OpenAI’s GPT-3 Language Model

Steve Omohundro, Ph.D.

Possibility Research

https://link.springer.com/chapter/10.1007/978-3-642-15760-8_23

GPT-3 is a “Language Model”

Assigns probabilities to word sequences: $P(s) \;=\; P(w_1, \ldots, w_m)$

Can factor as the product:

$$P(s) = \prod_{i=1\ldots m} P(w_i | w_1, \ldots, w_{i-1})$$

E.g. (n-1)-gram model:

$$P_{ngram}(s) = \prod_{i=1\ldots m} P(w_i | \mathbf{h}_i) = \prod_{i=1\ldots m} P(w_i | w_{i-n}, \ldots, w_{i-1})$$

GPT-3 is a 2048-gram model!

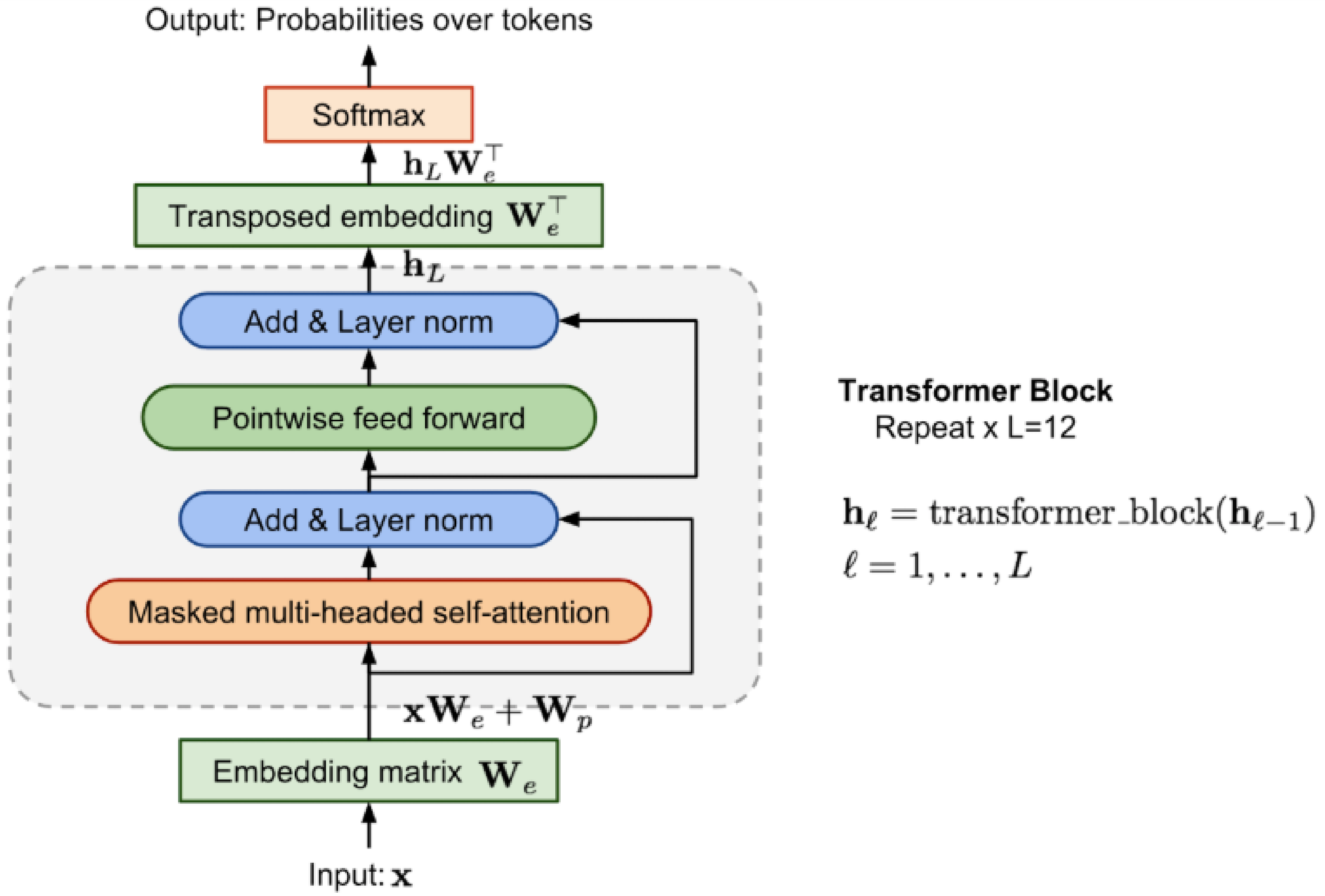
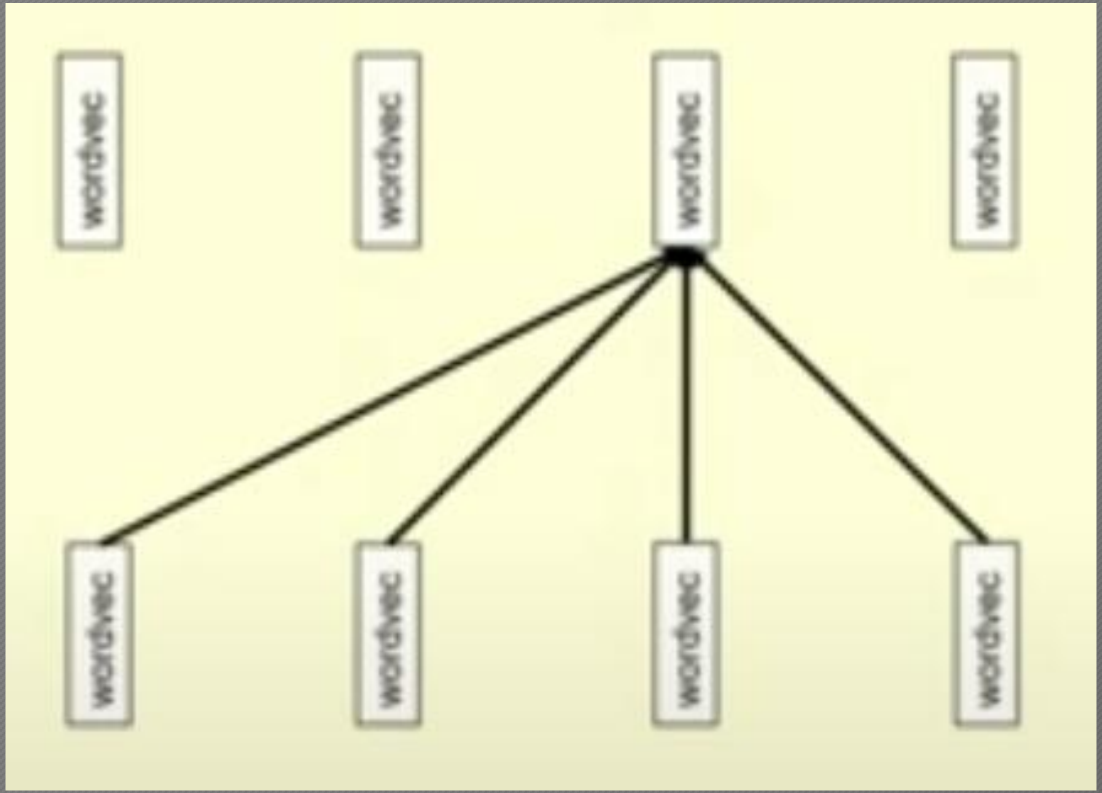
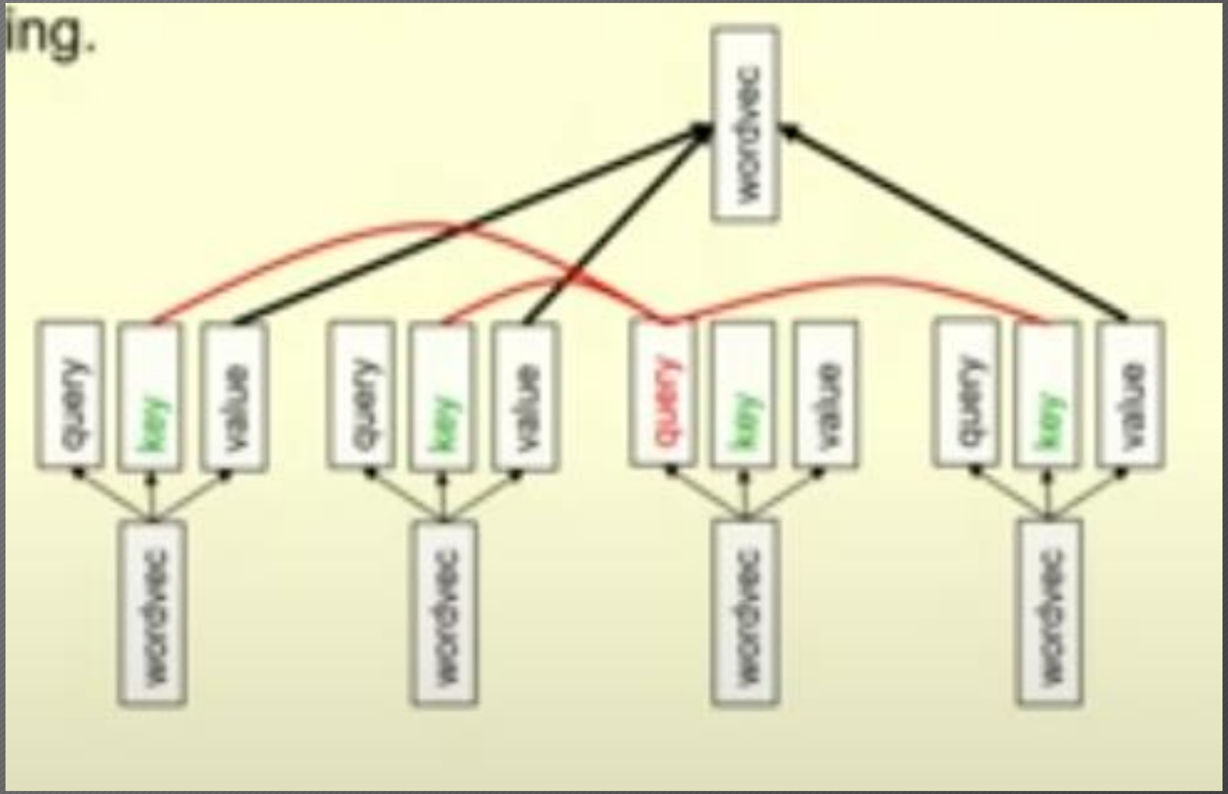

