} \log x$$



- ▶ From Baidu's scaling law paper: improved model architectures and optimizers can improve the power-law intercept, but not the exponent
- ▶ i.e., models for a single domain show the same learning curve steepness.

-
- ▶ Is power-law growth the same as exponential growth?

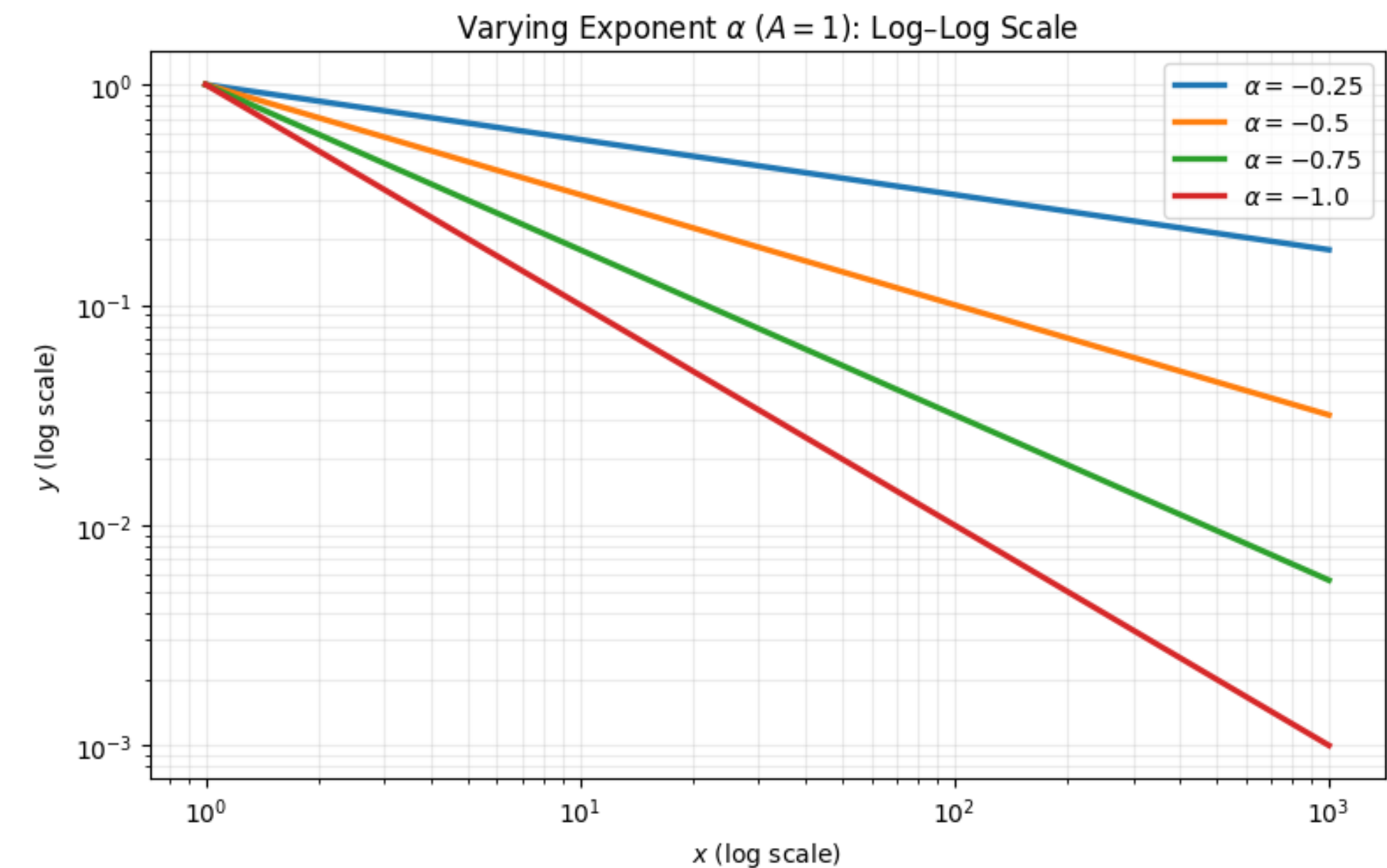
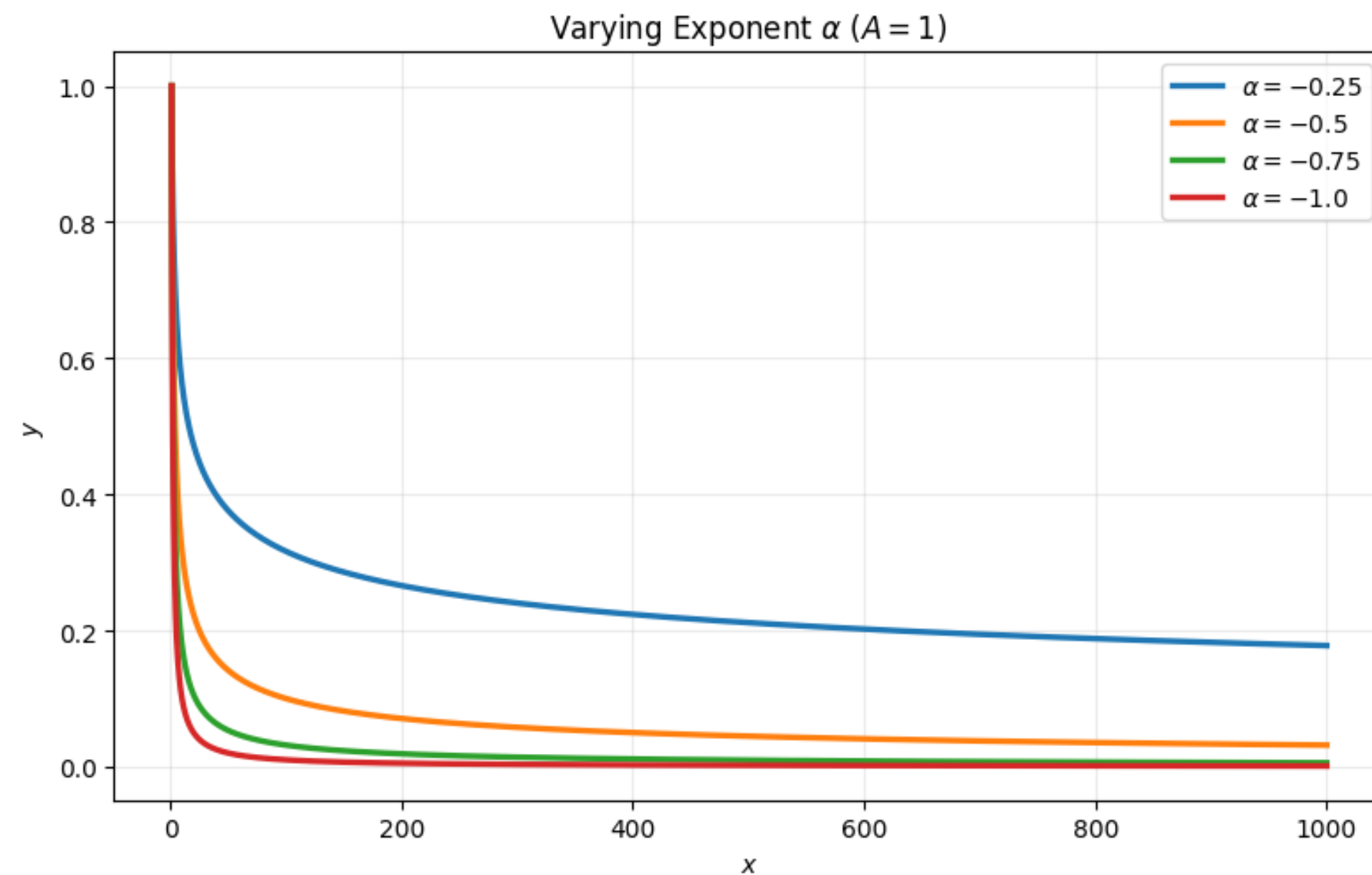


Power Law vs. Exponential

- ▶ Power law: $y = Ax^\alpha \implies \log y = \log A + \alpha \log x$
- ▶ Exponential: $y = Ae^{kx} \implies \log y = \log A + kx$

