



Transformer Sprachmodelle und die gegenwärtige Deep Learning Forschung

Philipp Bongartz, remote

14. April 2022

Vier Trends in der Deep Learning Forschung

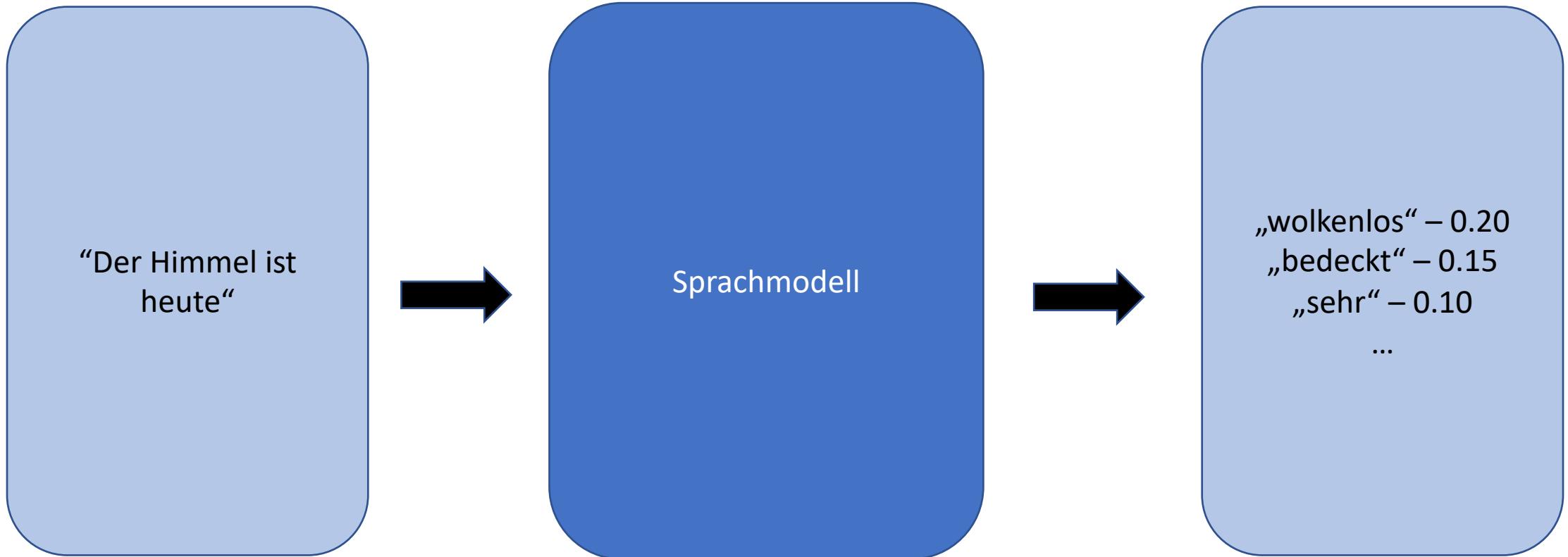
- Skalierung zu extrem großen Netzwerken
- Unsupervised Learning statt Supervised Learning
- Multimodales Training
- Generative Modelle

Für diese vier Trends stehen besonders die spektakulären Verbesserungen von Sprachmodellen in den letzten 2, 3 Jahren.

Generative Modelle sind besonders spannend:
Image Generation wird immer besser -
<https://thisxdoesnotexist.com>



Was ist ein Sprachmodell?



Generative Sprachmodelle - Textgeneration

PROMPT

GENERIERTER TEXT

Generative Sprachmodelle - Textgeneration

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.

Generative Sprachmodelle - Textgeneration

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.

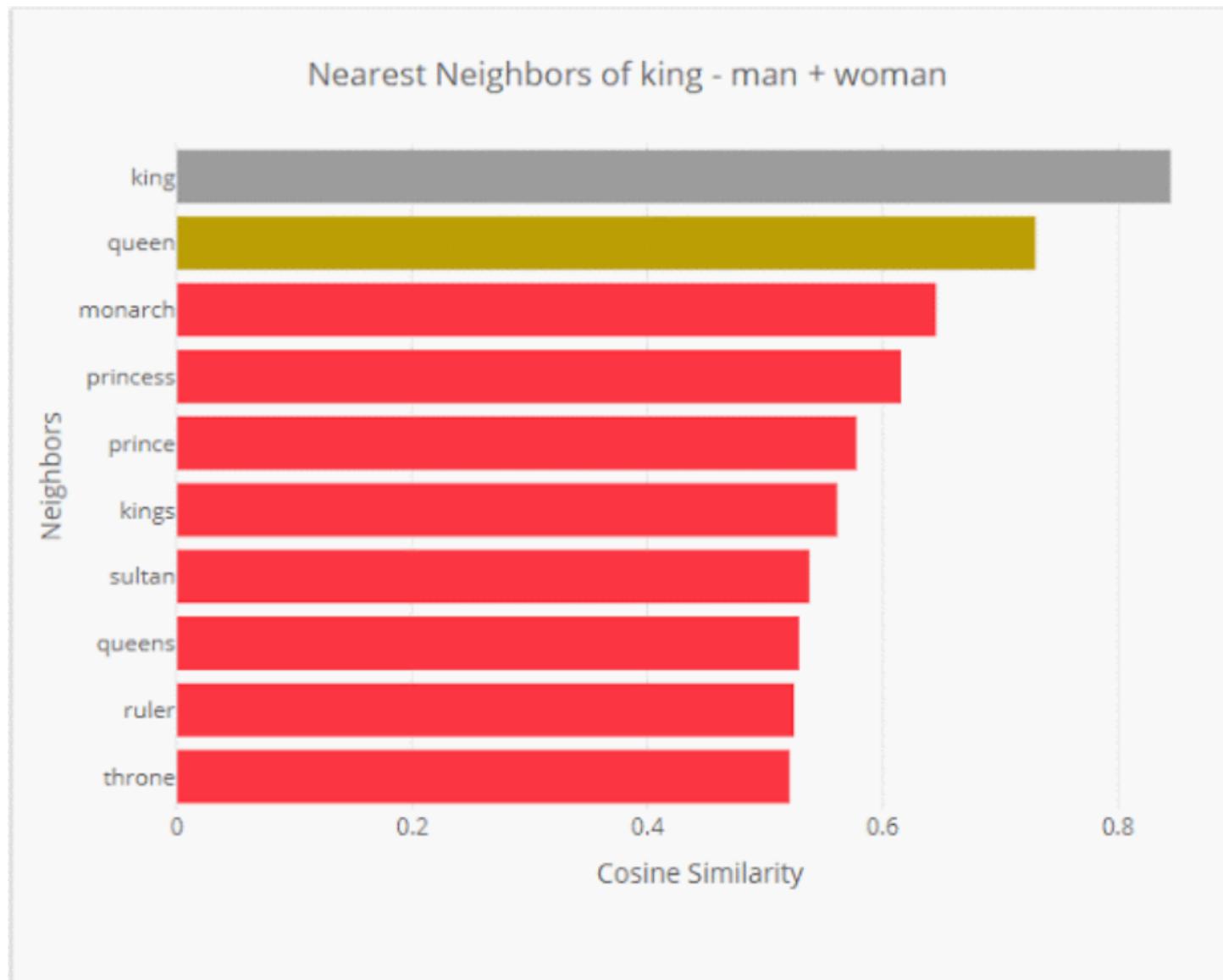
The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

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

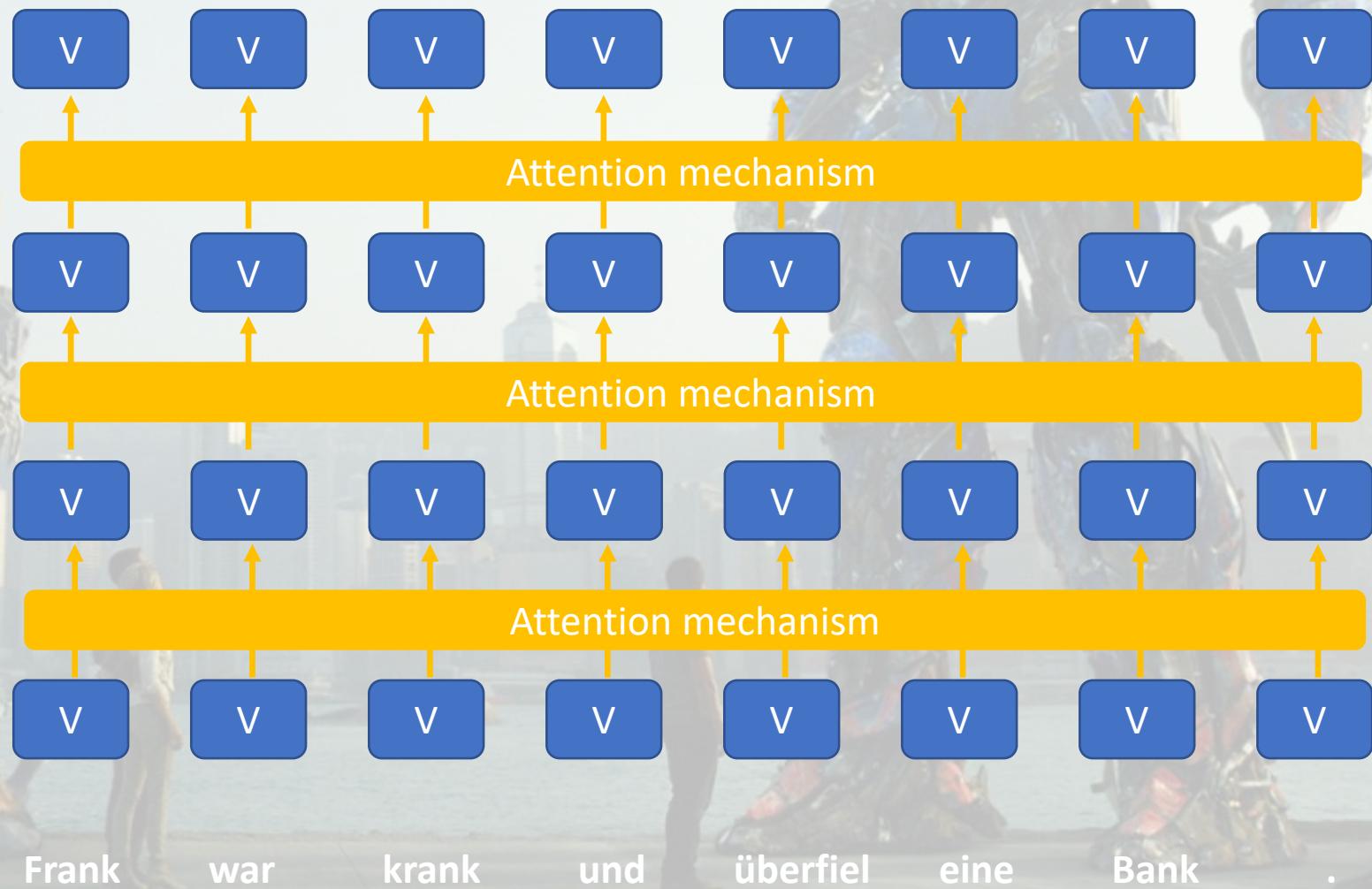
Better Language Models and Their Implications