Output: Probabilities over tokens
Softmax
$\mathbf{h}_L \mathbf{W}_e^\top$
Transposed embedding $\mathbf{W}_e^\top$
$\mathbf{h}_L$
Add & Layer norm
Pointwise feed forward
Add & Layer norm
Masked multi-headed self-attention
$\mathbf{x}\mathbf{W}_e + \mathbf{W}_p$
Embedding matrix $\mathbf{W}_e$
Input: $\mathbf{x}$
Transformer Block
Repeat x L=12
$\mathbf{h}_\ell = \text{transformer\_block}(\mathbf{h}_{\ell-1})$
$\ell = 1, \dots, L$

Fig. 7. The transformer decoder model architecture in OpenAI GPT.

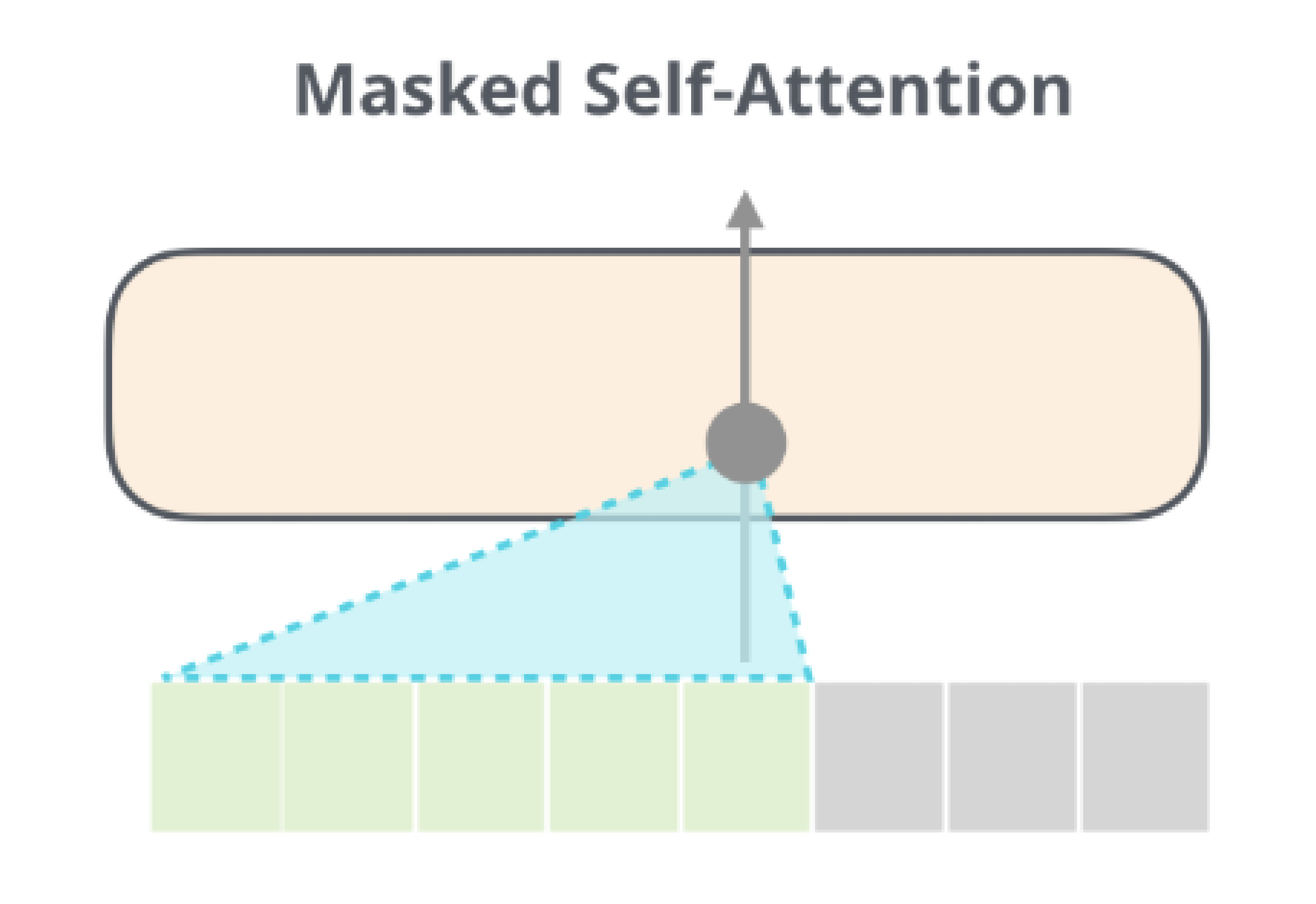
https://lilianweng.github.io/lil-log/2019/01/31/generalized-language-models.html

Convolutional Nets vs. Transformers





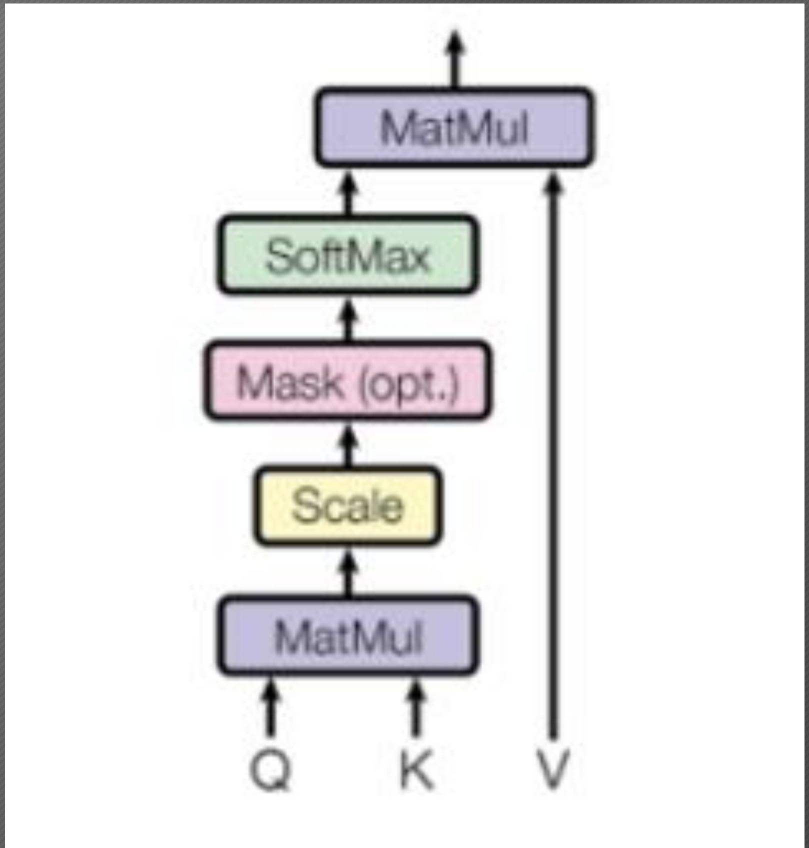
https://www.youtube.com/watch?v=UX8OubxsY8w&feature=youtu.be



http://jalammar.github.io/illustrated-gpt2/

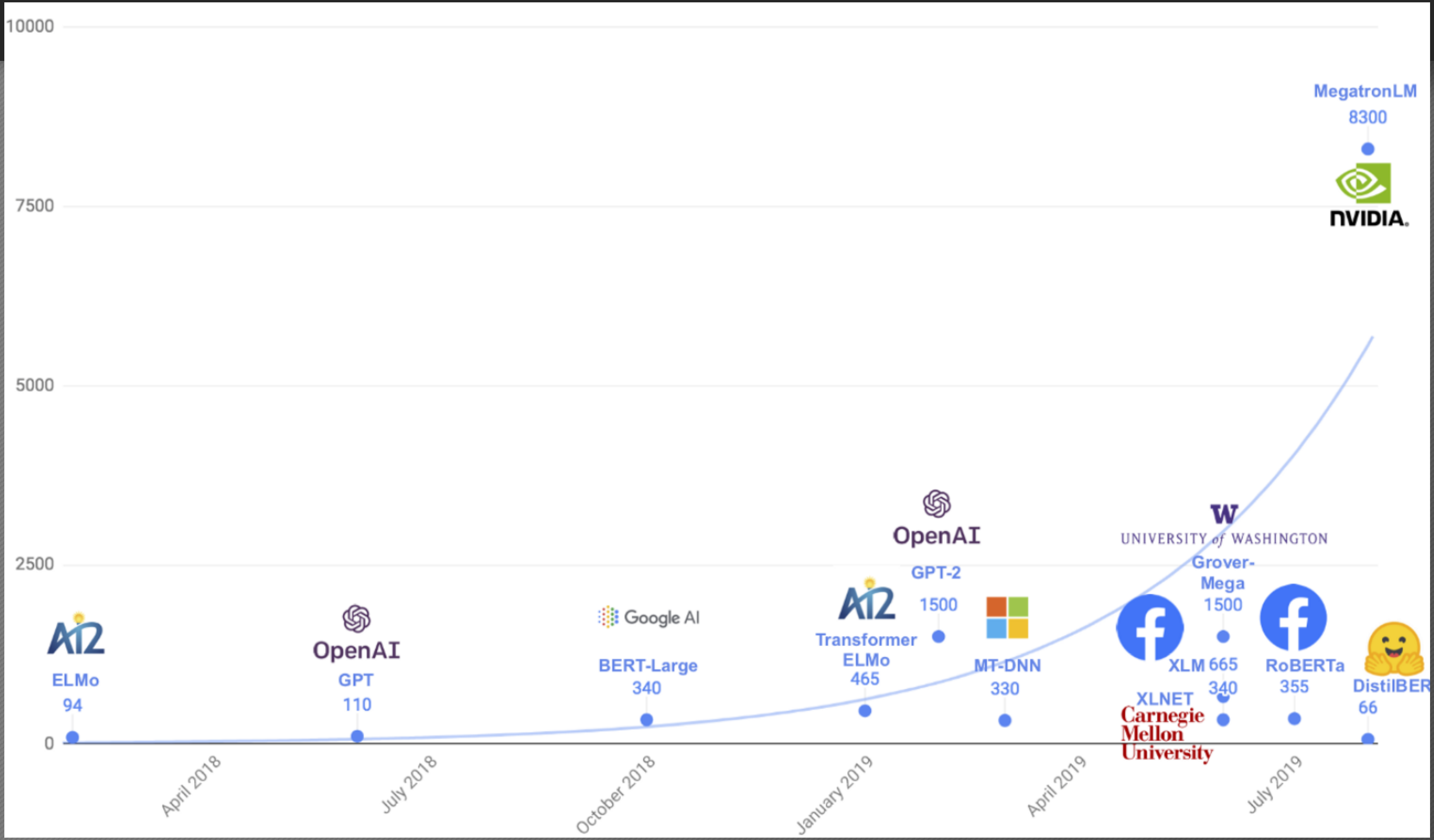
Scaled dot-product attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}})\mathbf{V}$$