Power Law

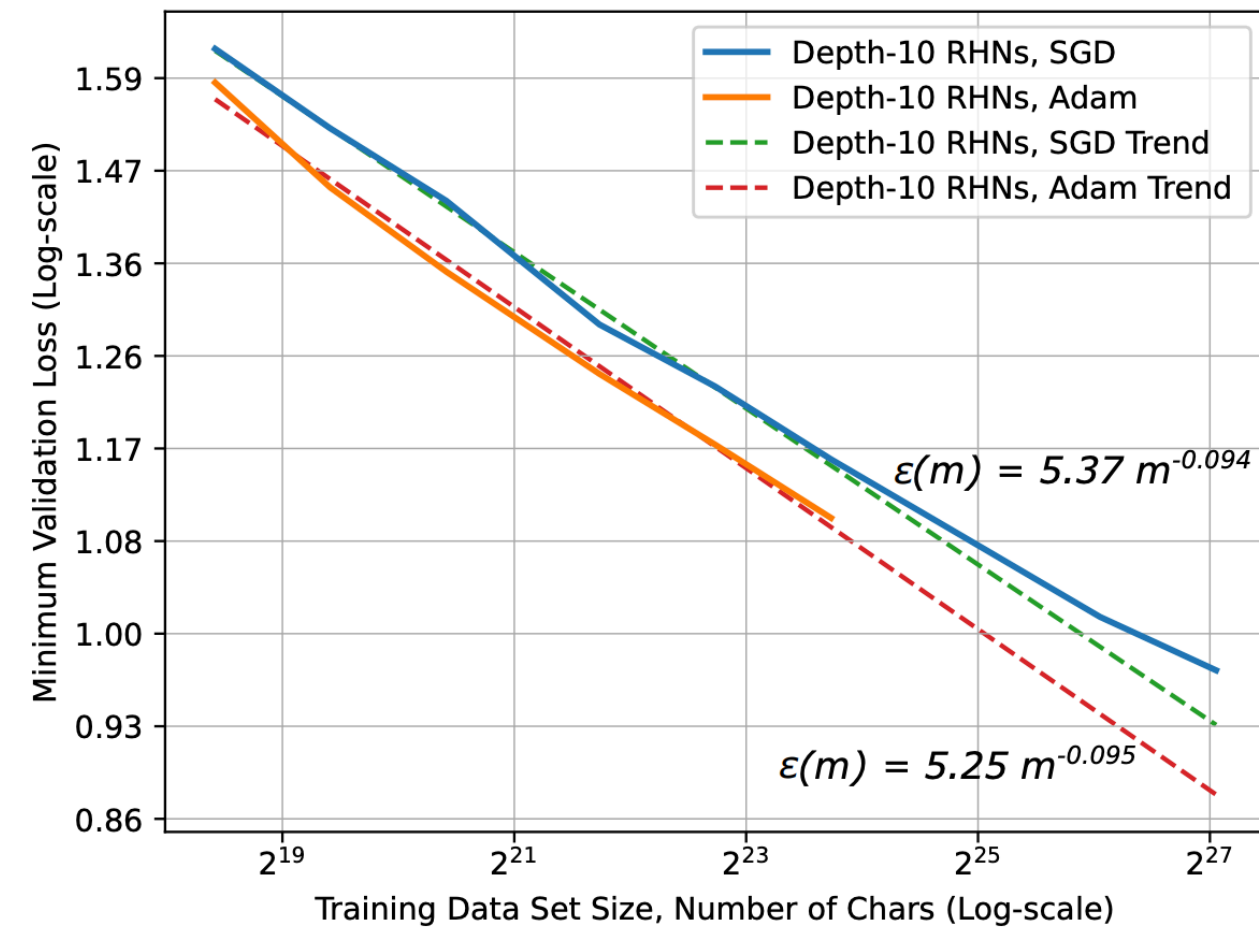
$$y = Ax^\alpha \implies \log y = \underbrace{\log A}_{\text{intercept}} + \underbrace{\alpha}_{\text{slope}} \log x$$



- More negative α means faster improvement (lower loss/error) with scale

Dataset Scaling Laws

Baidu (Hestness et al. 2017)



► General power law: $y = Ax^\alpha$

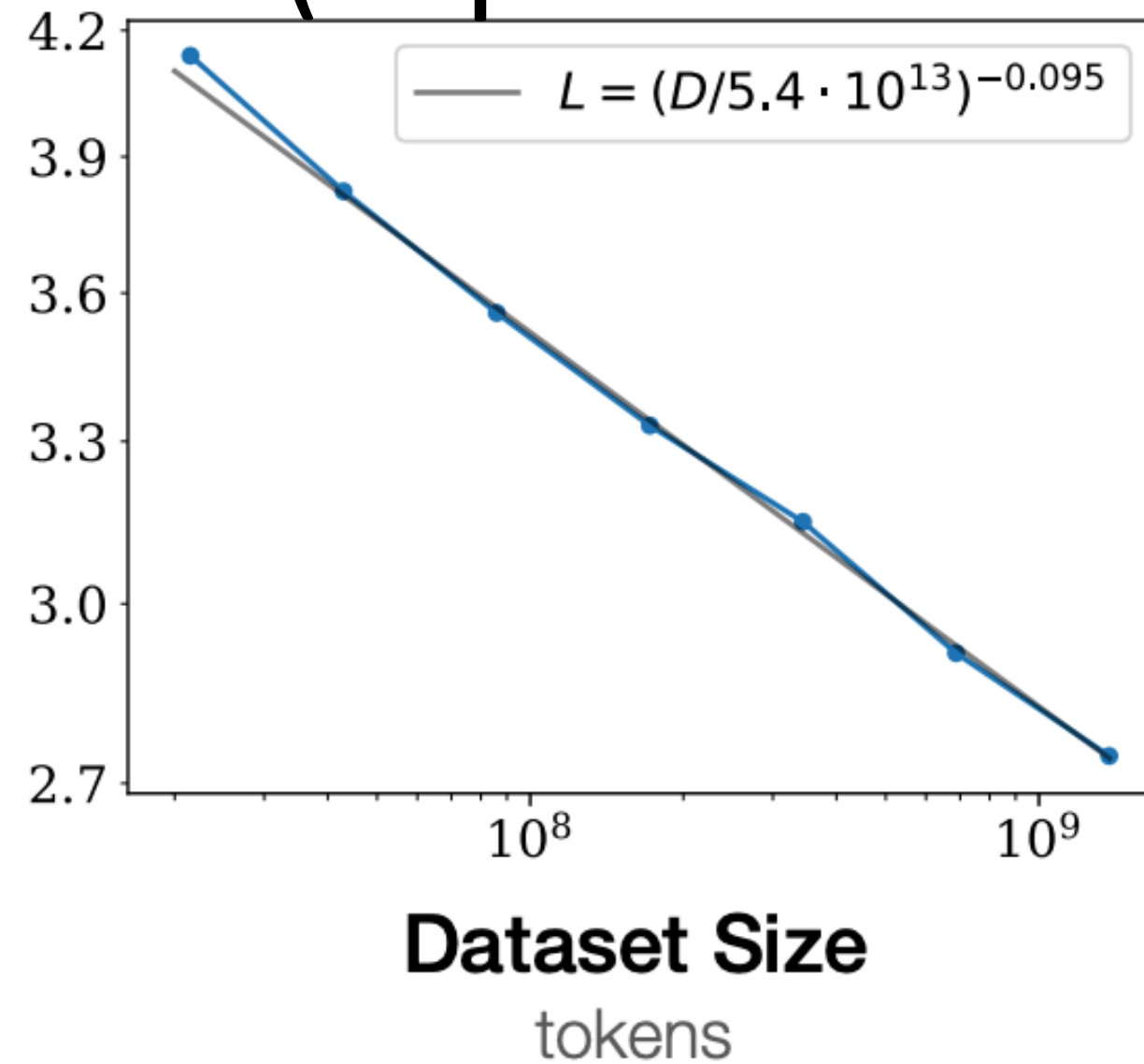
► Dataset scaling law:

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

$$= \underbrace{D_c^{\alpha_D}}_A \underbrace{D^{-\alpha_D}}_{x^\alpha}$$

$$\propto D^{-\alpha_D}$$

OpenAI (Kaplan et al. 2020)