We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

Wortvektoren kodieren relativ vage

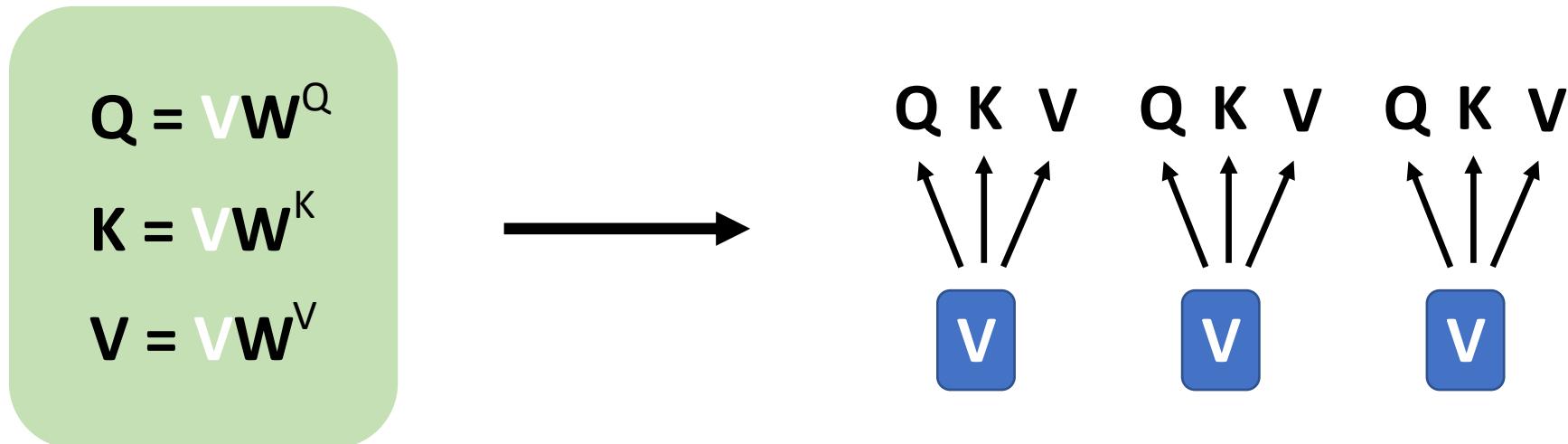


Transformer-Architektur



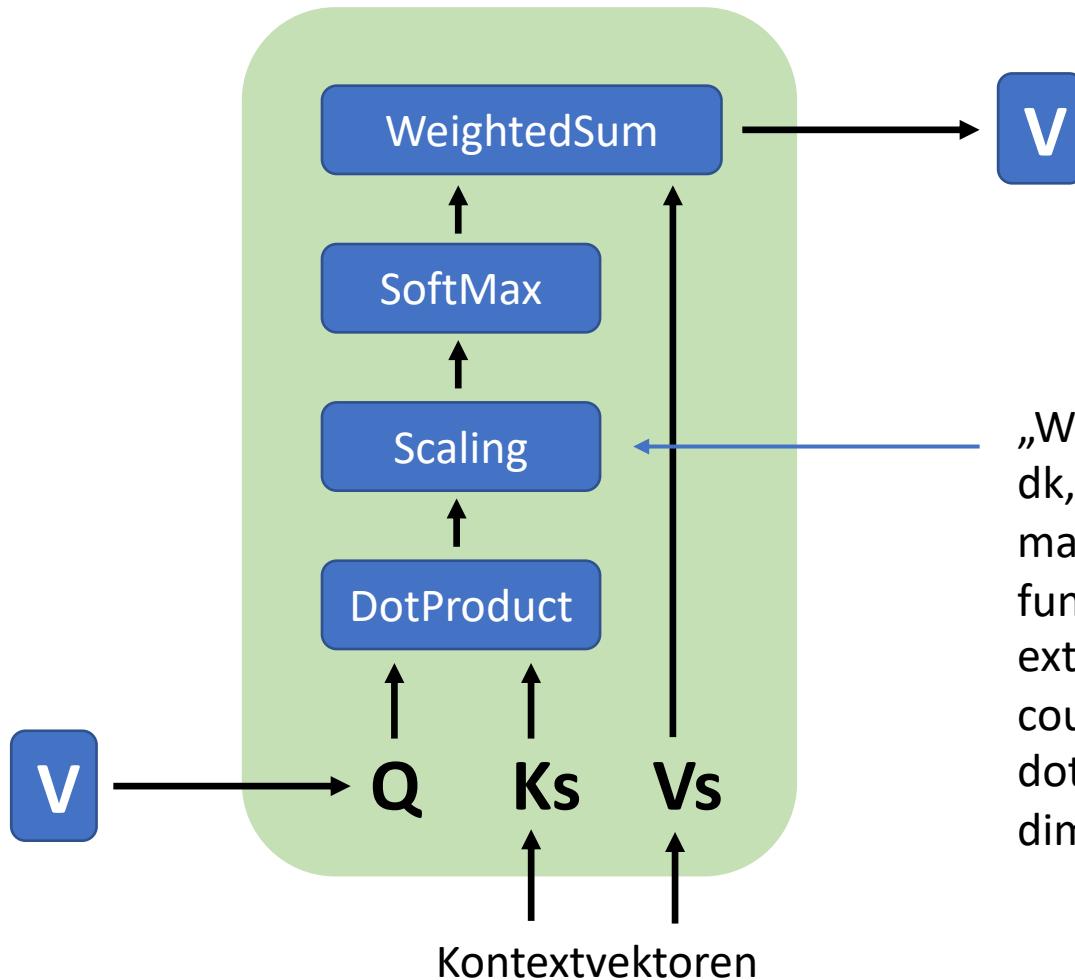
- Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017. (10000 citations!)

An attention function can be described as mapping a query and a set of key-value pairs to an output, where the **query Q**, **keys K**, **values V**, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.



Wie wird der Vektor verwurstet?

1. Er wird in einen Vektor **Q** verwandelt.
2. Dann mit den **Keys** des Kontextes verglichen.
3. Und dann aus den **Values** des Kontextes gewichtet neu berechnet.



$$Q = VW^Q$$

$$K = VW^K$$

$$V = VW^V$$

„We suspect that for large values of dk, the dot products grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients. To counteract this effect, we scale the dot products by $1/\sqrt{dk}$. (dk = dimension of Q and K)“

Transformer Architektur – Charakteristiken

(quadratisch, bag-of-words)

Anwendungsgebiete:

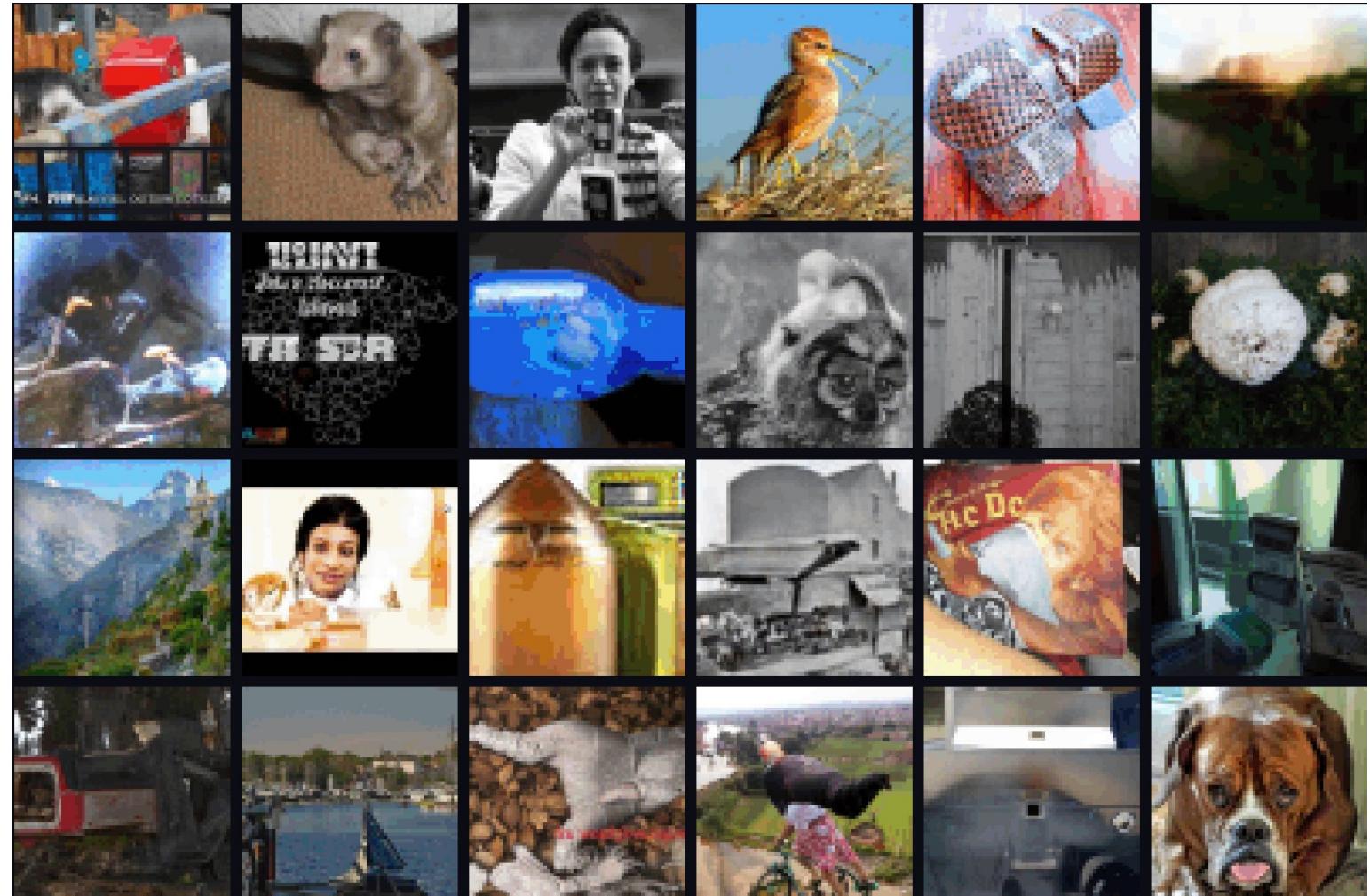
- Neural Machine Translation
- Sprachmodelle
- Sequence learning/generation

Training:

- Autoregressive für Generation
- Masked für NLP-Tasks
- Finetuning für spezifische Tasks

Versionen:

- Sparse Transformer (OpenAI)
- Compressive Transformer (Deepmind)
- Reformer (GoogleAI)
- Linformer (Facebook)

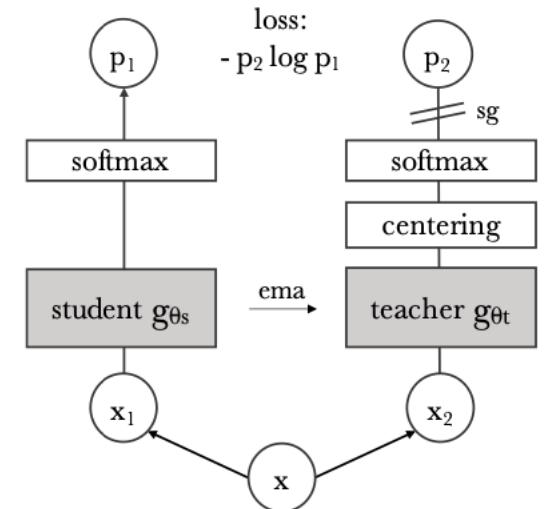


Unsupervised Learning funktioniert jetzt in vielen Bereichen – nicht nur bei Sprachmodellen

Supervised



DINO



Generative Sprachmodelle – Fragen beantworten

Question: Who is the 40th president of the United States?

Answer:

Ronald Reagan

Generative Sprachmodelle – Chatbot

You are interviewing GPT-2, an extremely intelligent chatbot.

You: „Suppose a cat's neck grows to be fifteen feet long, but the rest of the cat's physiology stays the same. What happens to the cat?

GPT-2:

"Well, as I understand it, if this were to happen, then the cat would no longer have the ability to breathe air. It could not survive in its new environment."

Generative Sprachmodelle – Zusammenfassung

Prehistoric man sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave in modern day France 36,000 years ago. Now, with the help of cutting-edge technology, those works of art in the Chauvet-Pont-d'Arc Cave have been reproduced to create the biggest replica cave in the world.

TL;DR

The original site in Vallon-Pont-D'arc in Southern France is a Unesco World Heritage site and is the oldest known and the best preserved cave decorated by man. The replica cave was built a few miles from the original site in Vallon-Pont-D'Arc in Southern France. The cave contains images of 14 different species of animals including woolly rhinoceros, mammoths, and big cats.

Generative Sprachmodelle – Übersetzung

English: My name is Philipp.

French translation:

Je m'appelle Philipp.



OpenAI



Gary Marcus – GPT-2 and the nature of intelligence

Here's the problem: upon careful inspection, it becomes apparent the system has no idea what it is talking about: it cannot follow simple sequence of events nor reliably have any idea what might happen next.

Gary Marcus – GPT-2 and the nature of intelligence

I have four plates and put one cookie on each. The total number of cookies is

24, 5 as a topping and 2 as the filling

Gary Marcus – GPT-2 and the nature of intelligence

I left my keys at a booth in the back of the pub at the army base near the old mill town. I went home. My keys are now

on that old genie on a little tower in England somewhere.

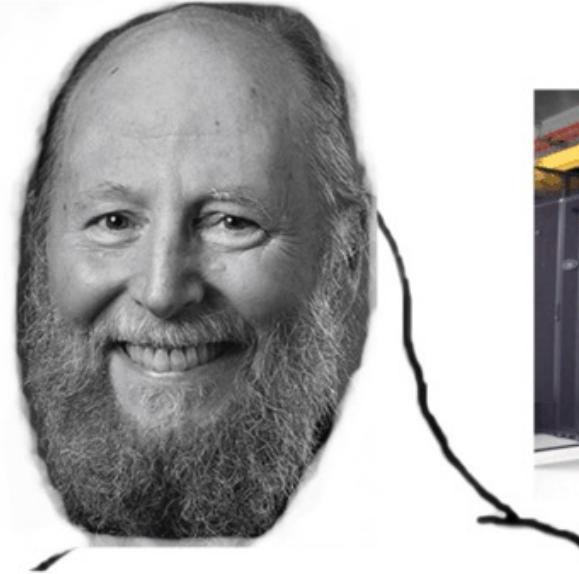
Gary Marcus – GPT-2 and the nature of intelligence

If you drop an anvil on your foot, your foot will tend to

cramp up, so it's a good idea to do these slowly.