https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html#full-architecture https://medium.com/@Moscow25/the-best-deep-natural-language-papers-you-should-read-bert-gpt-2-and-looking-forward-1647f4438797

Transformer Variants Getting Rapidly Larger

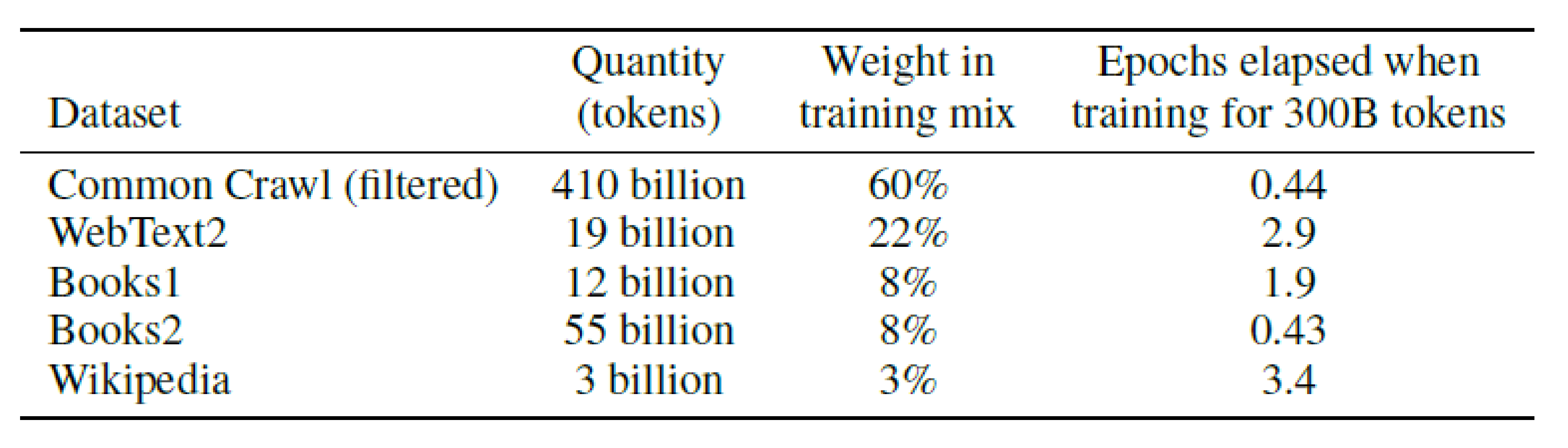


GPT-3 175B Model Architecture

- Input reversibly encoded as tokens with “Byte Pair Encoding” of UTF-8 bytes
- Context window of 2,048 tokens
- 96 transformer layers
- 96 self-attention heads, each 128 dimensional
- 12,288 units in bottleneck layer, 49,152 in feed forward layer
- Batch size of 3.2M samples
- Learning rate .6*10^-4

GPT-3 Training Data

- Trained on 499 Billion tokens
- Would require 355 years and $4,600,000 train on cheapest GPU cloud



Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

https://lambdalabs.com/blog/demystifying-gpt-3/

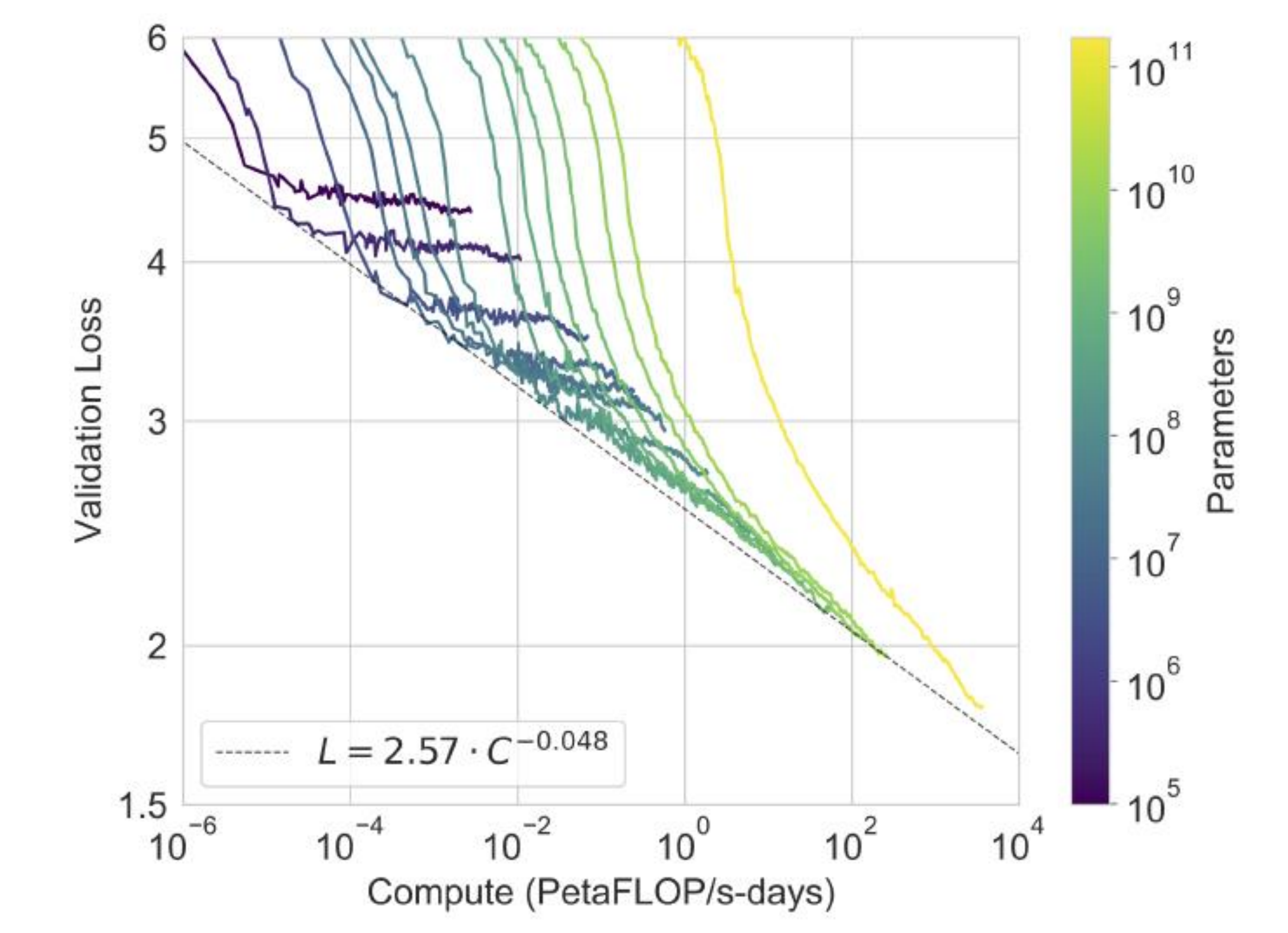
Microsoft-built AI Supercomputer

- NVIDIA V100 GPUs in a high-bandwidth cluster
- 285,000 CPU cores
- 10,000 GPUs
- 400 gigabits per second network connectivity for each GPU server
- Trained on cuDNN accelerated PyTorch models

https://blogs.microsoft.com/ai/openai-azure-supercomputer/

https://arxiv.org/abs/2005.14165

GPT Validation Loss vs. Compute



Perplexity

Perplexity of p is 2^H(p) where H is the entropy of p

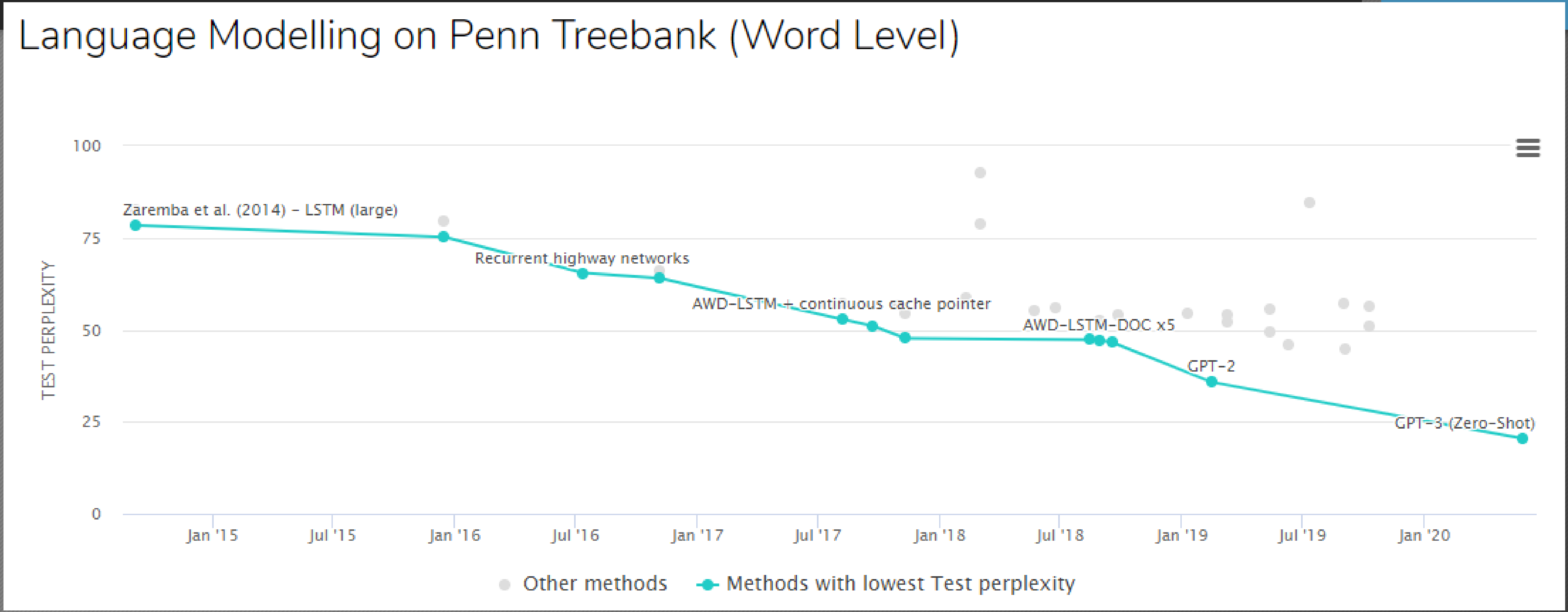
It’s k if the uncertainty of the next word is like a k-sided dice