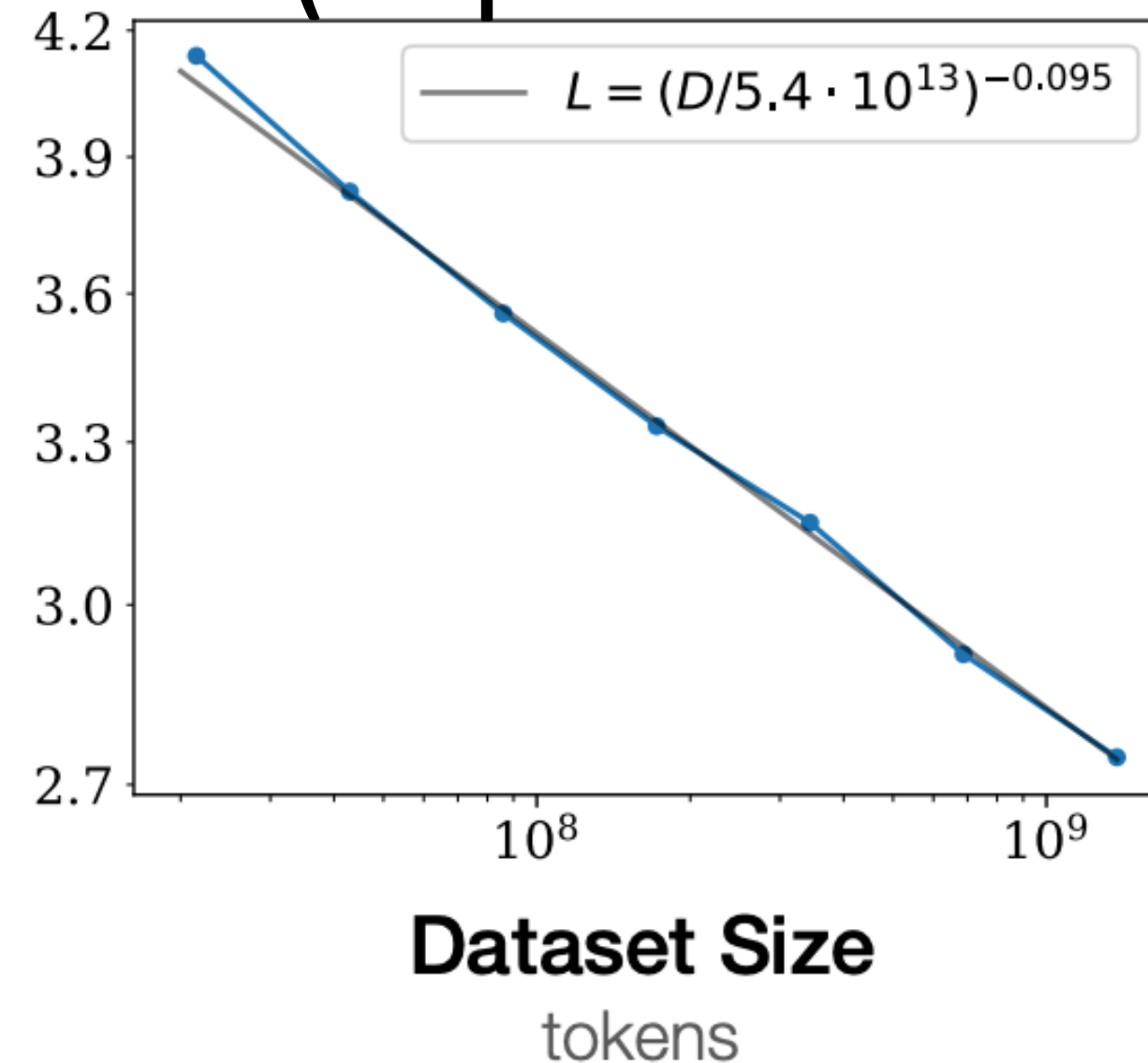


Dataset Scaling Laws

- ▶ The exponent captures the fundamental scaling behavior:
i.e. how quickly loss decreases as D increases?

$$L(D) \approx \left(\frac{D_c}{D}\right)^{\alpha_D} = \left(\frac{D}{D_c}\right)^{-\alpha_D} = D_c^{\alpha_D} D^{-\alpha_D} \propto D^{-\alpha_D}$$

OpenAI (Kaplan et al. 2020)

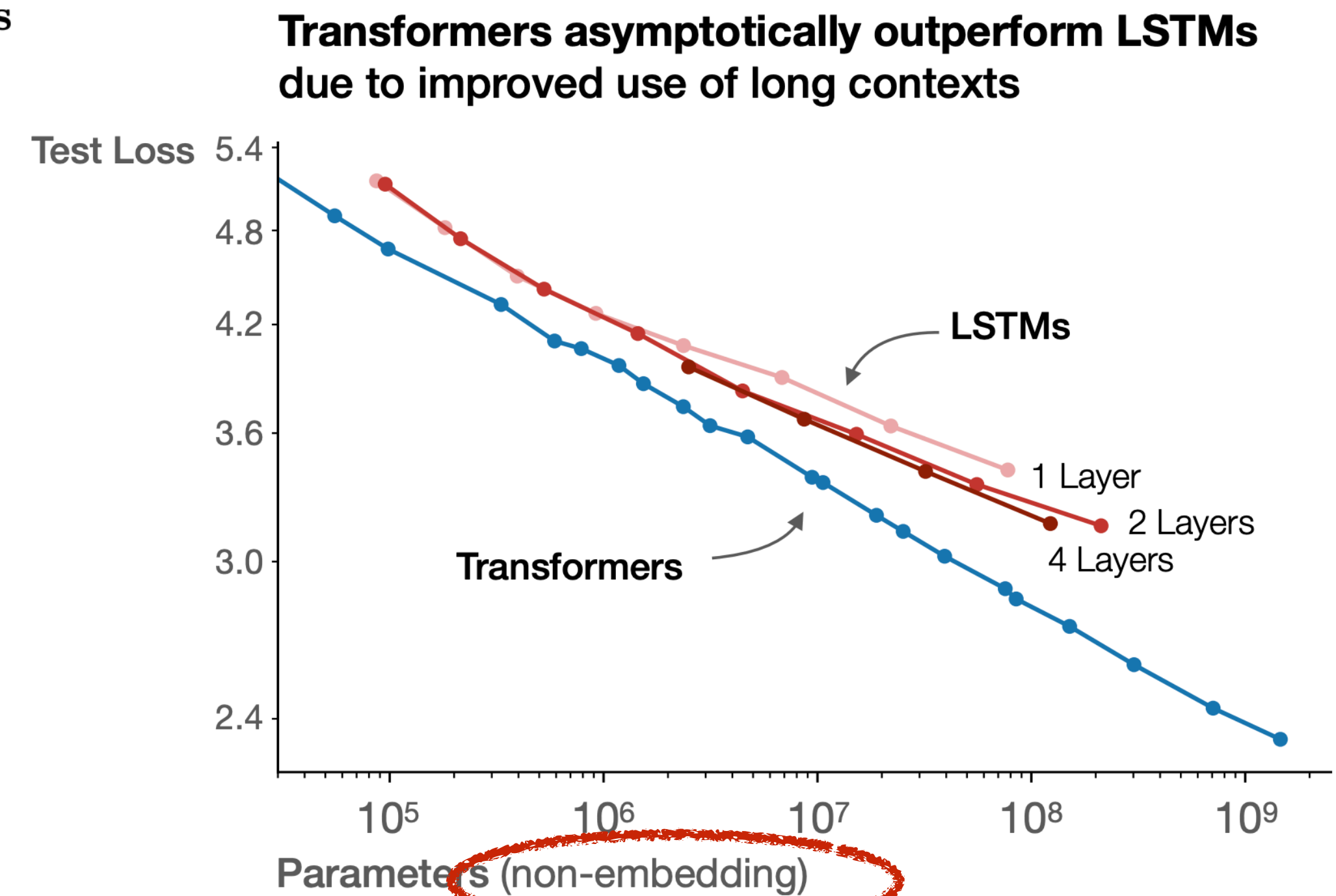
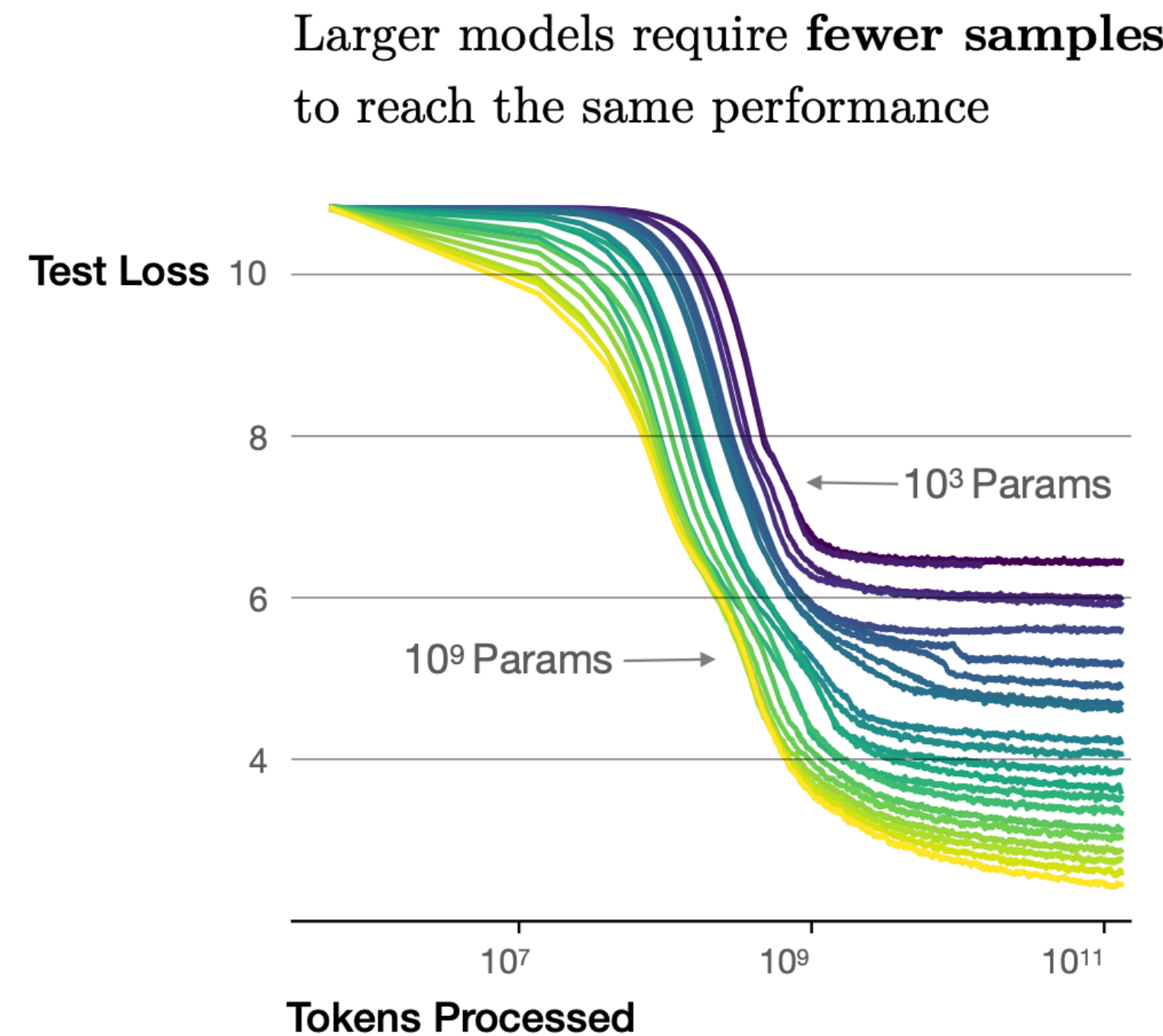
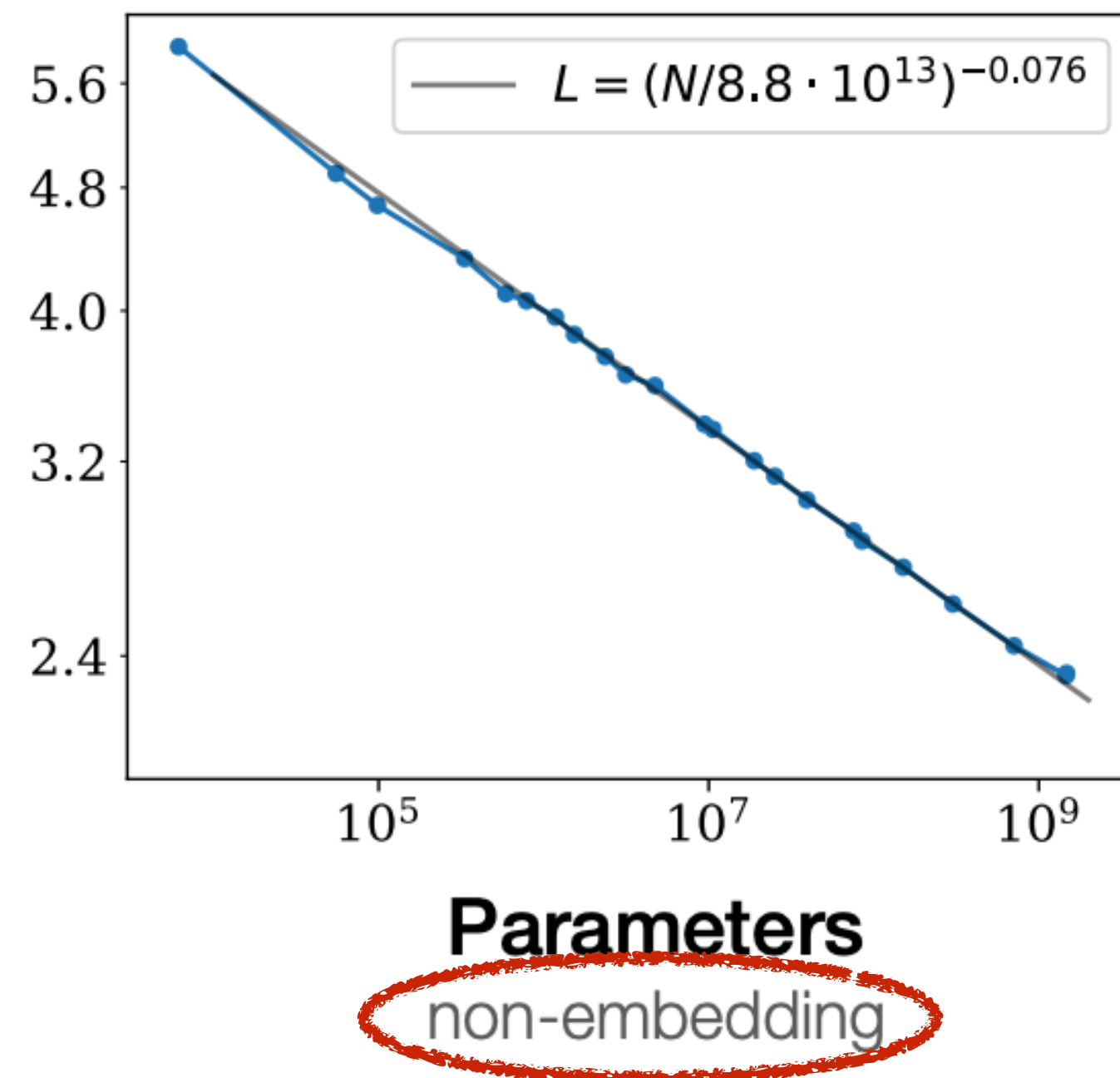