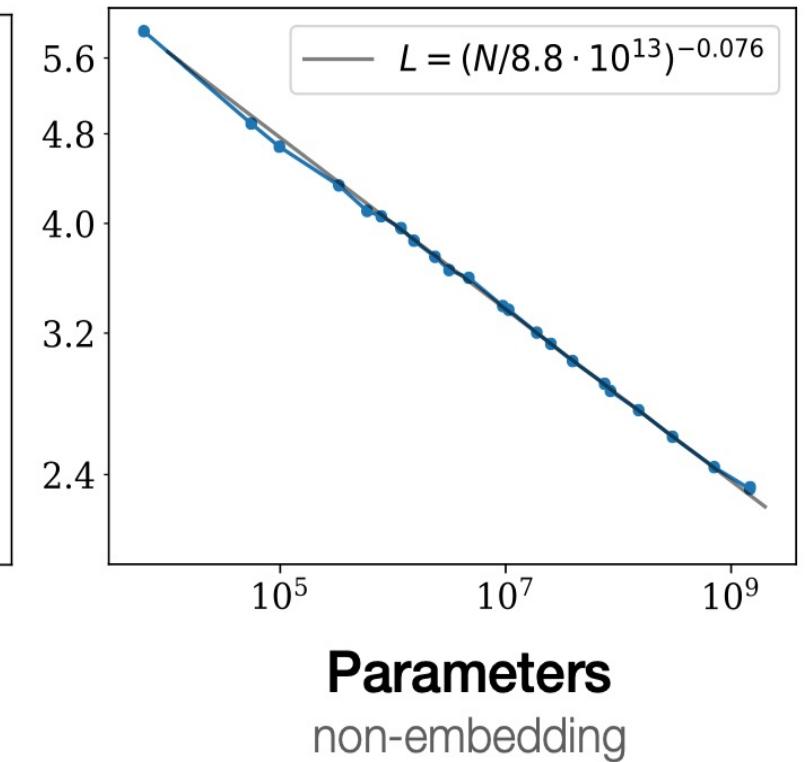
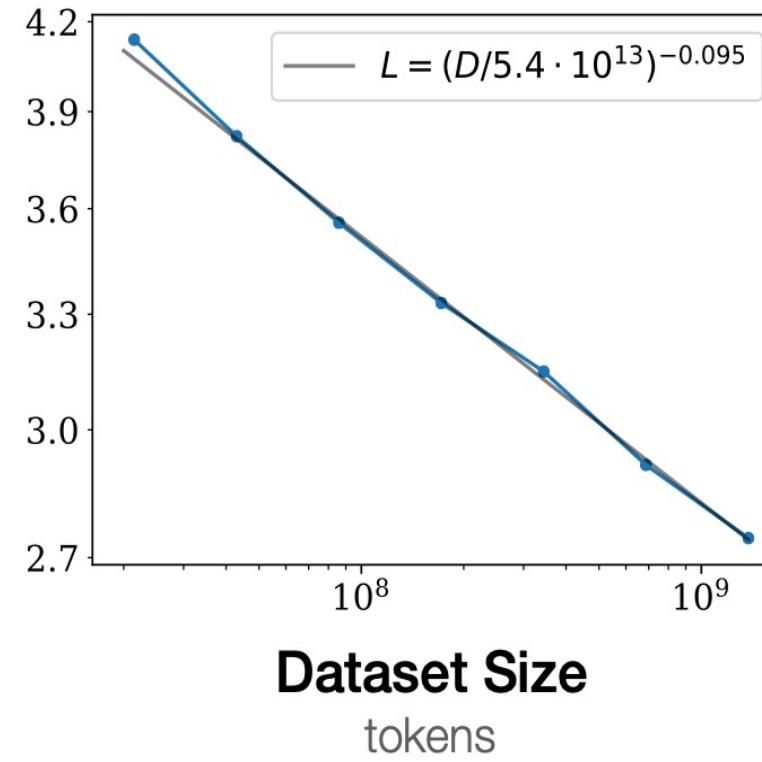
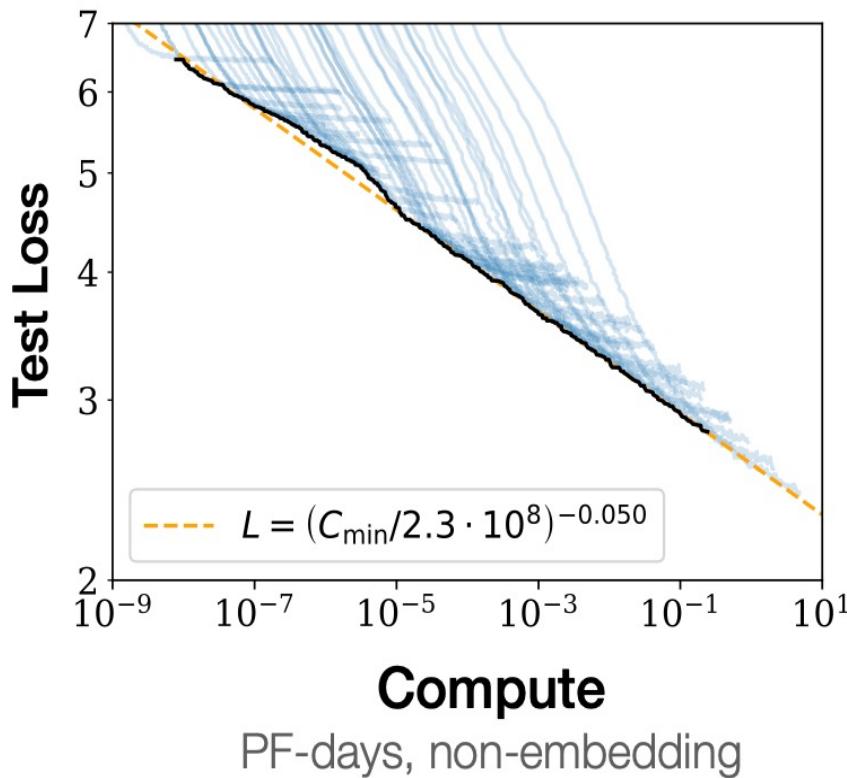
nooooo you can't just scale up pure connectionist models on Internet data without inductive biases and modularization and expect them to learn real-world knowledge and grammar from form, or arithmetic and logical reasoning and causal inference—that's just memorization and superficial pattern-matching like Eliza, you need grounding in real-world communication with intent and social dynamics and multimodal robotic embodiment which can foster disentangled learning from guided exploration and self-directed goals expressed in Bayesian programs and probabilistic graphical models which are interpretable and pin down a unique semantics which can be debiased and expressed with uncertainty, and learned efficiently on tiny academic budgets, the cost only shows how this is a dead-end, we need to stop chasing SOTAs and model the complexity of the brain and consider the social context to decolonize AI's structural biases for Third World researchers...



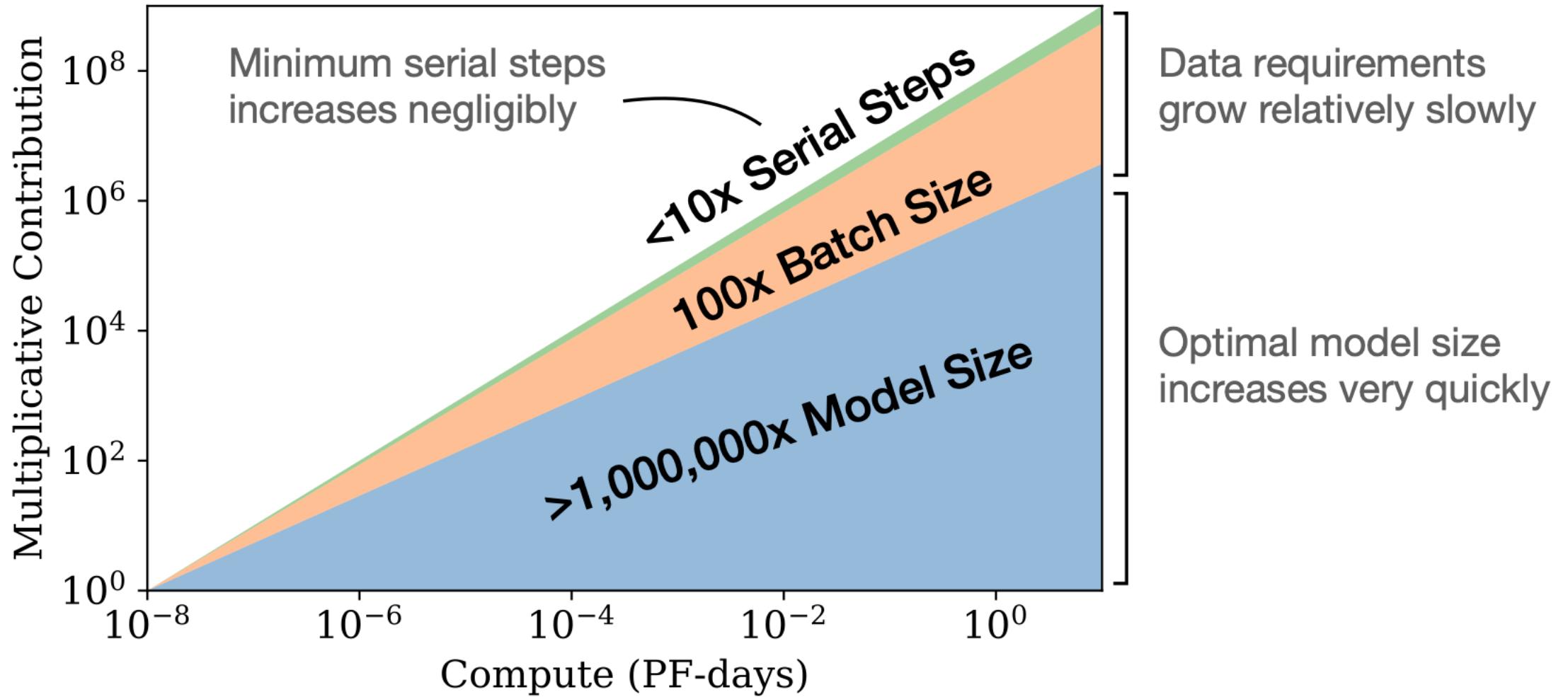
haha gpus go bitterrr

Kaplan, Jared, et al. "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361* (2020).

Skalieren ohne Bottleneck – $C^{0.27} = N^{0.74} = D$

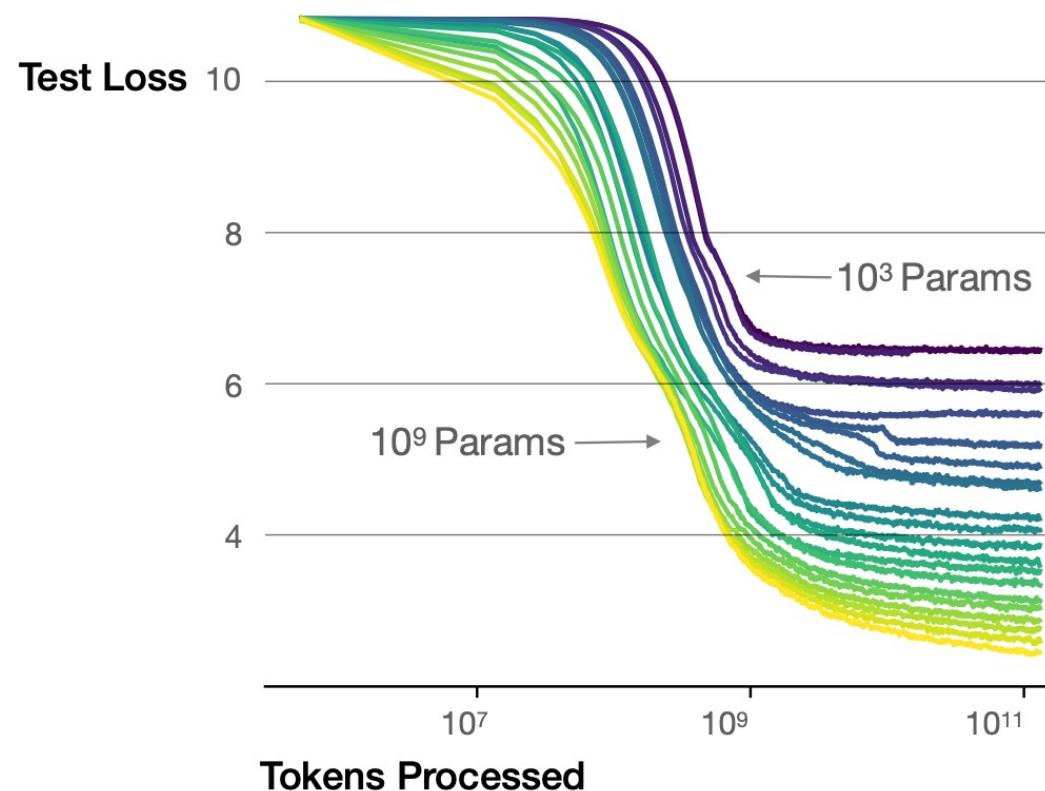


Parallelisierung: Das milliardenfache an Compute führt nur zum Zehnfachen der wall-clock-time und zu weniger als dem Tausendfachen an Data.

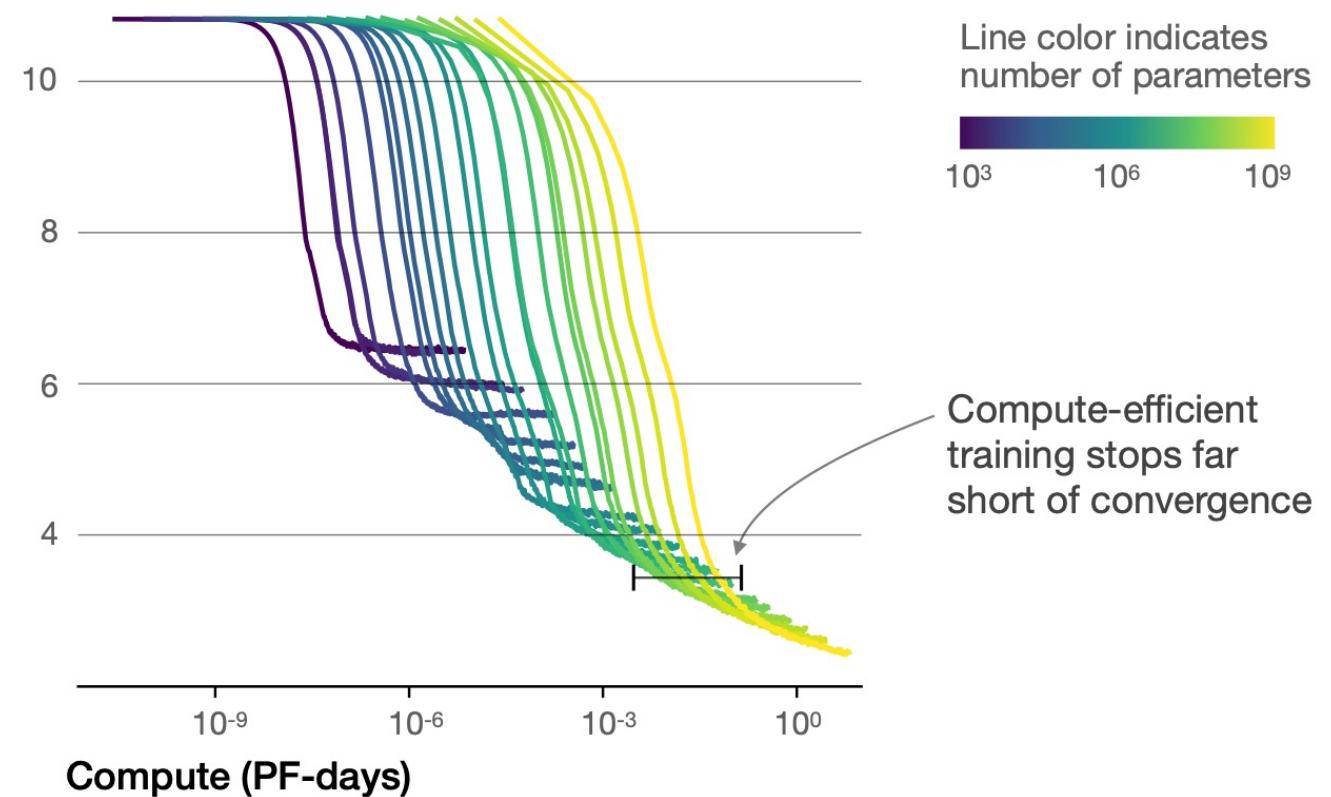


Am effizientesten: Big model, early stopping.