- Unigram: 962
- Bigram: 170
- Trigram: 109
- GPT-2: 35.8
- GPT-3: 20.5
- Human: 12

https://web.stanford.edu/~jurafsky/slp3/

https://www.isca-speech.org/archive/Interspeech_2017/abstracts/0729.html

SOTA Perplexity on Penn Treebank



Language Modelling on Penn Treebank (Word Level)

https://paperswithcode.com/sota/language-modelling-on-penn-treebank-word

GPT-3 solves a version of the Turing Test

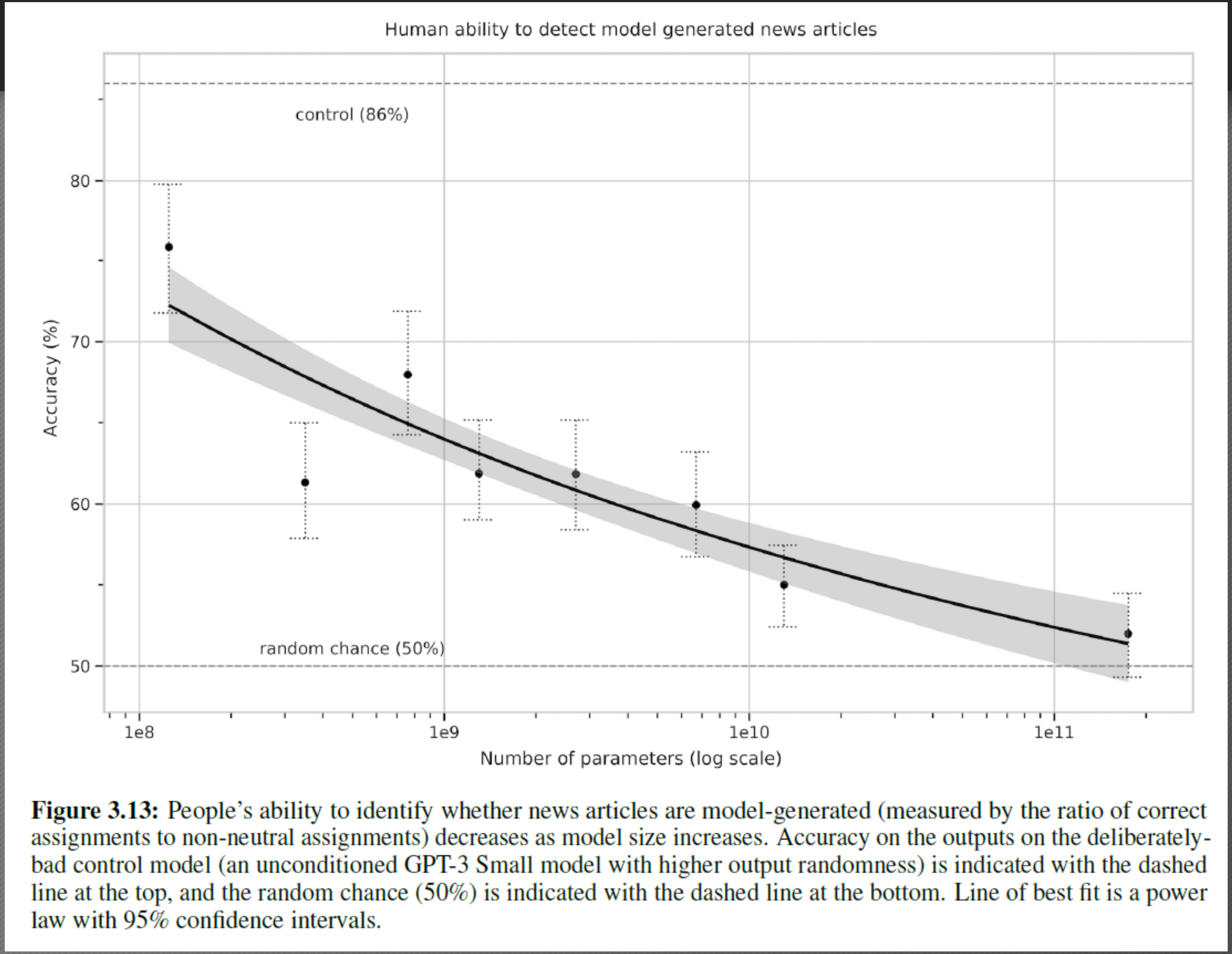


Figure 3.13: People's ability to identify whether news articles are model-generated (measured by the ratio of correct assignments to non-neutral assignments) decreases as model size increases. Accuracy on the outputs on the deliberately-bad control model (an unconditioned GPT-3 Small model with higher output randomness) is indicated with the dashed line at the top, and the random chance (50%) is indicated with the dashed line at the bottom. Line of best fit is a power law with 95% confidence intervals.

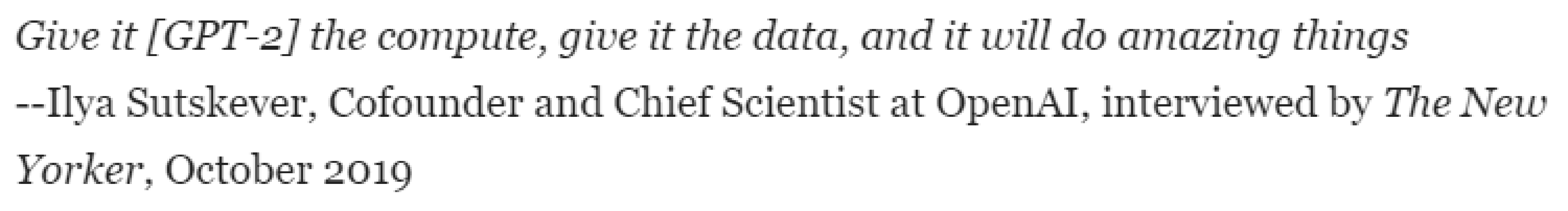
Rich Sutton’s “Bitter Lesson”

- Simple AI leveraging compute power beats clever AI built using human knowledge
- Deep Blue chess machine based on search
- NLP translation based on n-grams
- Scaling of NLP transformer models
- AlphaGo based on search and self-play

- Gwern: OA5, BigGAN, BiT, ViLBERT, AlphaStar, MetaMimic, StyleGAN, GQN, Dactyl, DD-PPO, Procgen, AlphaZero, MuZero