- ▶ For example, $L(D) \propto D^{-0.095}$
- ▶ Increasing D by 10x, will reduce loss by ~20%

$$\frac{L(10D)}{L(D)} = \frac{(10D)^{-0.095}}{D^{-0.095}} = 10^{-0.095} \approx 0.803$$

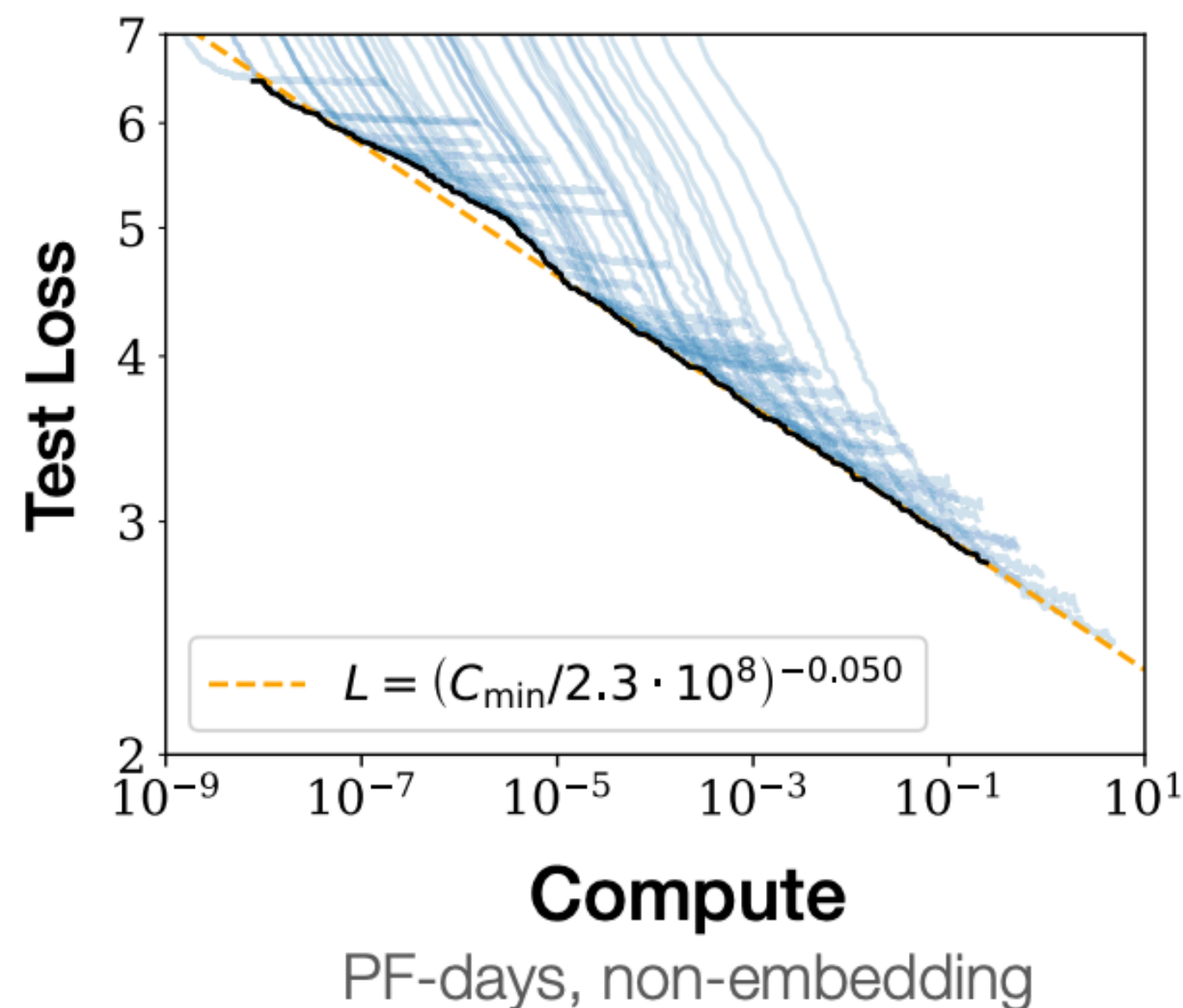
Model Size Scaling Laws

- ▶ Similar to the dataset scaling laws



OpenAI (Kaplan et al. 2020)