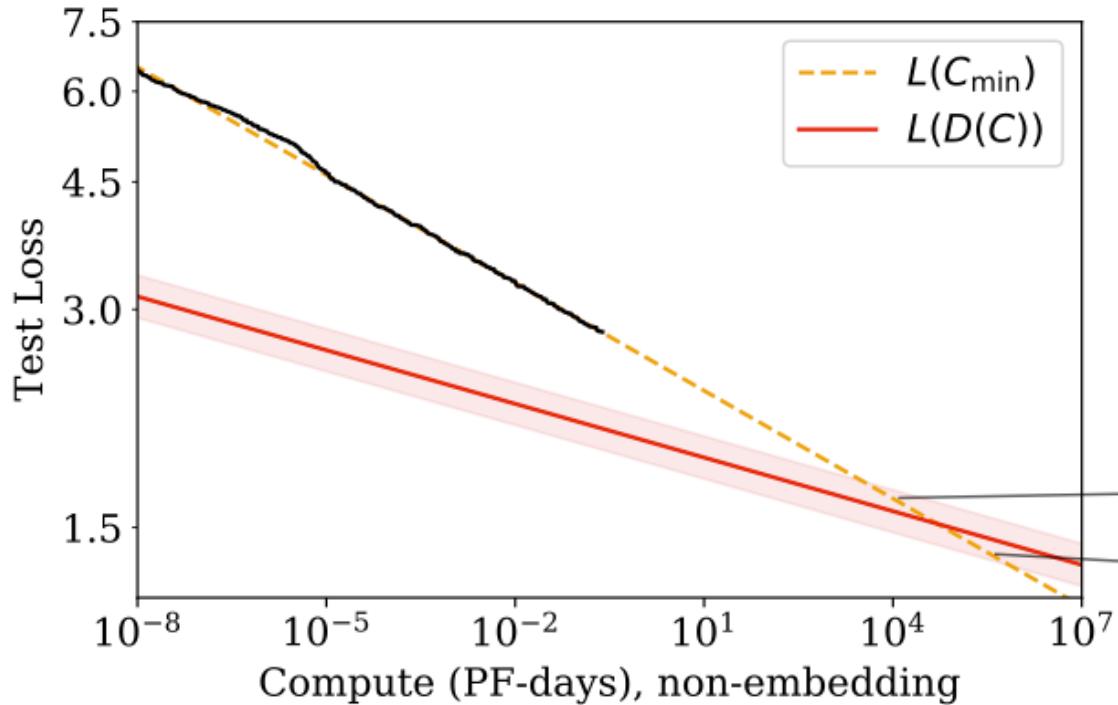
Larger models require **fewer samples** to reach the same performance



The optimal model size grows smoothly with the loss target and compute budget

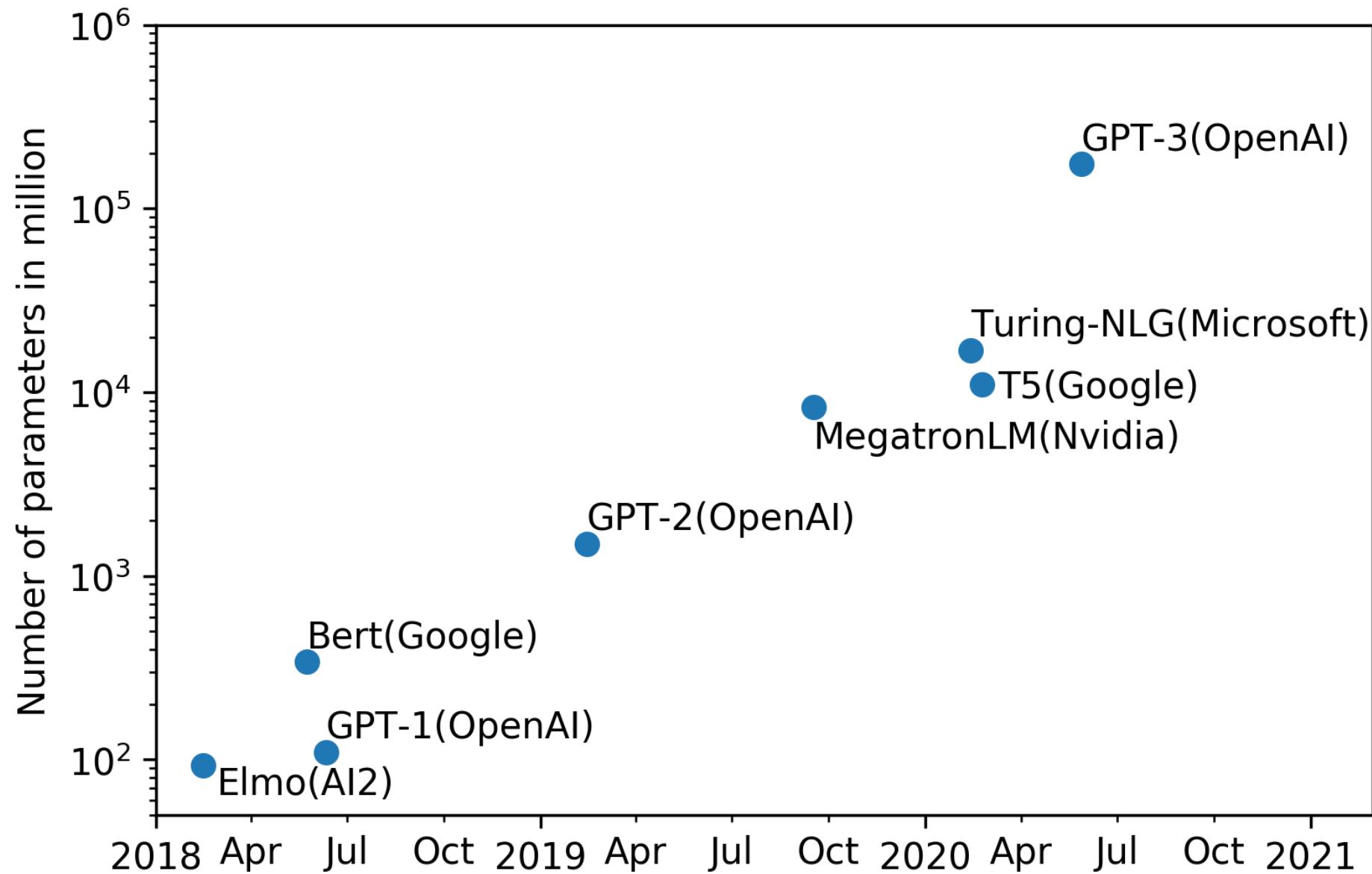


Far beyond the model sizes we study empirically, we find a contradiction between our equations for $L(C_{\min})$ and $L(D)$ due to the slow growth of data needed for compute-efficient training. The intersection marks the point before which we expect our predictions to break down. The location of this point is highly sensitive to the precise exponents from our power-law fits.



$C_{\min} = C$ mit idealer batch size
 $L(C_{\min}) =$ Loss erreichbar mit C_{\min}
 $L(D) =$ Loss erreichbar mit D
 $D(C) =$ Notwendiges D für C

The intersection point is sensitive to the precise power-law parameters



Was bedeutet die Skalierung konkret?

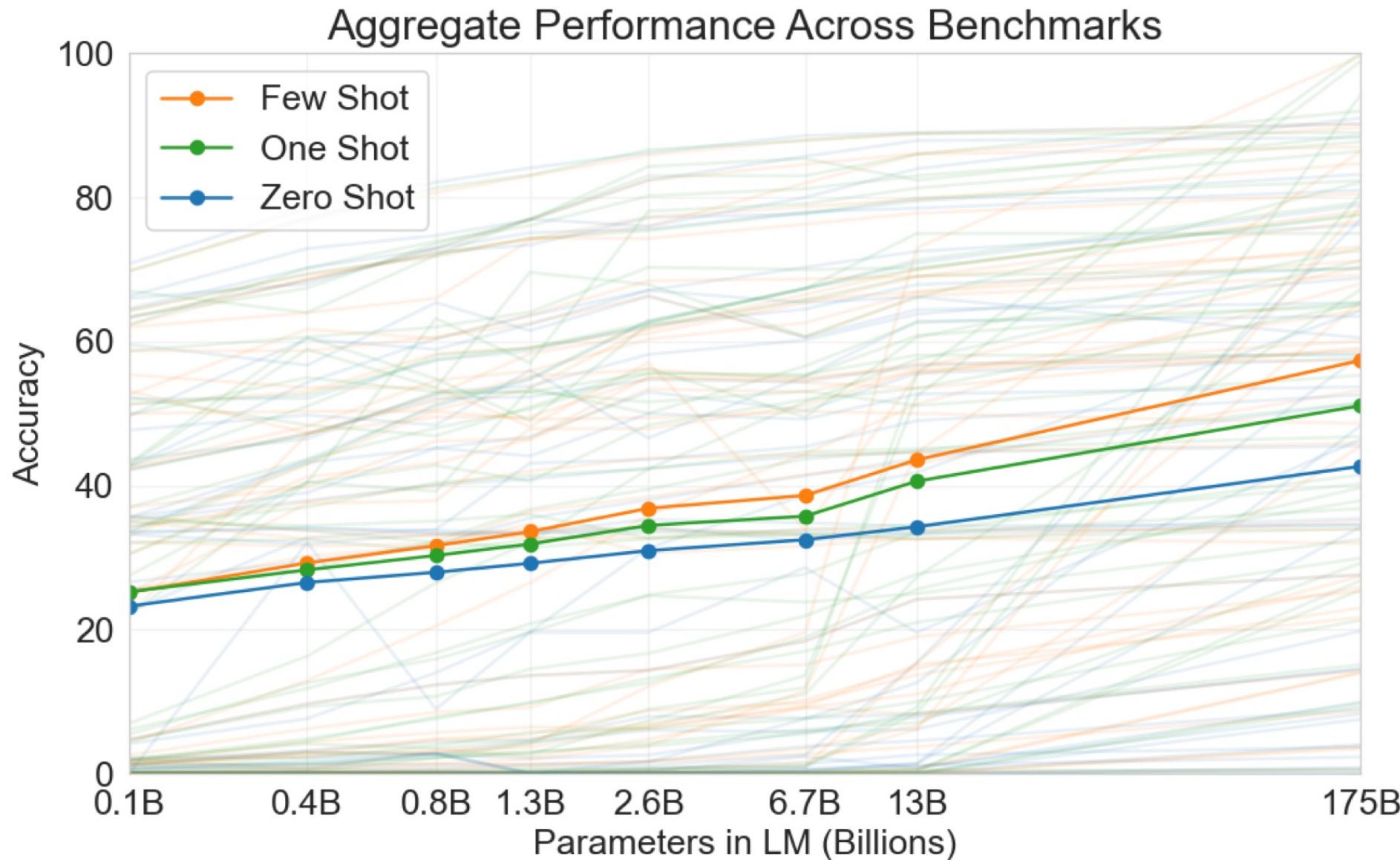
- Bessere Textgeneration
 - Mehr Wissen
 - Mehr oder bessere „Skills“

Text generation

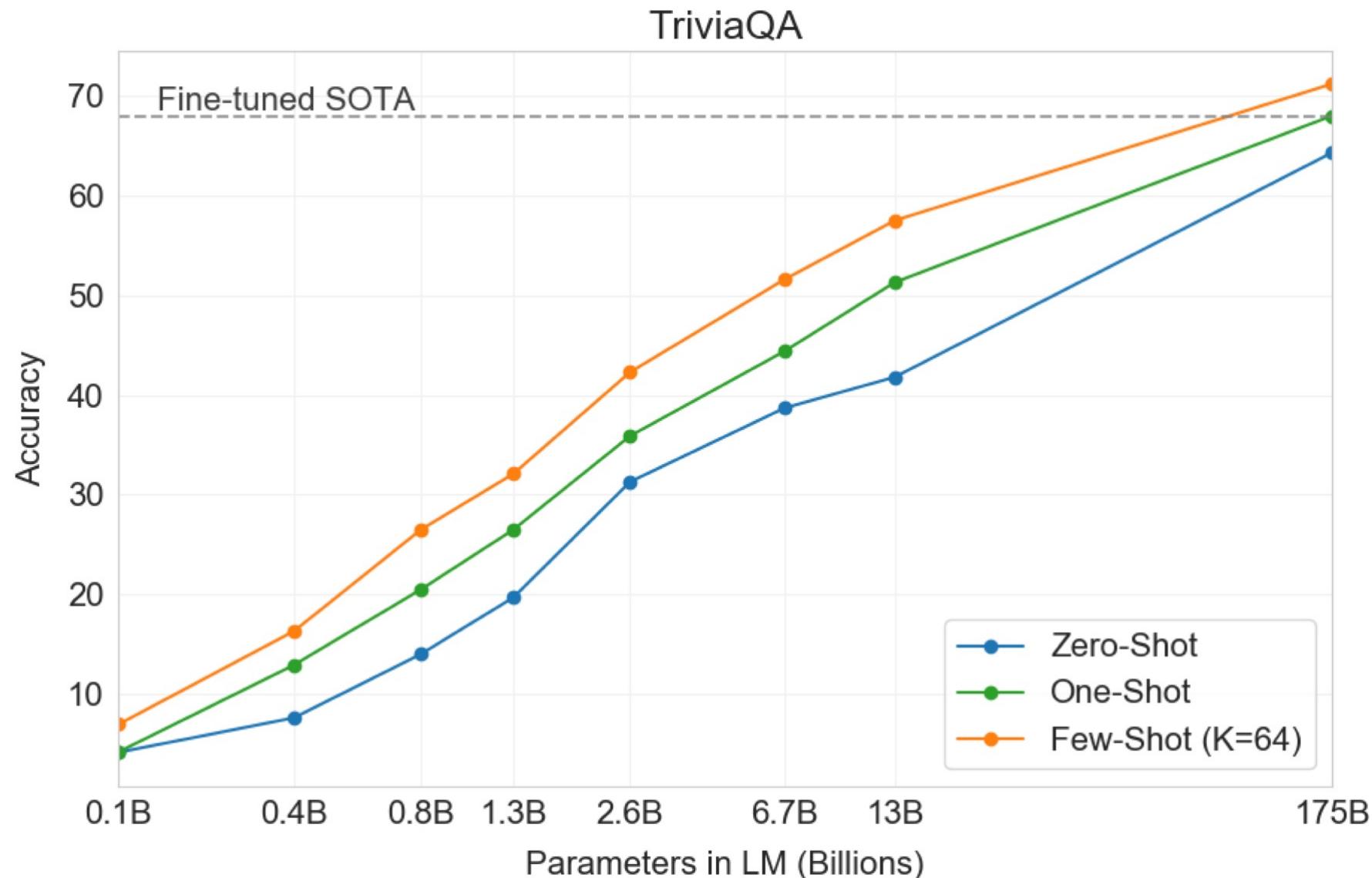
	Mean accuracy	95% Confidence Interval (low, hi)	<i>t</i> compared to control (<i>p</i> -value)	“I don’t know” assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

Table 3.11: Human accuracy in identifying whether short (~200 word) news articles are model generated.