https://www.gwern.net/newsletter/2020/05

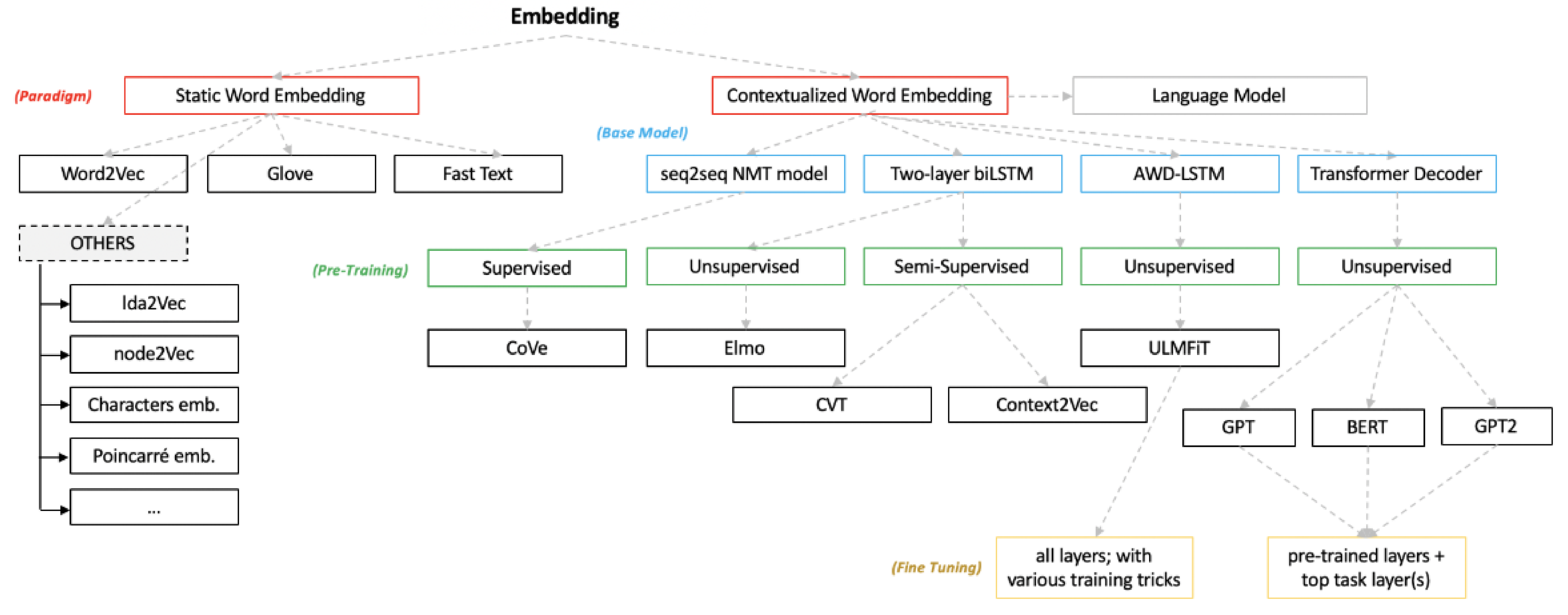
http://www.incompleteideas.net/IncIdeas/BitterLesson.html



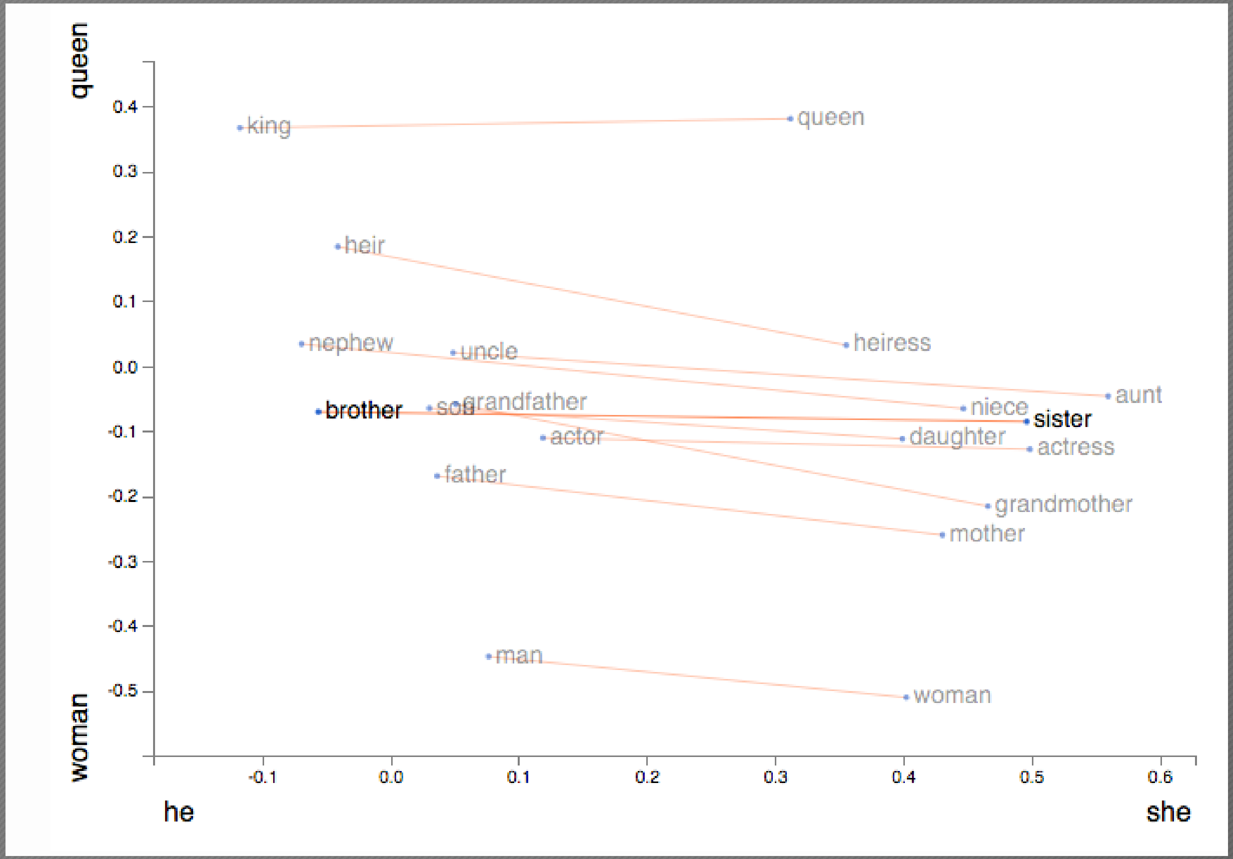
Give it [GPT-2] the compute, give it the data, and it will do amazing things