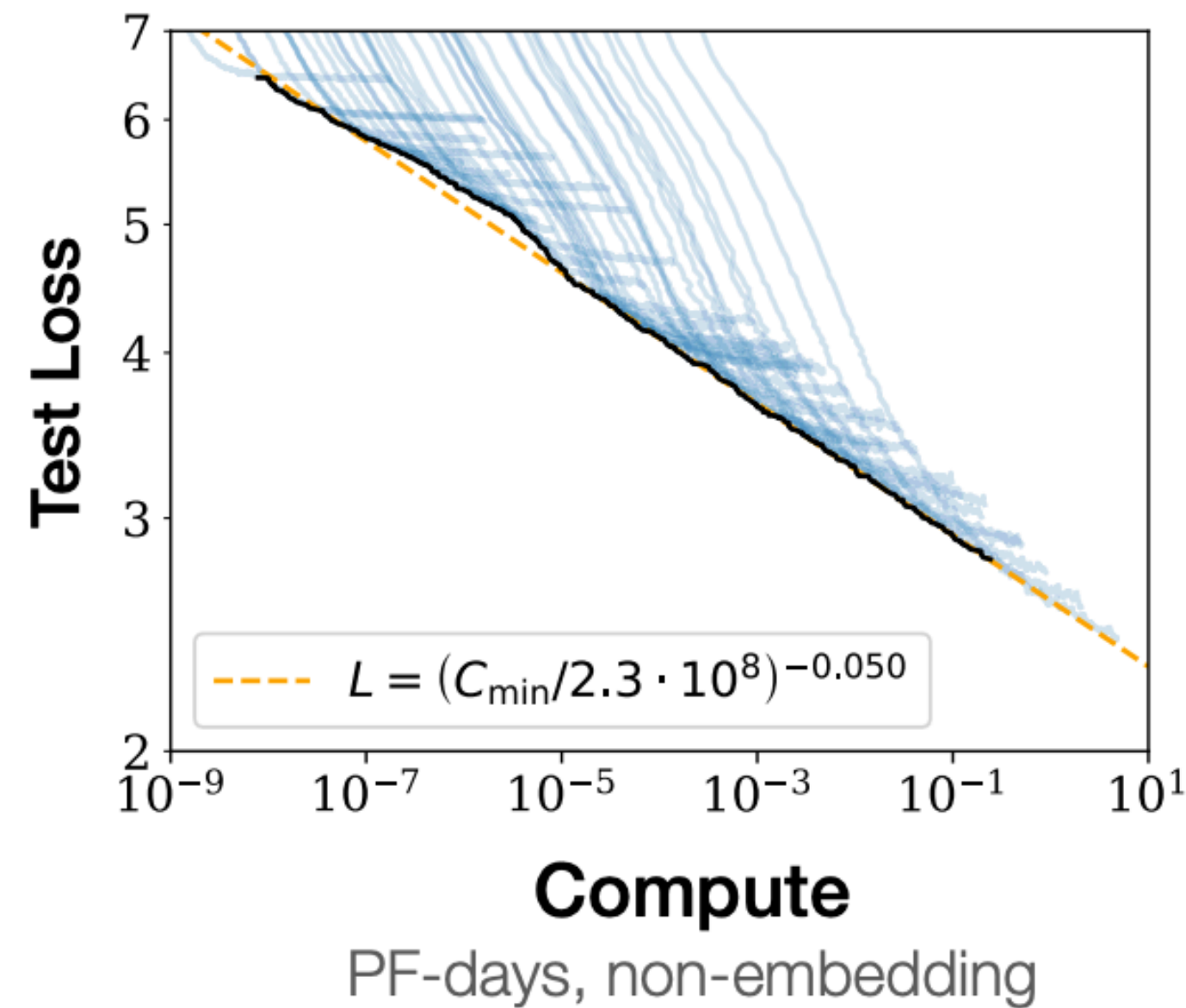
Compute Scaling Laws



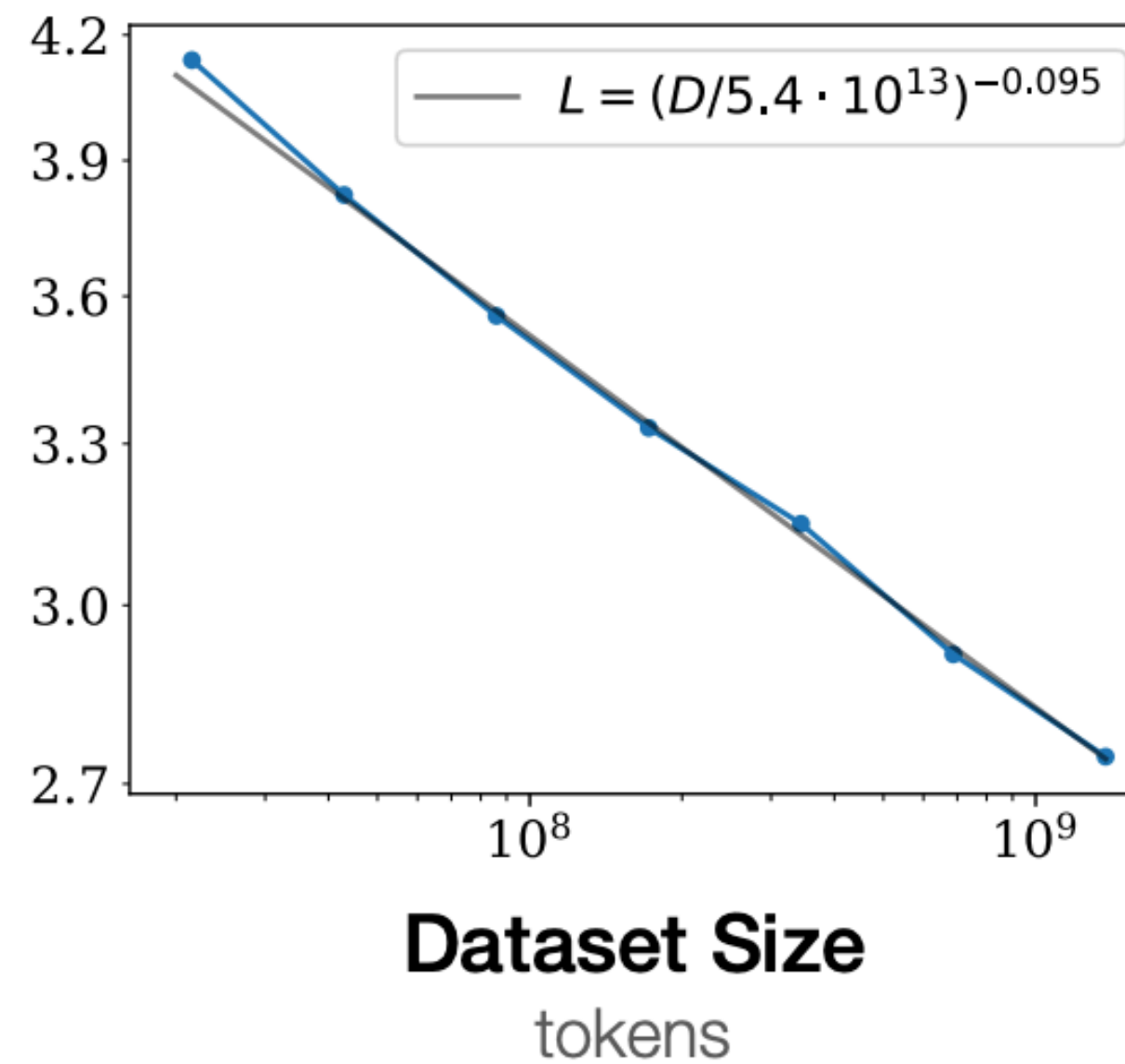
- ▶ **Blue** lines: model of different sizes N , trained on increasing amount of data D
- ▶ **Black** line - “compute-efficient frontier”: lowest test loss achieved by any model at a given compute budget
- ▶ **Orange** dashed line: use log-space regression to fit a power law