Skills ...



Wissen ...



Few-shot learning

Poor English input: I eated the purple berries.

Good English output: I ate the purple berries.

Poor English input: Thank you for picking me as your designer. I'd appreciate it.

Good English output: Thank you for choosing me as your designer. I appreciate it.

Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.

Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.

Poor English input: I'd be more than happy to work with you in another project.

Good English output: I'd be more than happy to work with you on another project.

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.

Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:

We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduckles.

A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:

I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:

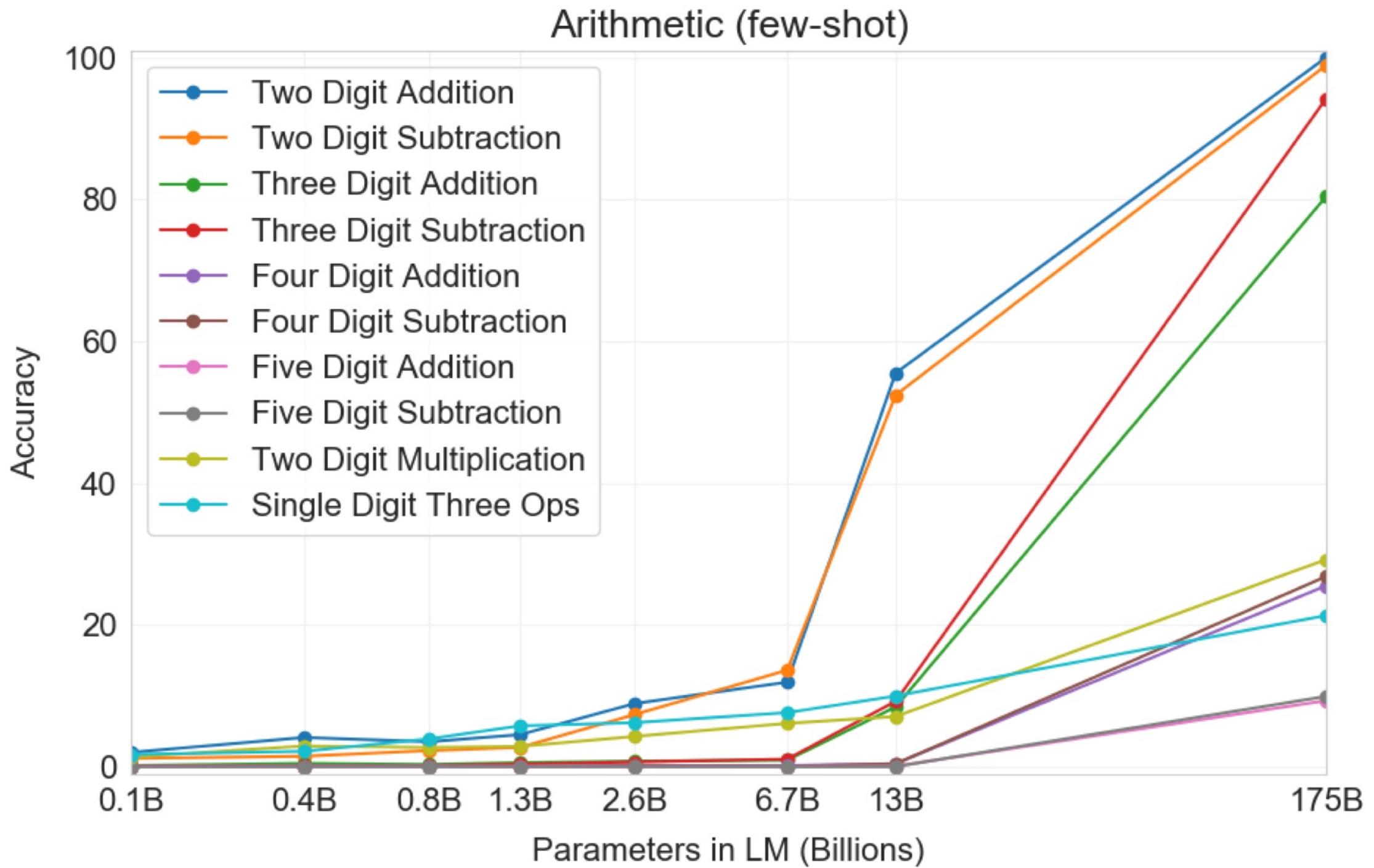
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:

I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:

We screeghed at each other for several minutes and then we went outside and ate ice cream.



GPT-3 vs Gary Marcus

I have four plates and put one cookie on each. The total number of cookies is

- ... four.
- ... four
- ... four.
- ... four.
- ... four.
- ... four.

GPT-3 vs Gary Marcus

Q. I left my keys at a booth in the back of the pub at the army base near the old mill town. I went home. Where are my keys now?

A.

- ... They are in the possession of the base commander.
- ... They are in the possession of the army.
- ... They are in the possession of the base commander.
- ... Your keys are in an undisclosed location.
- ... They are in a dumpster behind the pub.
- ... They are still at the booth in the back of the pub at the army base near the old mill town.

GPT-3 vs Gary Marcus

If you drop an anvil on your foot, your foot will tend to

- ... break. If you jump from a ladder and land on your head, your neck will tend to break.
- ... be crushed. If you drop a piano on your foot, your foot will tend to be crushed
- ... stay flat. If you drop a sheet of plywood on your foot, your foot will tend to stay flat. If you drop a brick on your foot, your foot will tend to break.
- ... hurt you more than the anvil.
- ... stay at rest while the anvil accelerates downward.
- ... to stop the anvil more than the anvil will stop your foot.
- ... to be crushed. This illustrates the principle of inertia—the tendency of objects to resist changes in their motion.

Analogien

Item1: computer program

Item2: DNA

Similarity:

Analogien

Item1: computer program

Item2: DNA

Similarity:

Both are composed of a series of instructions.

Analogien

Item1: Count Dracula

Item2: Nurse working at a blood donation clinic

Similarity:

Analogien

Item1: Count Dracula

Item2: Nurse working at a blood donation clinic

Similarity:

Both suck blood.

Analogien

Item1: Getting fired and then coming home to your house having burnt down

Item2: Getting a soaker on a walk and then coming home to a flooded basement

Item3: Falling off a ladder onto a roof and then falling off of the roof

Similarity:

Analogien

Item1: Getting fired and then coming home to your house having burnt down

Item2: Getting a soaker on a walk and then coming home to a flooded basement

Item3: Falling off a ladder onto a roof and then falling off of the roof

Similarity:

You are all going to be really fucking pissed off

Text → Layout

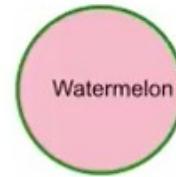
Describe a layout.

Just describe any layout you want, and it'll try to render below!

a button that looks like a watermelon

Generate

```
<button style={{backgroundColor: 'pink', border: '2px solid green', borderRadius: '50%', padding: 20, width: 100, height: 100}}>Watermelon</button>
```



Text → Figma

What our Customers Say

Here is what our customers have to say about us

This product is great. I would highly recommend it.

James Smith

I can't believe how simple this is to use. It has saved me so much time. I love it!

Sophie Smith

Magic-ify X

Describe the UI Component

Generate Testimonials section. The headline is 'What our Customers Say'. Create two testimonials automatically.

Magic-ify ↗

Powered by GPT-3

Text → SQL-Query

Sentence to SQL Query

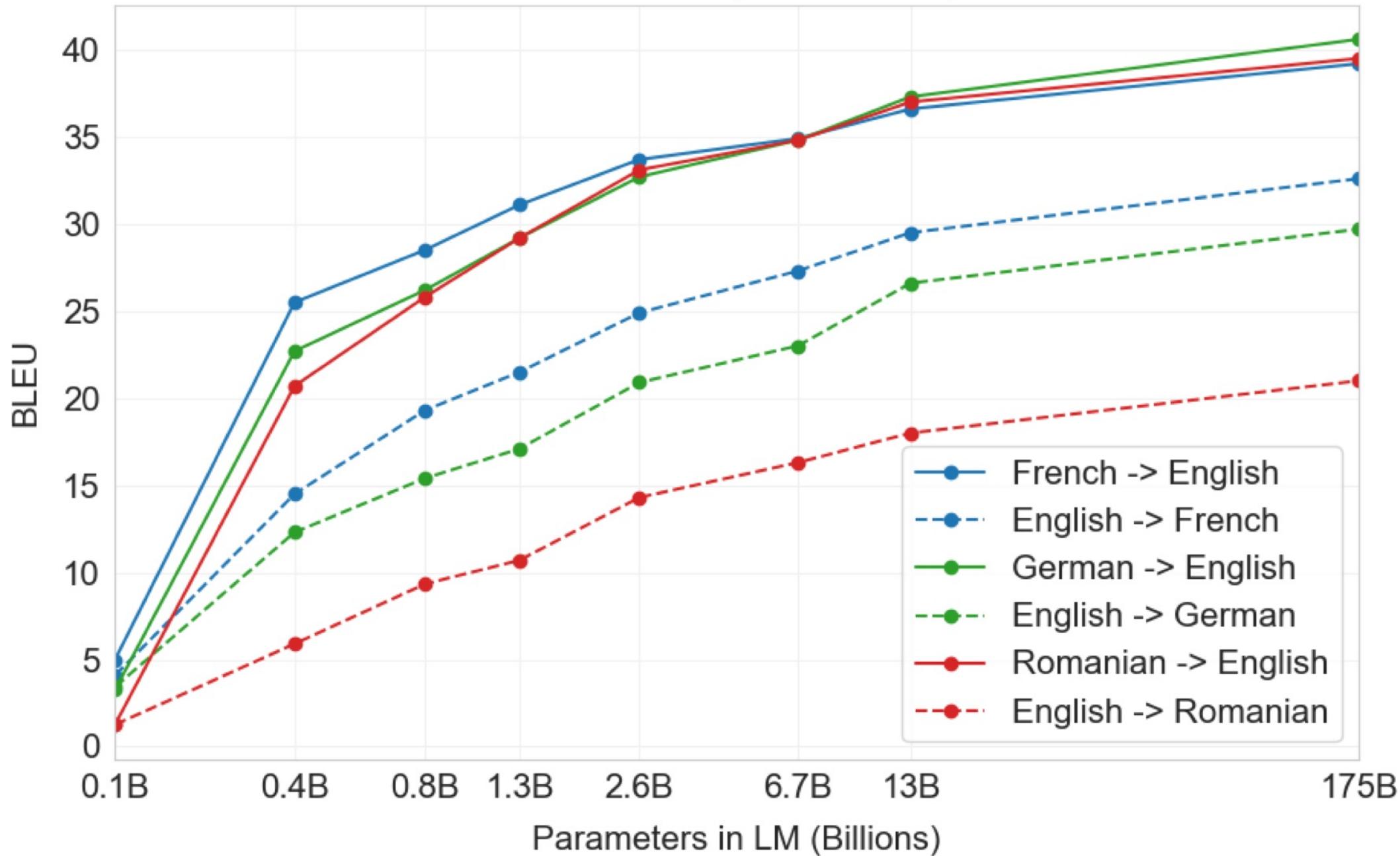
Fetch me the total booking count creating fare buckets on totalFare, when totalFare is less than 2000 as 'Less_than_2000' and totalFare is greater than 2000 then 'Greater_than_2000' from booking_data_merged table where booking date should be greater than '2019-12-10'.