GPT Scaling Laws

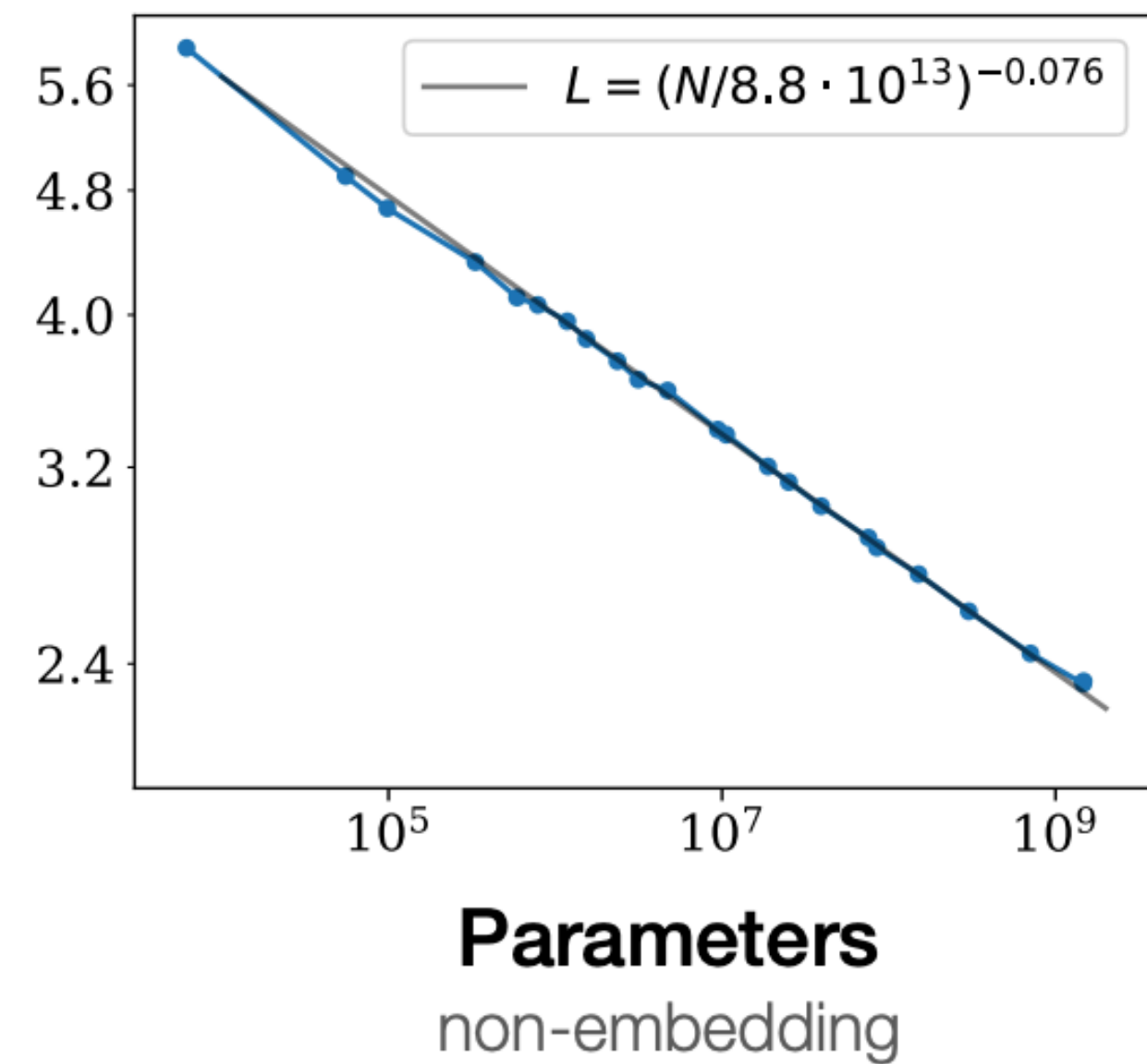
- ▶ With more compute, larger models get further down the loss “frontier”
- ▶ Building a bigger model (increasing compute) will decrease test loss!



$$L(C) \propto C^{-0.050}$$



$$L(D) \propto D^{-0.095}$$

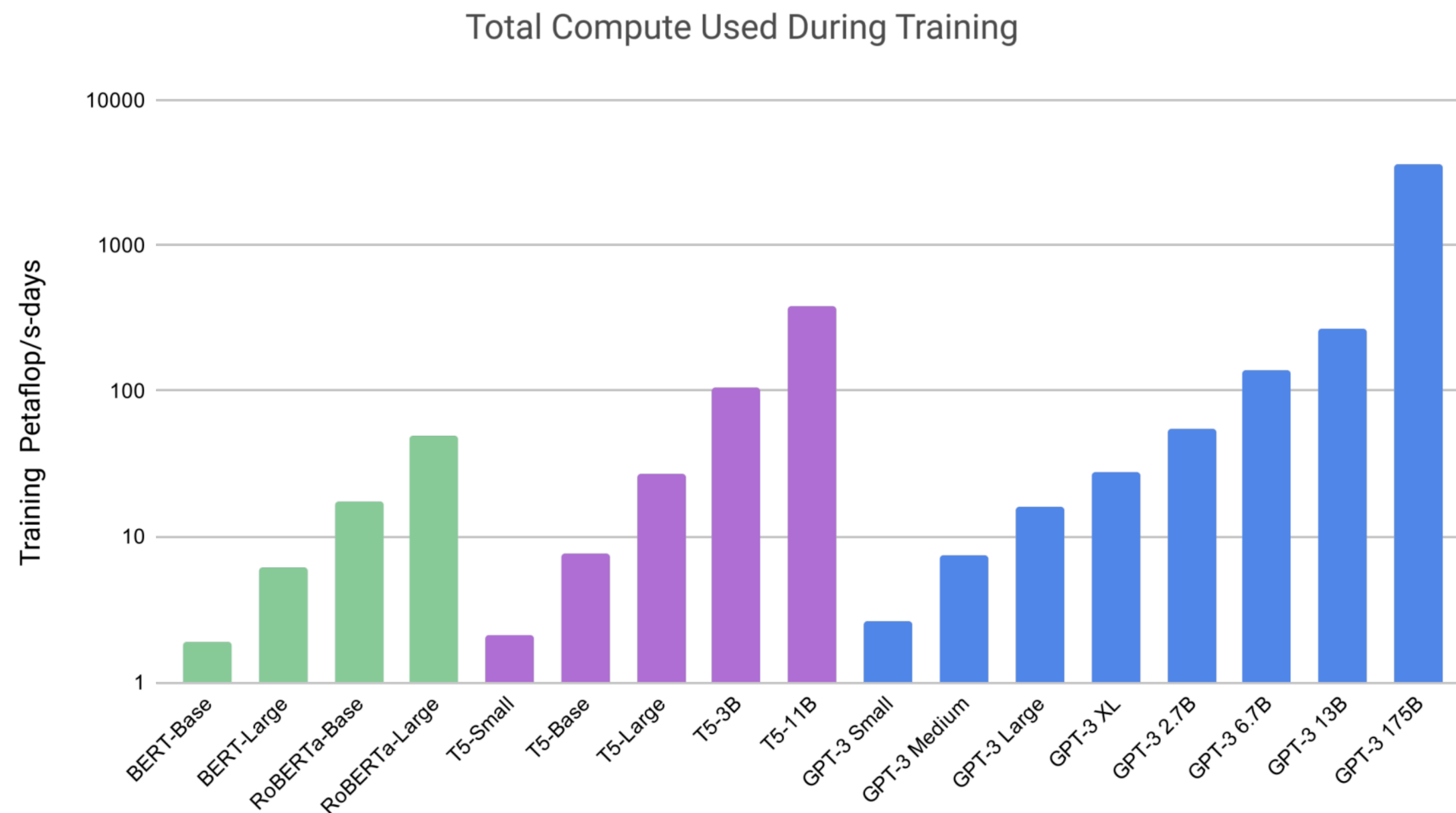


$$L(N) \propto N^{-0.076}$$

GPT Scaling Laws

- For a fixed compute budget, Kaplan et al. (2020) argued for scaling model size **N** much faster than training data size **D** (train bigger models!):

$$N_{\text{opt}} \propto C^{0.73}, \quad D_{\text{opt}} \propto C^{0.27}$$

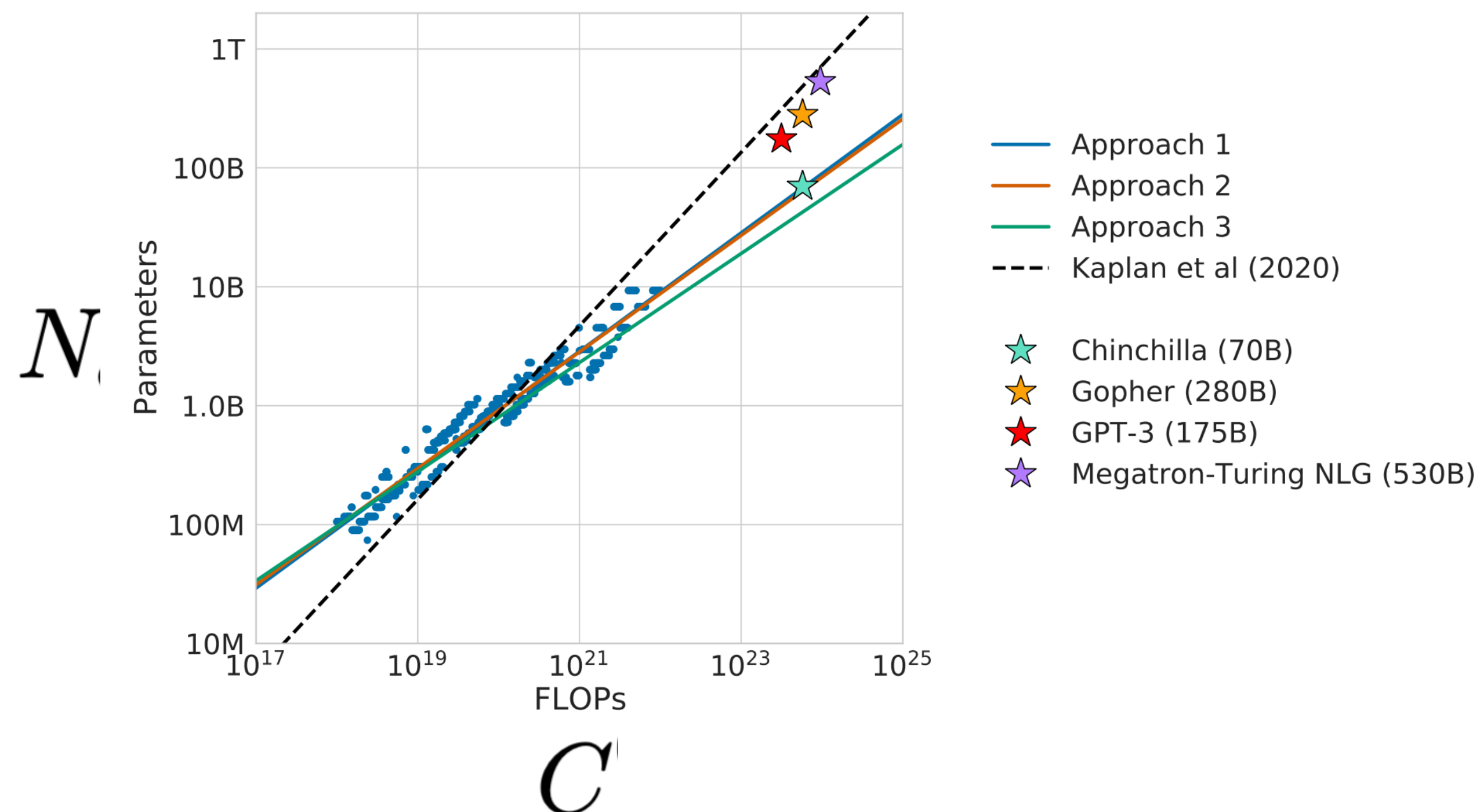


Kaplan et al. (2020)

Chinchilla Scaling Laws

- In contrast, Hoffmann (2022) argued for scaling model size **N** and training data equally **D** (training models on more data, avoid under-training):

$$N_{\text{opt}} \propto C^{0.50}, \quad D_{\text{opt}} \propto C^{0.50}$$



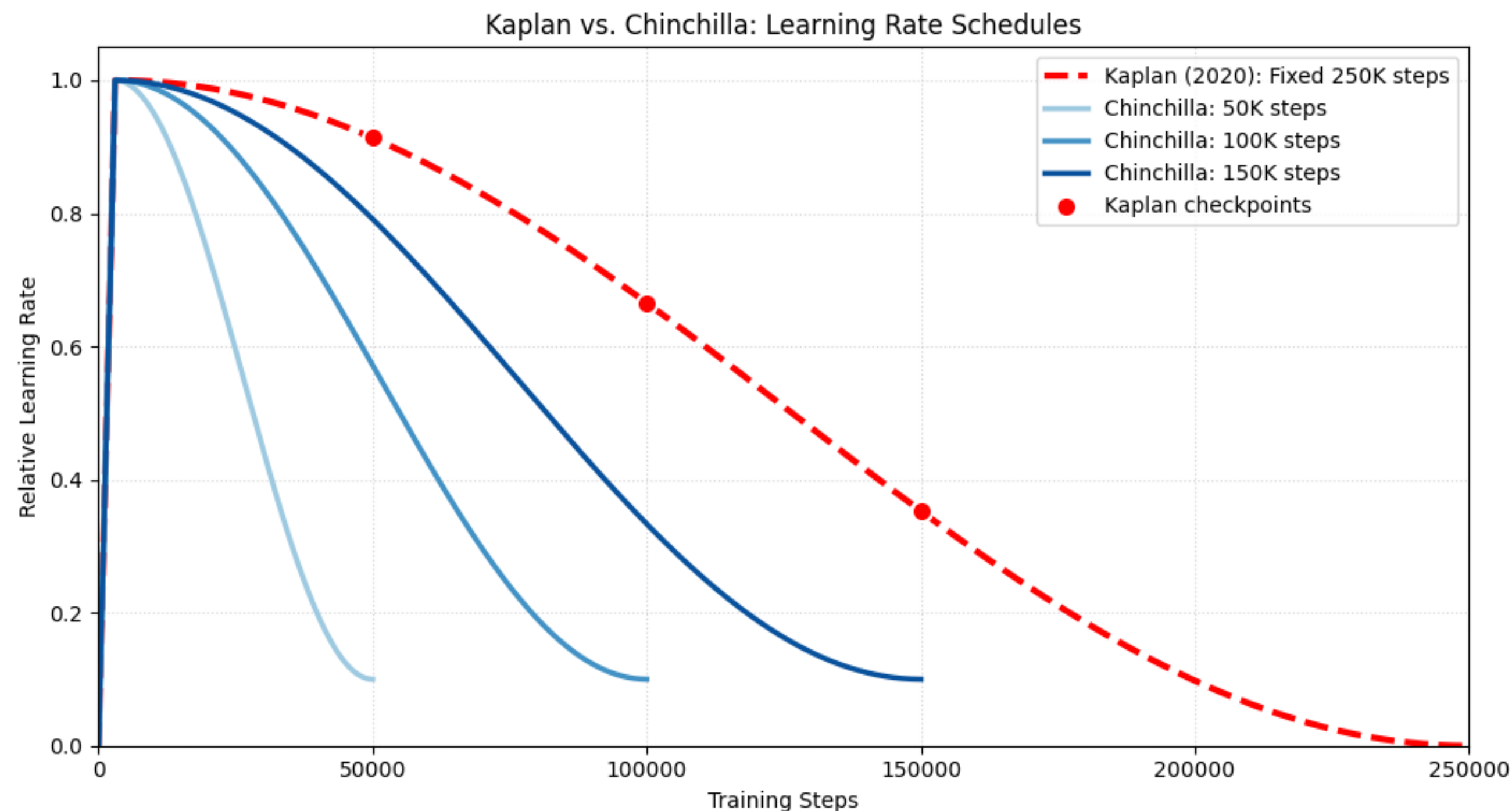
MMLU results:

Random	25.0%
Average human rater	34.5%
GPT-3 5-shot	43.9%
<i>Gopher</i> 5-shot	60.0%
<i>Chinchilla</i> 5-shot	67.6%
Average human expert performance	89.8%

Hoffmann et al. (2022)

Chinchilla Scaling Laws

- ▶ Learning rate: adaptive schedule vs. Kaplan+ (2020)'s fixed schedule
- ▶ Model size: up to 70B vs. Kaplan+ (2020)'s 1.5B non-embedding parameters



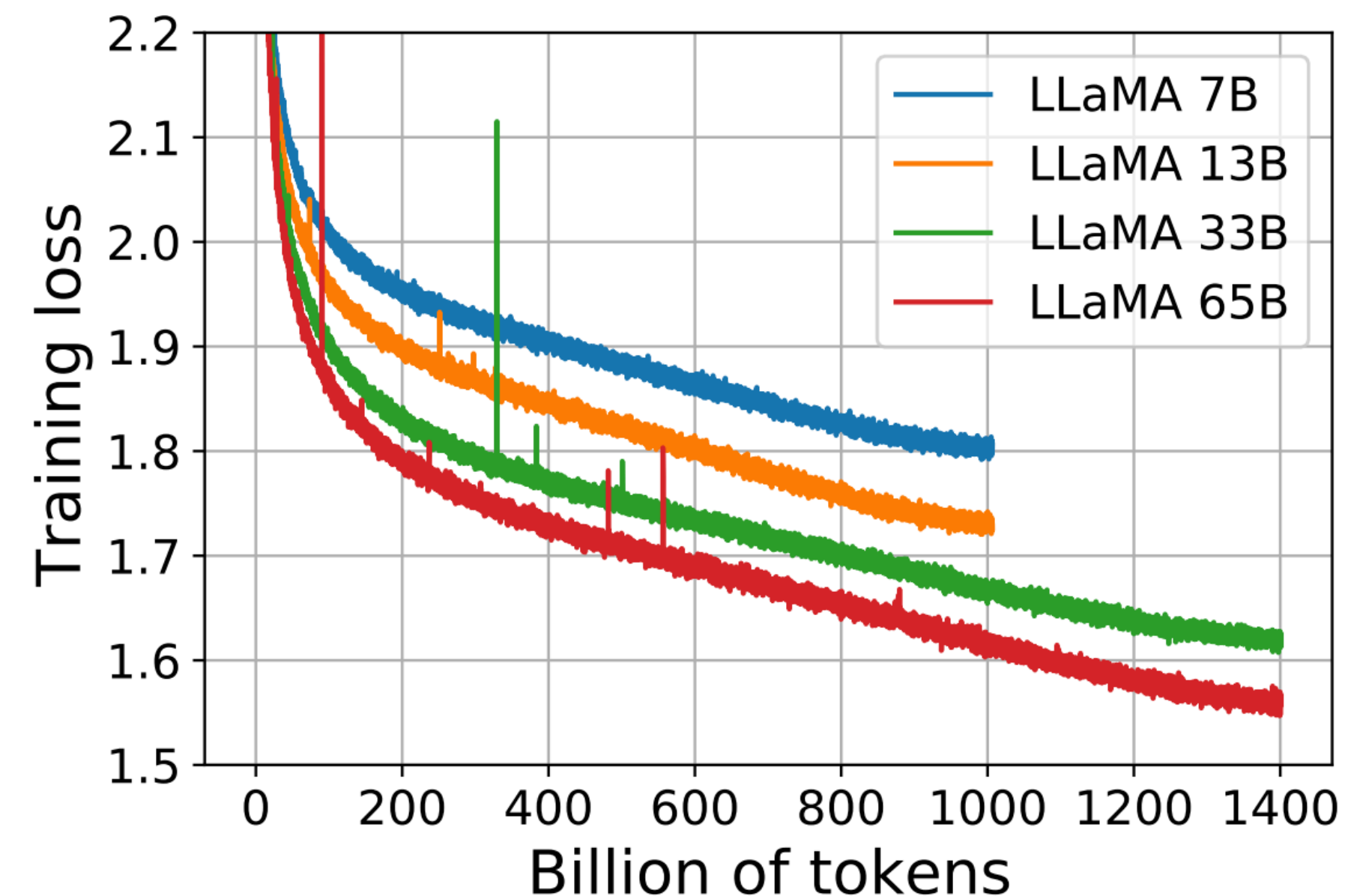
Hoffmann et al. (2022)