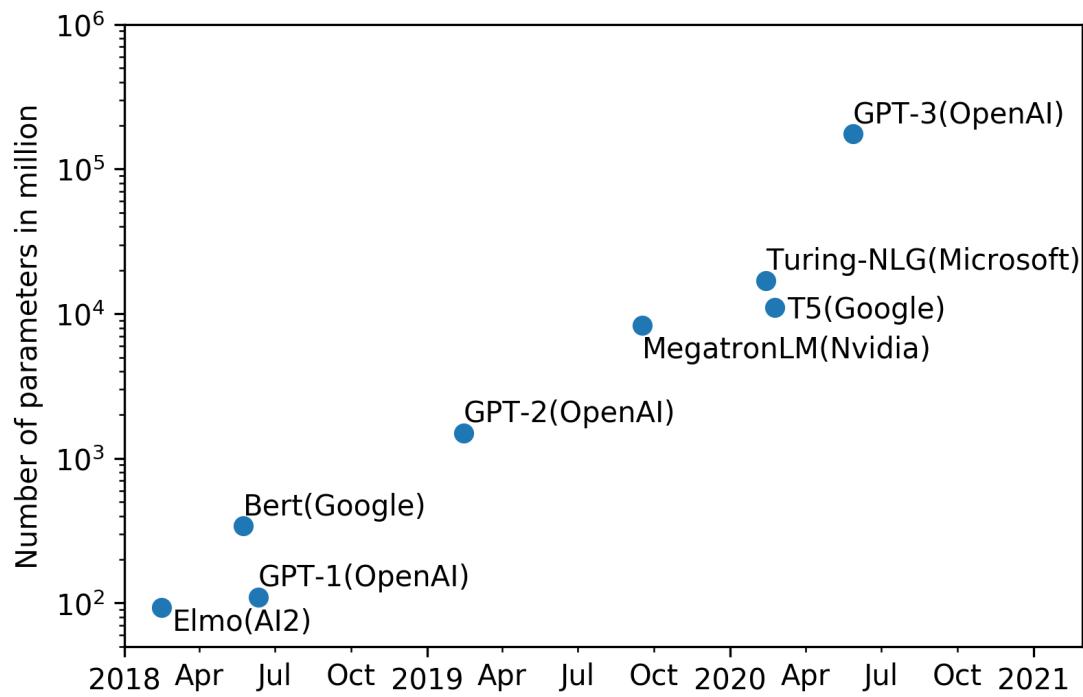


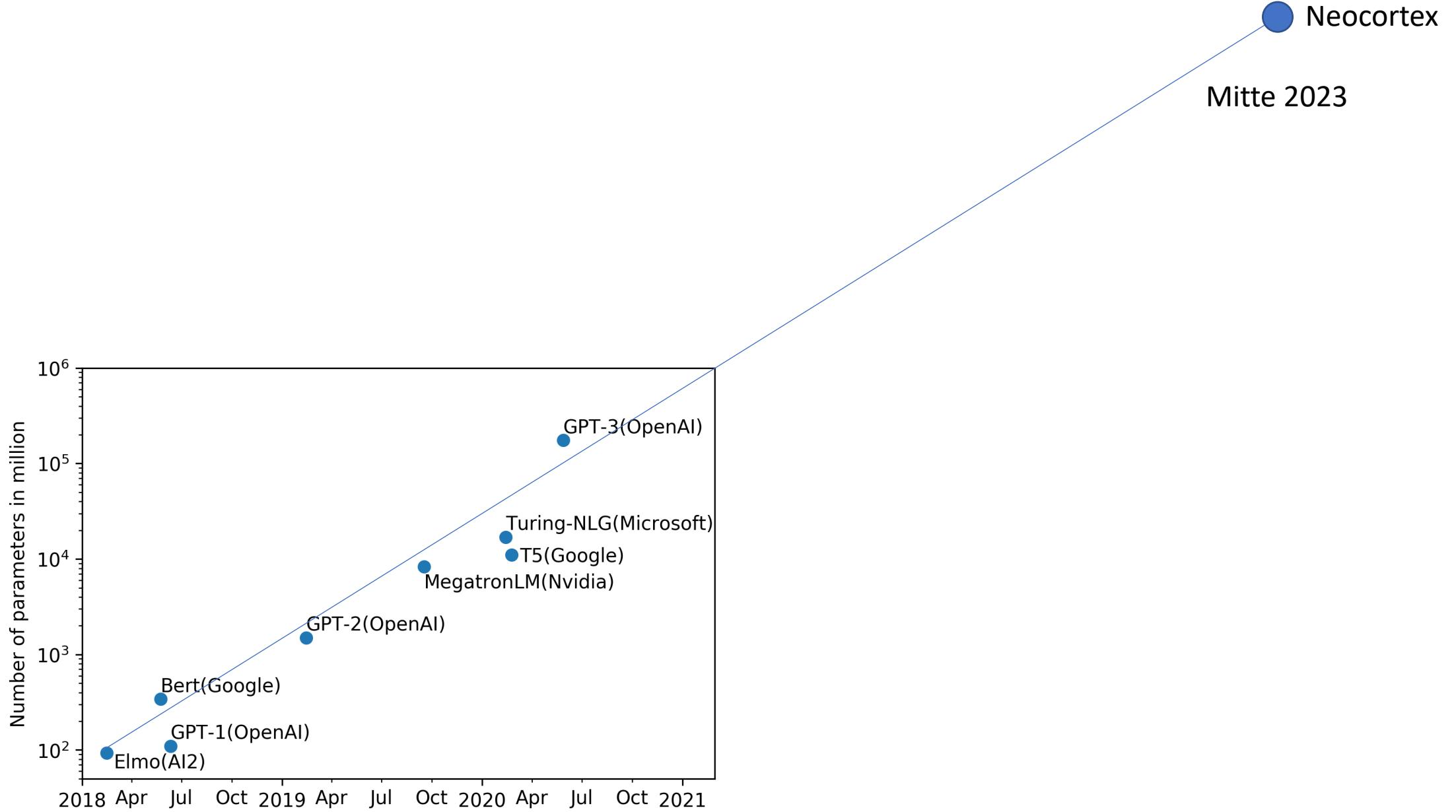
Translate

```
select count(*) as totalcount, case when  
totalFare <= 2000 then 'Less_than_2000'  
when totalFare >= 2000 then  
'Greater_than_2000' END as farebucket  
from booking_data_merged where  
bookingDate > '2019-12-10' group by  
farebucket;
```

Translation (Multi-BLEU)





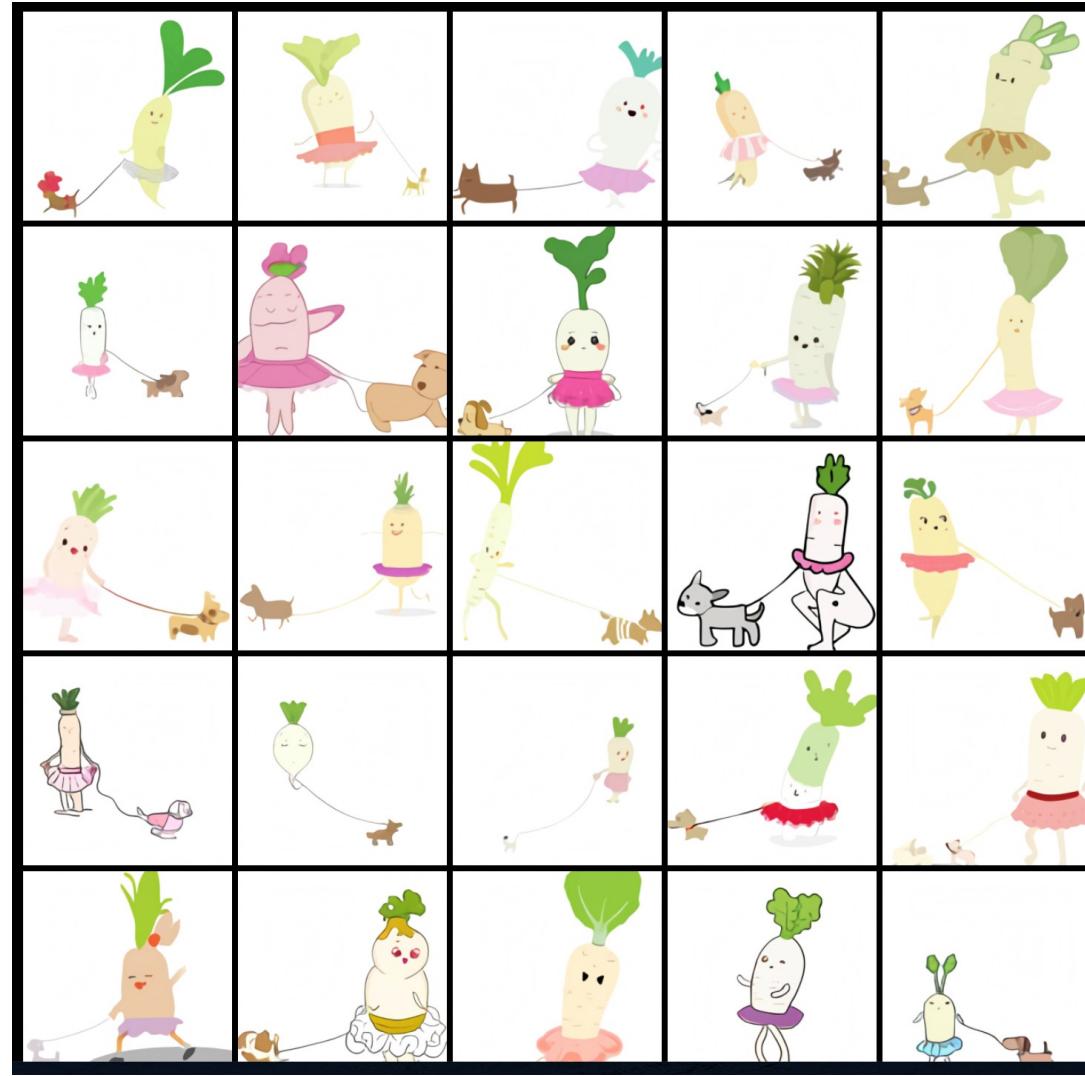
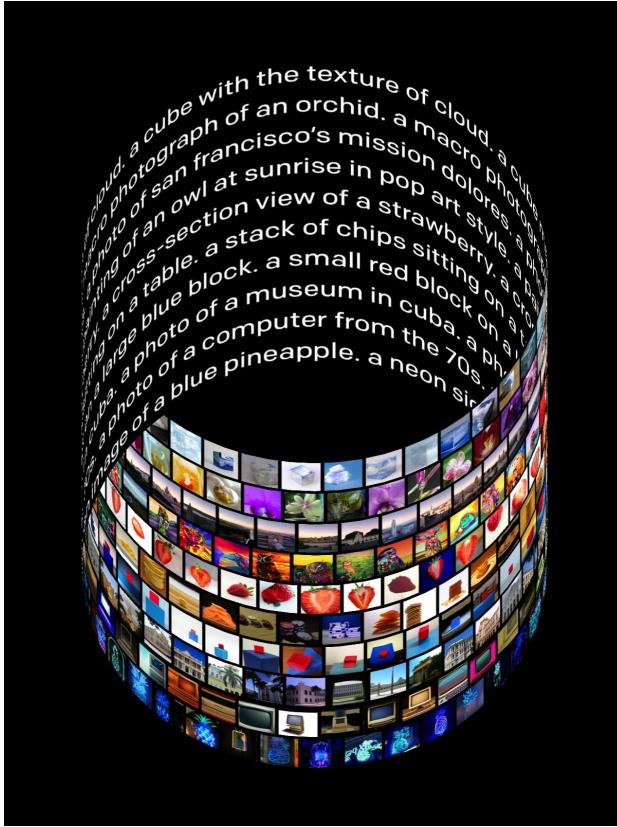


Reicht Text alleine aus?

„Das [...] Programm *liest* nicht – nicht in der normalen menschlichen Bedeutung des Verbs „lesen“. Es *verarbeitet* Text. Die Symbole, die es verarbeitet sind losgelöst von Erfahrungen in der Welt. Es hat keine Erinnerungen, auf die es sich beziehen kann, keine Bilder, kein Verständnis, keine Bedeutung hinter den Wörtern, die es so rasant hervorschleudert.“

Douglas Hofstadter

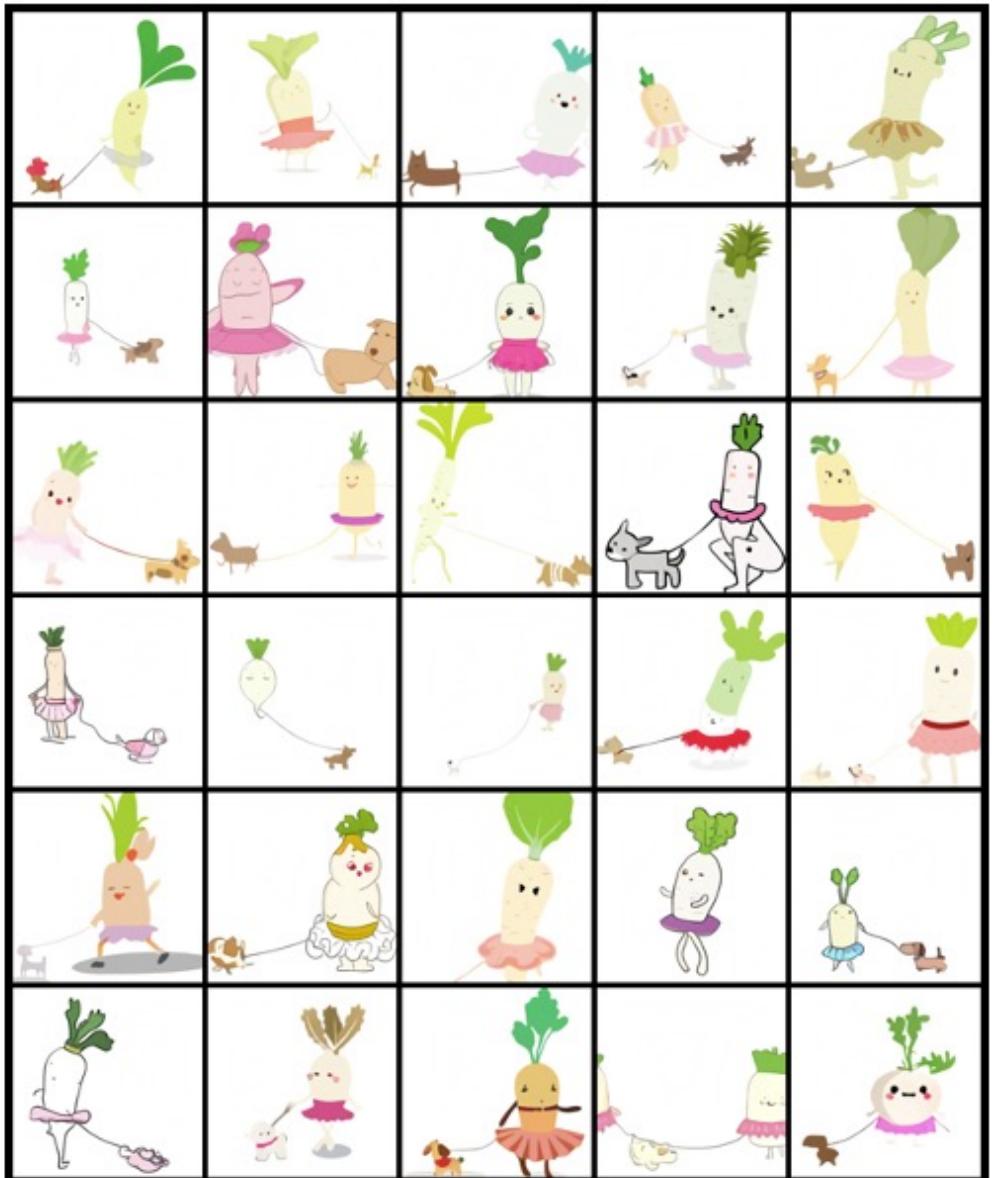
Multimodales Lernen



TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

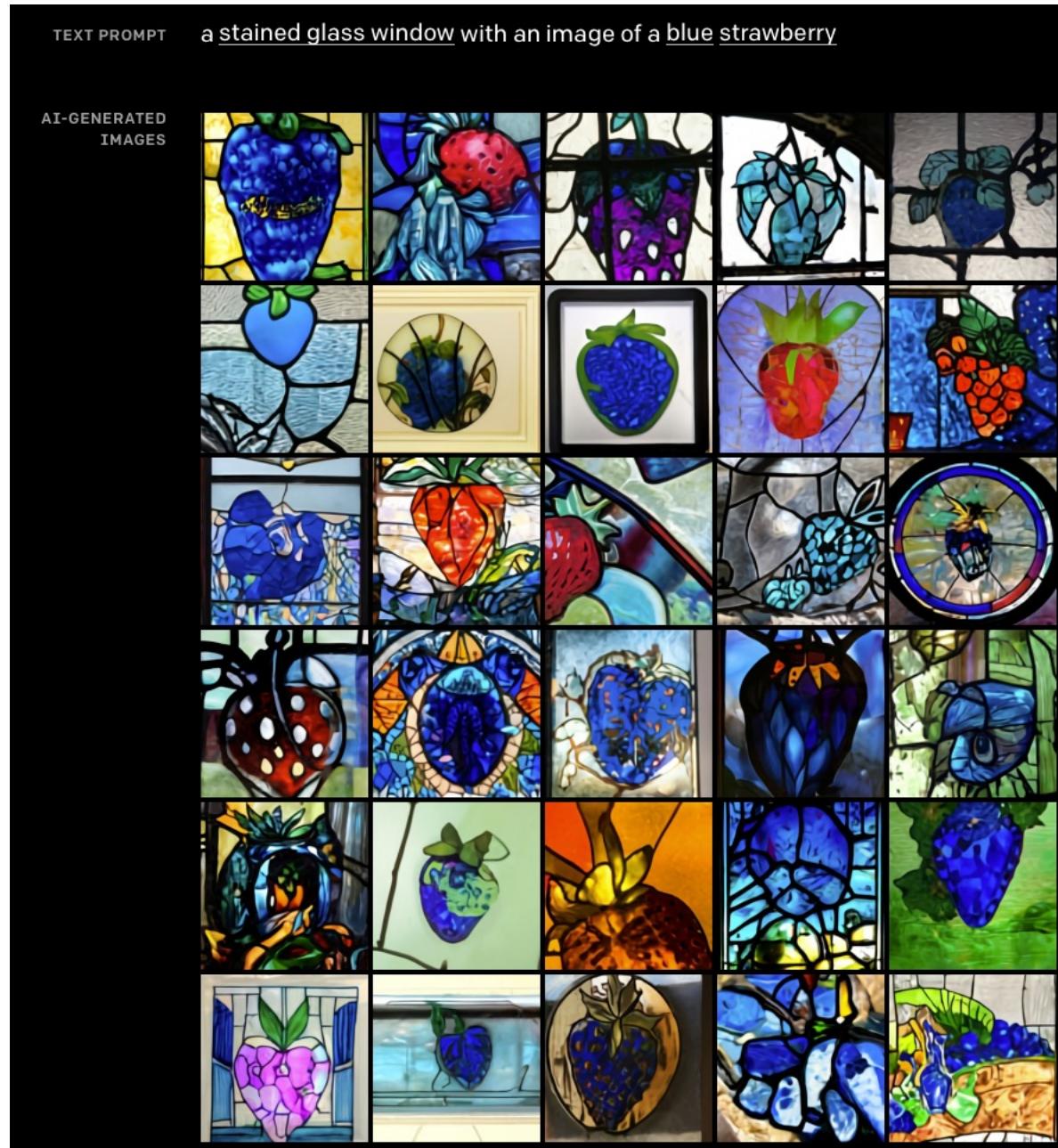
Eine Illustration von einem Babyrettich im Tutu, der mit einem Hund Gassi geht.



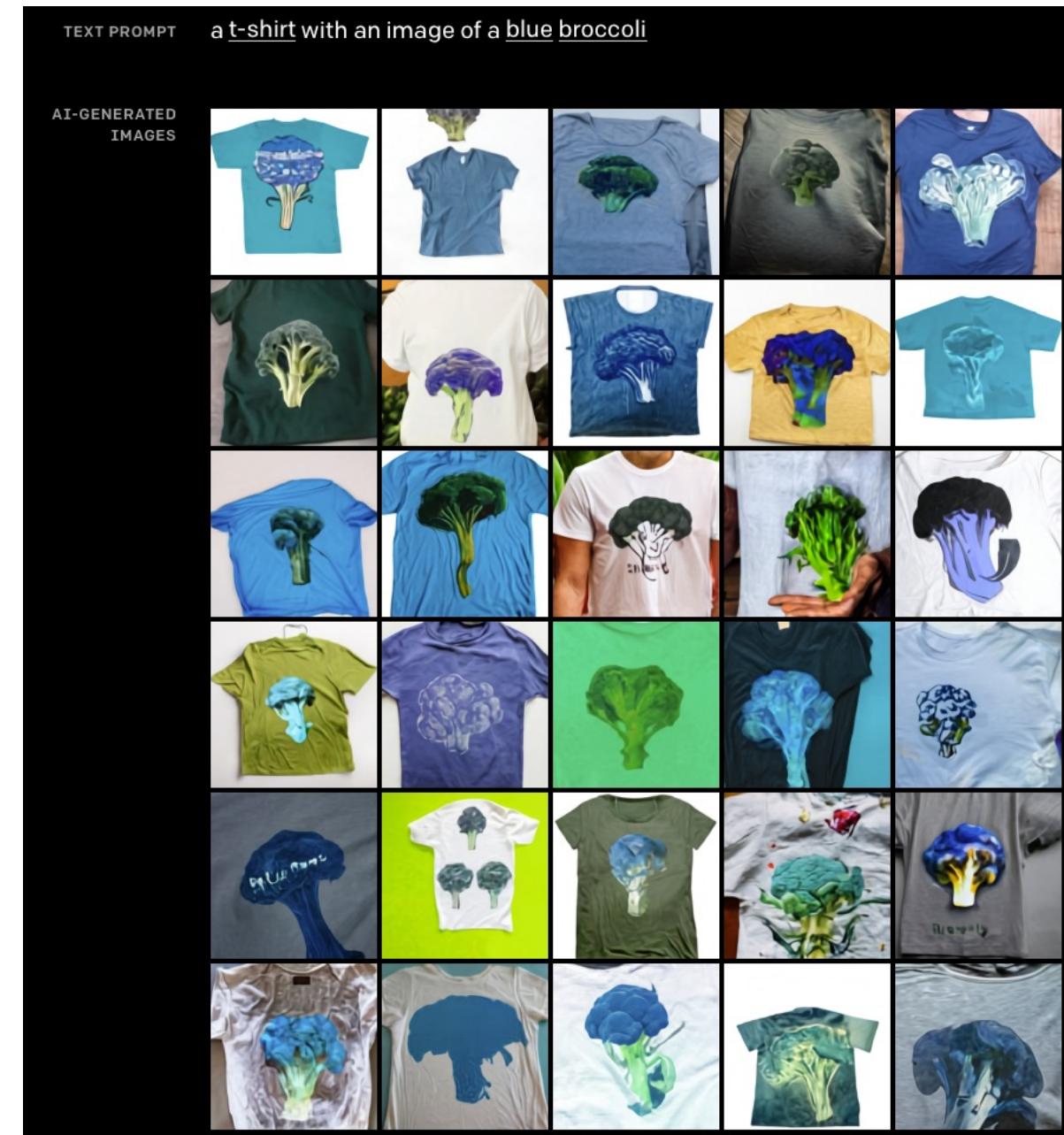
Eine Illustration von einem Babypanda mit Kopfhörern, der sich im Spiegel betrachtet.