Llama

- Besides scale dataset and model sizes for **training** compute budget, it is also important to consider the **inference (test-time)** compute budget

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

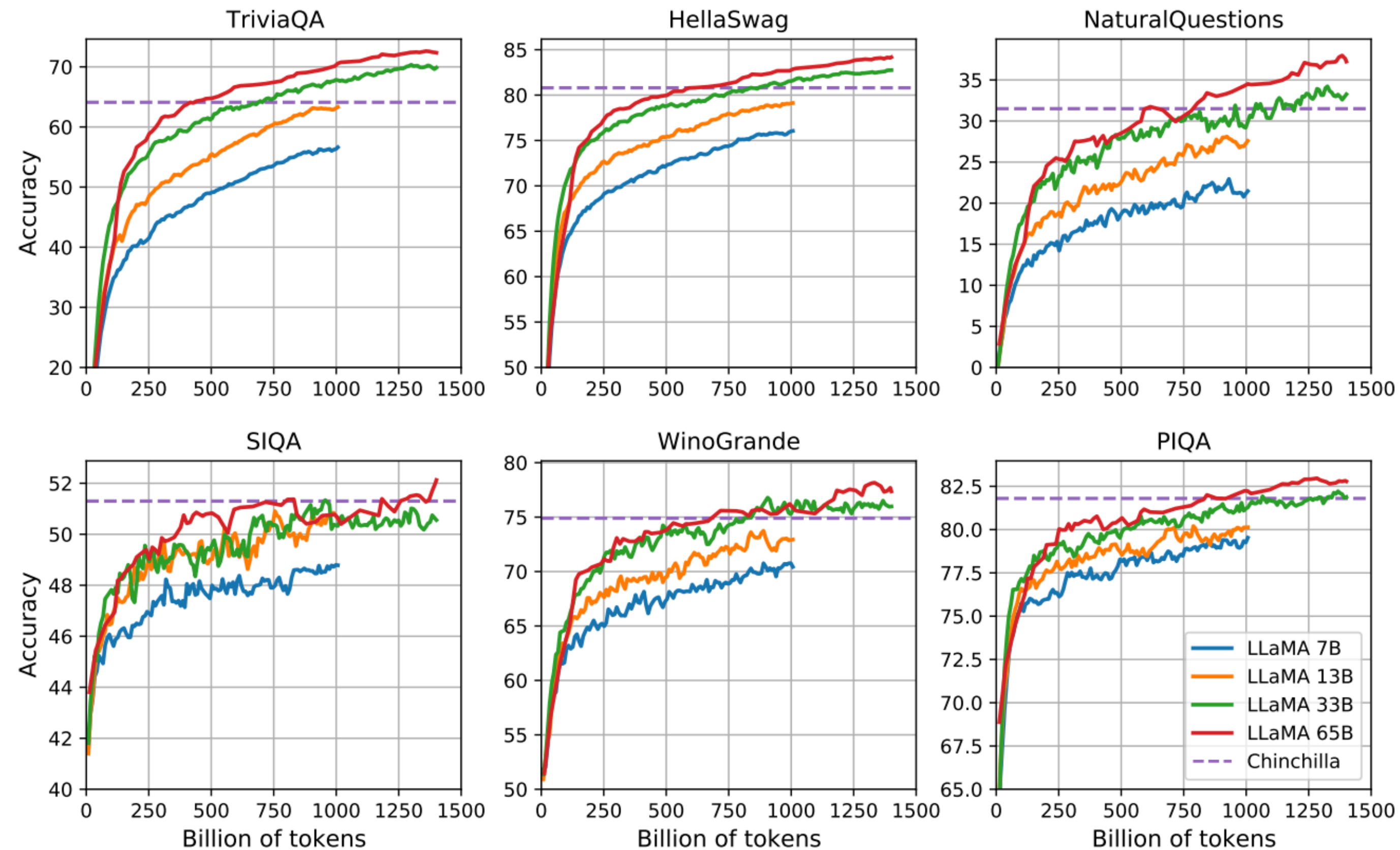
trained on 1~1.4T tokens



Touvron et al. (2023)

Llama

- Smaller models continue to improve when trained longer on substantially more tokens.



Touvron et al. (2023)

Baidu Scaling Laws

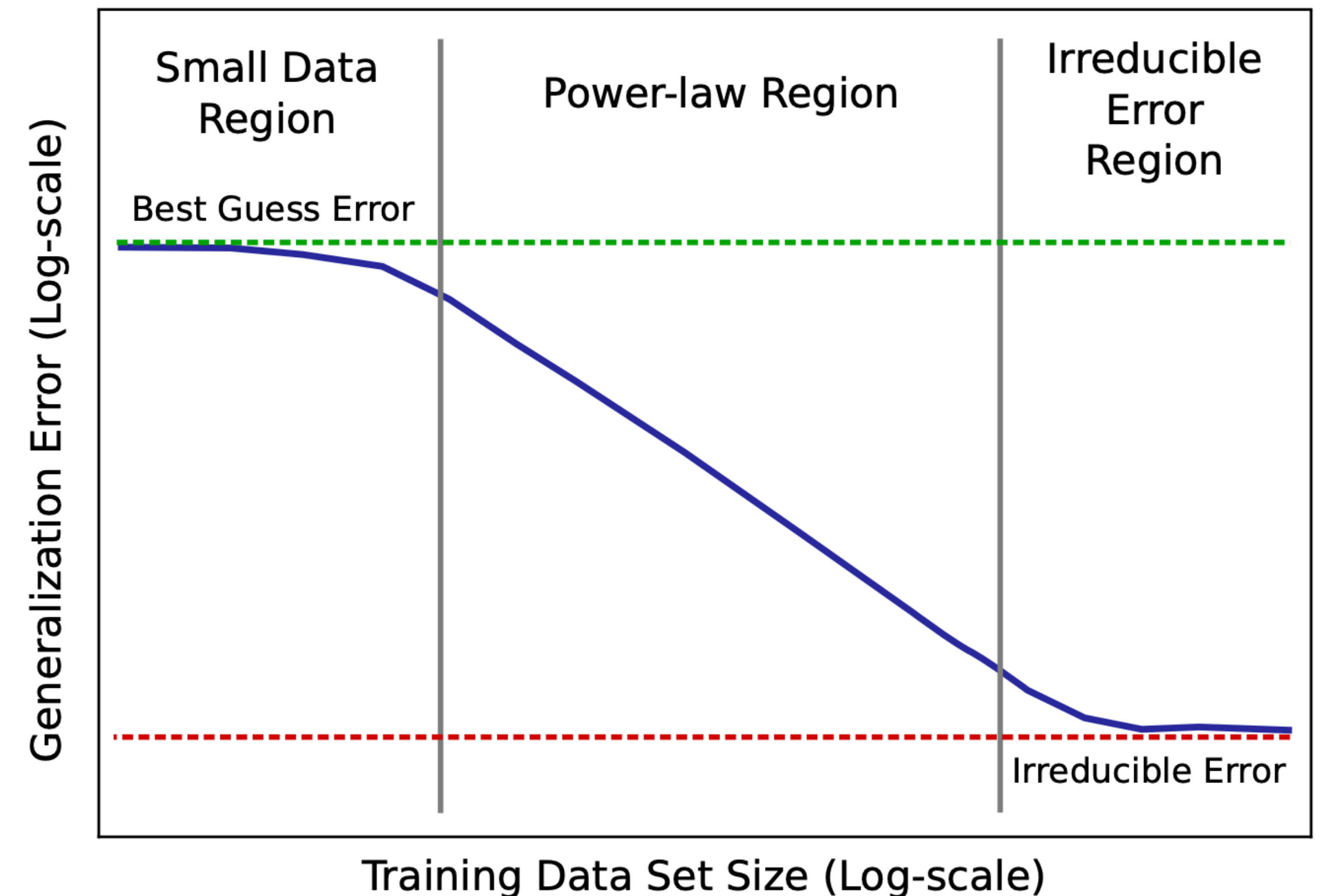
- ▶ Scaling behavior wasn't unique to Transformers or LLMs.
- ▶ Irreducible error:

$$L(D) = L_{\infty} + AD^{-\alpha_D}$$

- ▶ e.g., cross-entropy loss in LM:

$$H(p, q) = H(\underset{\substack{\uparrow \\ \text{true}}}{p}) + D_{\text{KL}}(\underset{\substack{\uparrow \\ \text{predicted}}}{p} || q)$$

“I want a cup of _____”



Hestness et al. (2017)

Summary: Scaling Laws

- ▶ Hestness+ 2017: Scaling behavior wasn't unique to Transformers or LLMs.
- ▶ Kaplan+ 2020 (GPT-3): Prioritize model size, make models bigger
- ▶ Hoffmann+ 2022 (Chinchilla): Increase training data & model size equally
- ▶ Touvron+ 2023 (Llama-1): Smaller models continue to improve when trained on substantially more data, benefiting inference efficiency.
- ▶ Later: data-constrained scaling, sparse/MoE, test-time scaling, ...
- ▶ And more: synthetic data, code, multimodal data, data filtering/curation, repeated data, post-training, reinforcement learning