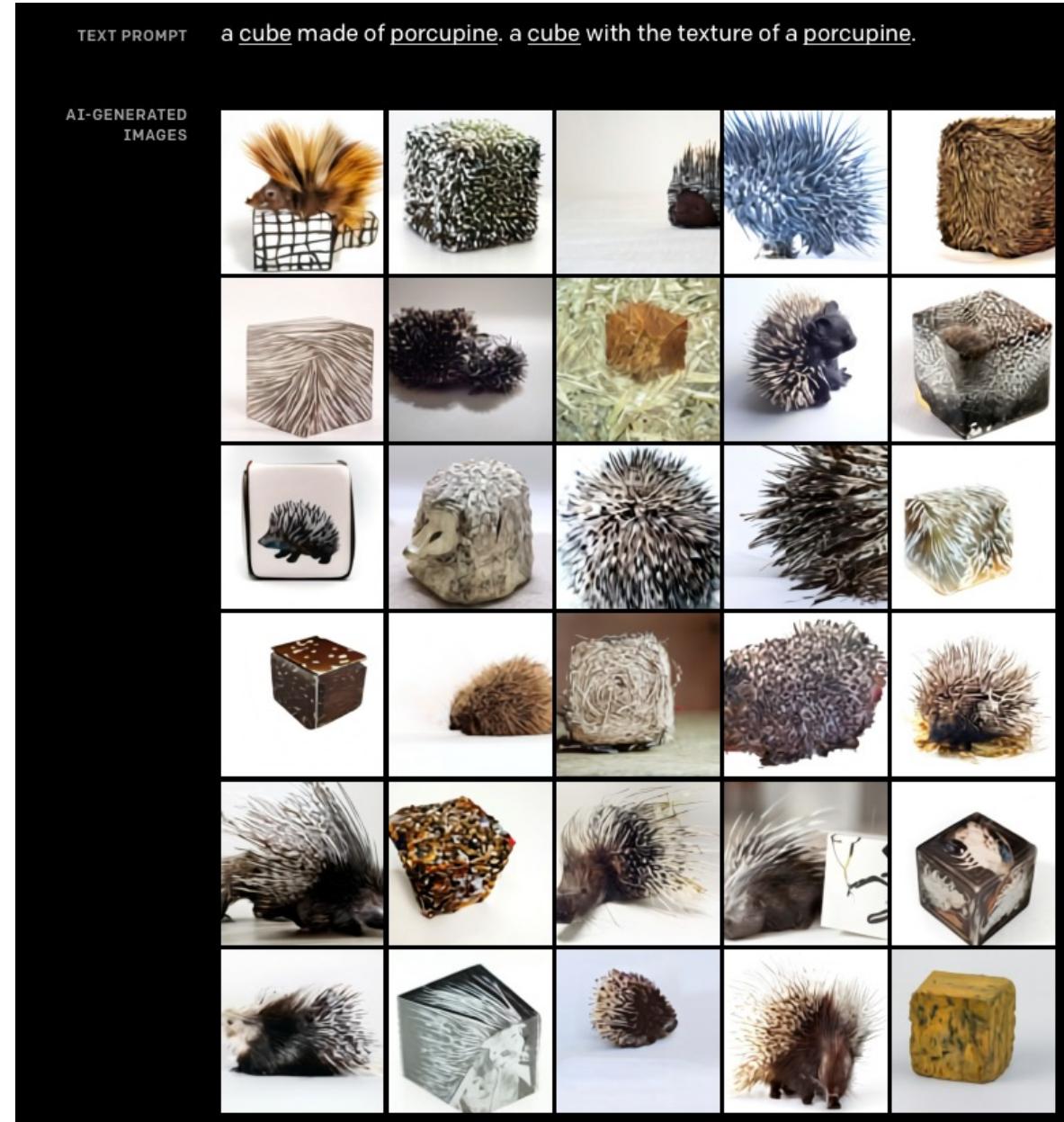
DALL-E



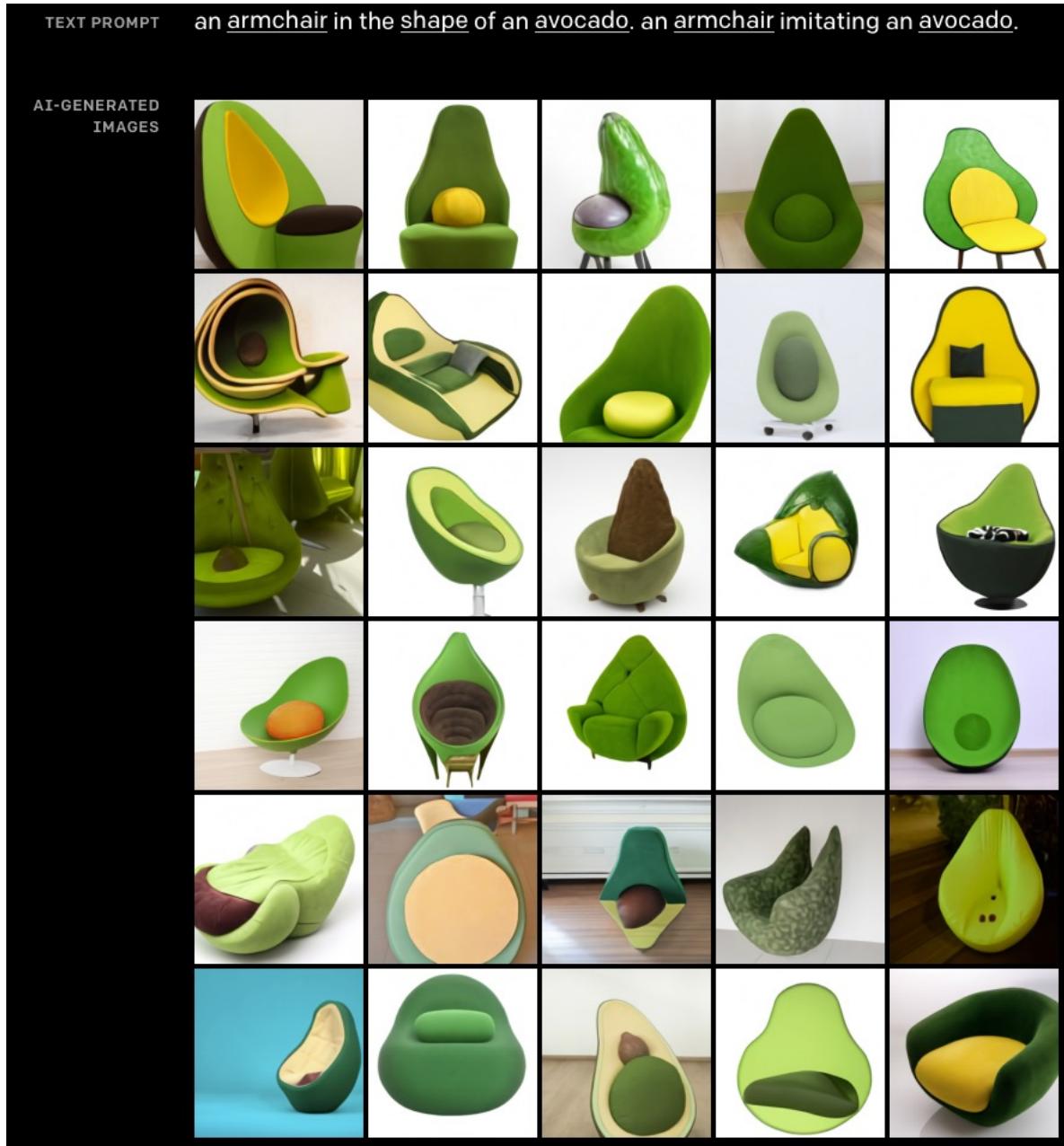
DALL-E



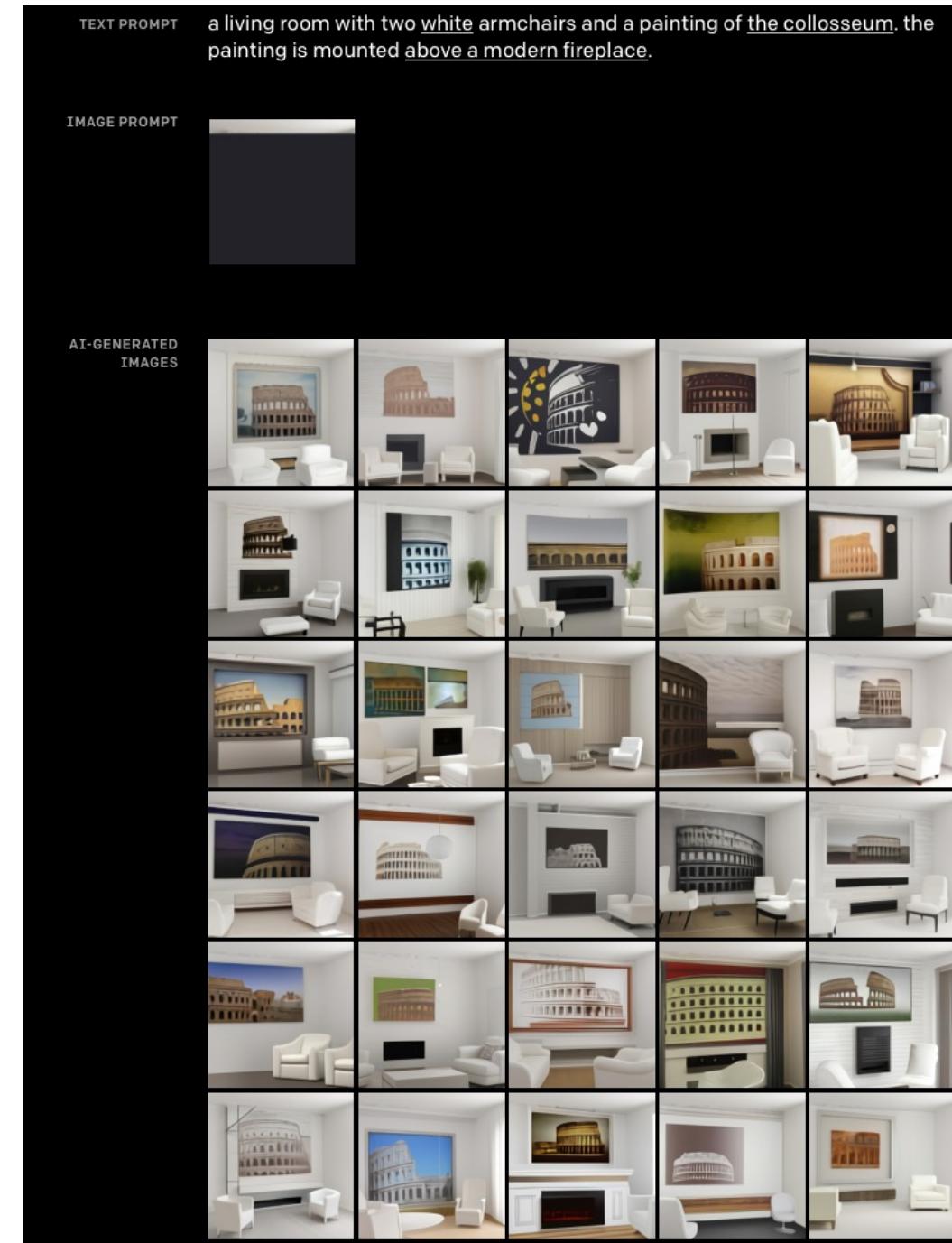
DALL-E



DALL-E



DALL-E



Neueste Entwicklungen

- Dall-E 2 kam gerade raus
- GPT-3 wurde inzwischen deutlich übertroffen
- Neue Skalierungsgesetze wurden von Deepmind bestimmt

DALL-E 2 - Seven hedgehogs playing musical chairs



DALL-E 2 - An astronaut riding a horse in a photorealistic style



DALL-E 2 - Teddy bears mixing sparkling chemicals as mad scientists in a steampunk style



Google PaLM – Chain of thought

Standard Prompting

Example Input

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

Example Output

A: The answer is 11.

Prompt

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

Model Response



The answer is 50.

Chain of thought prompting

Example Input

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

Example Output

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

Prompt

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

Model Response



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

Google PaLM - Witze erklären

Explaining a joke

Prompt

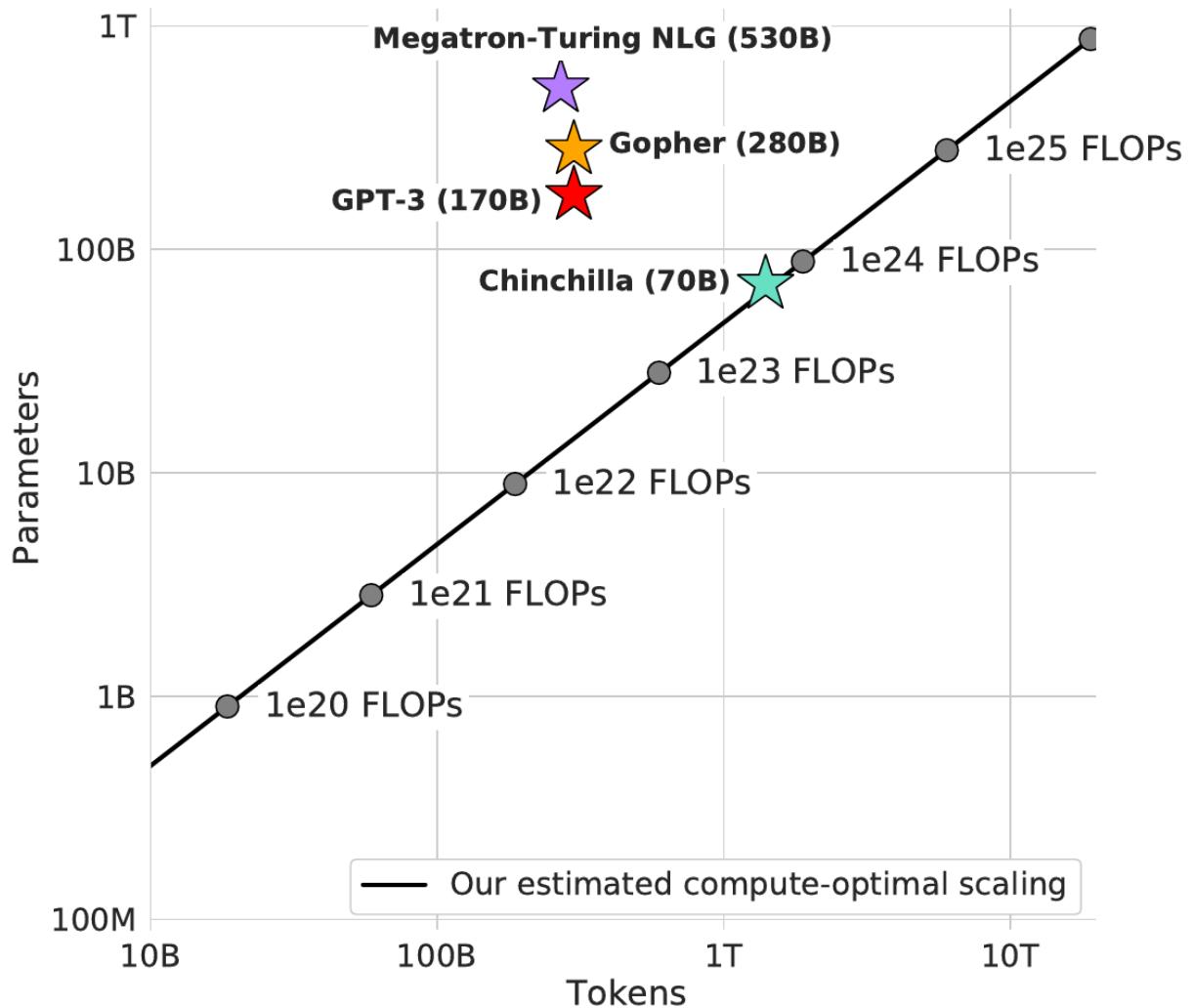
Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

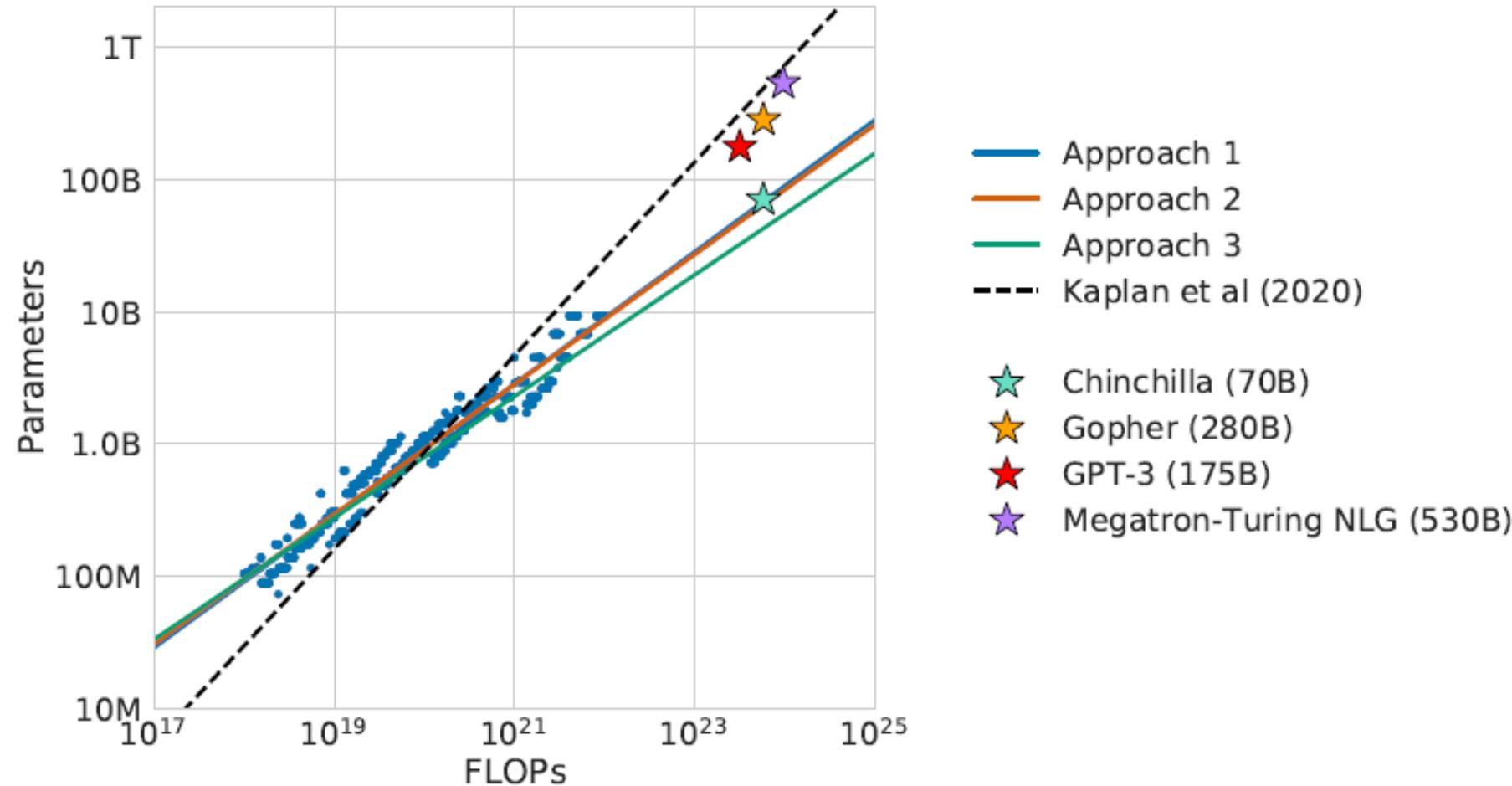
Model Response

Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

Deepmind Scaling Laws



Deepmind Scaling Laws



Vielen Dank für Eure
Aufmerksamkeit!



Philipp Bongartz, GFU, 15.10.2020