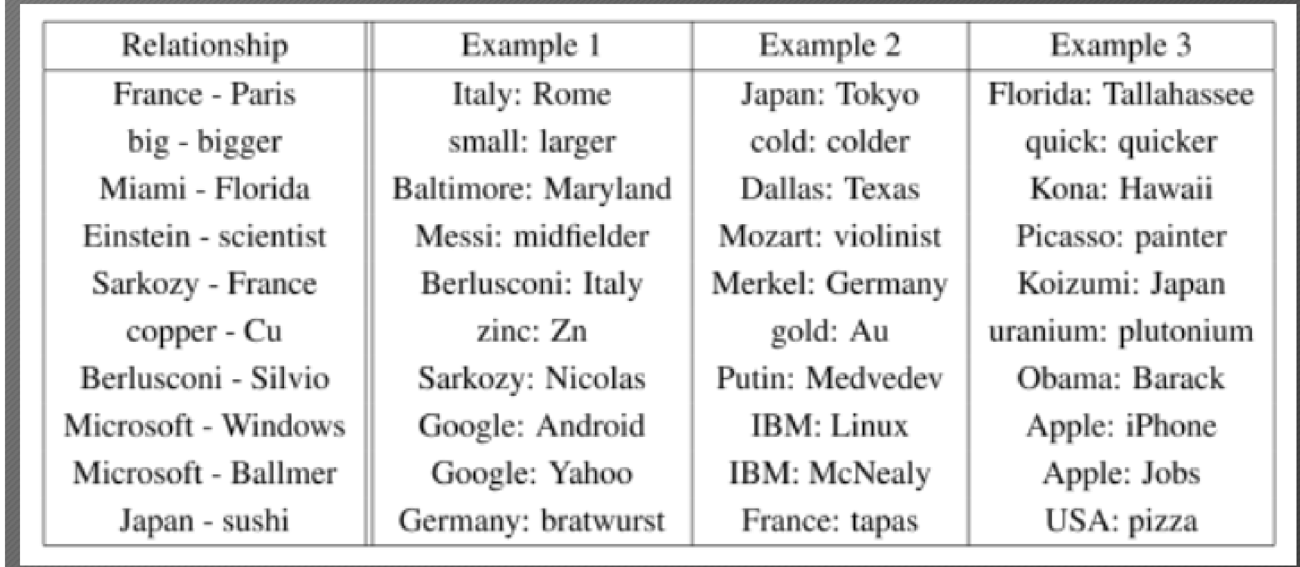
--Ilya Sutskever, Cofounder and Chief Scientist at OpenAI, interviewed by *The New Yorker*, October 2019 https://towardsdatascience.com/from-pre-trained-word-embeddings-to-pre-trained-language-models-focus-on-bert-343815627598

Word Embeddings



King – Man + Woman = Queen

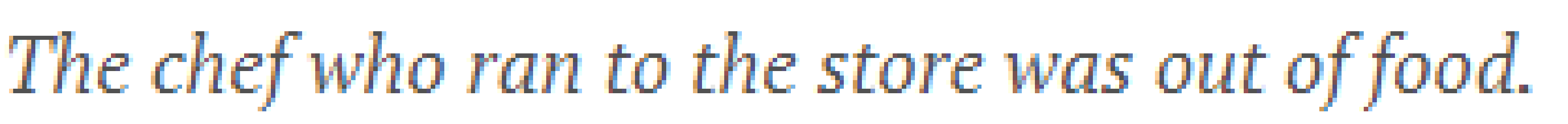




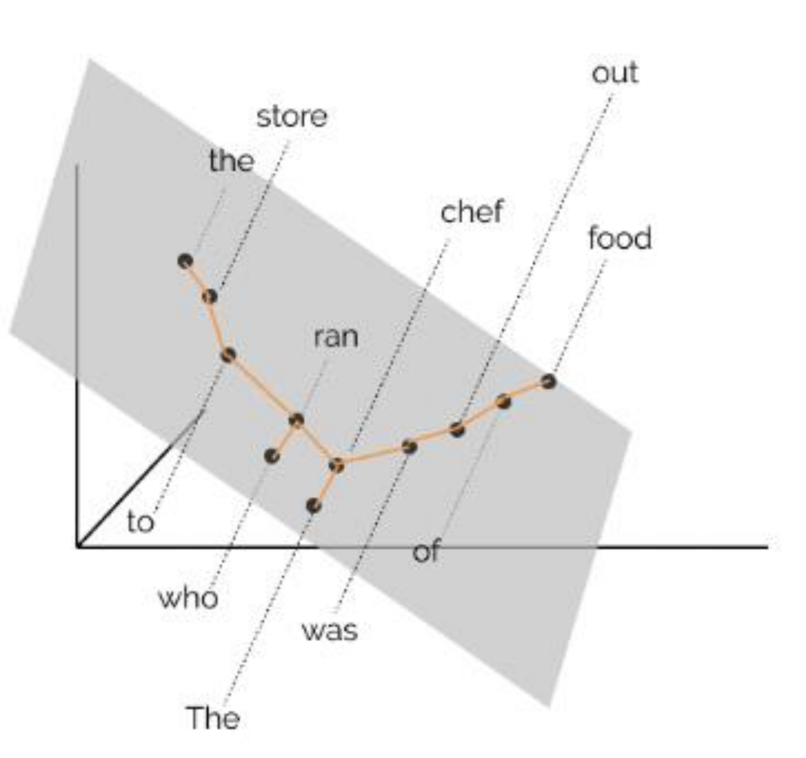
Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

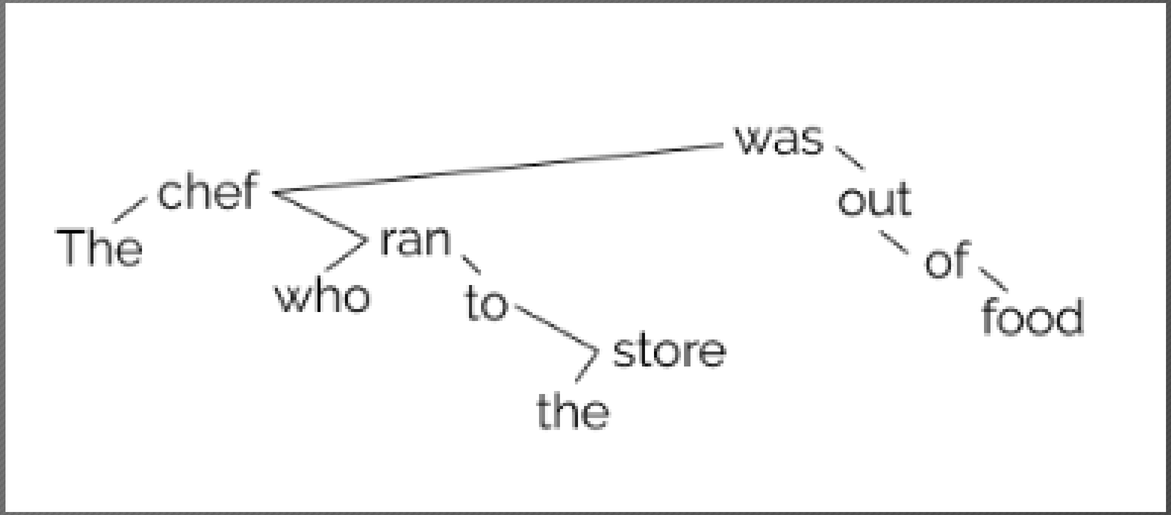
https://lamyiowce.github.io/word2viz/ https://www.kdnuggets.com/2016/05/amazing-power-word-vectors.html

Parse trees encoded in embeddings



The chef who ran to the store was out of food.



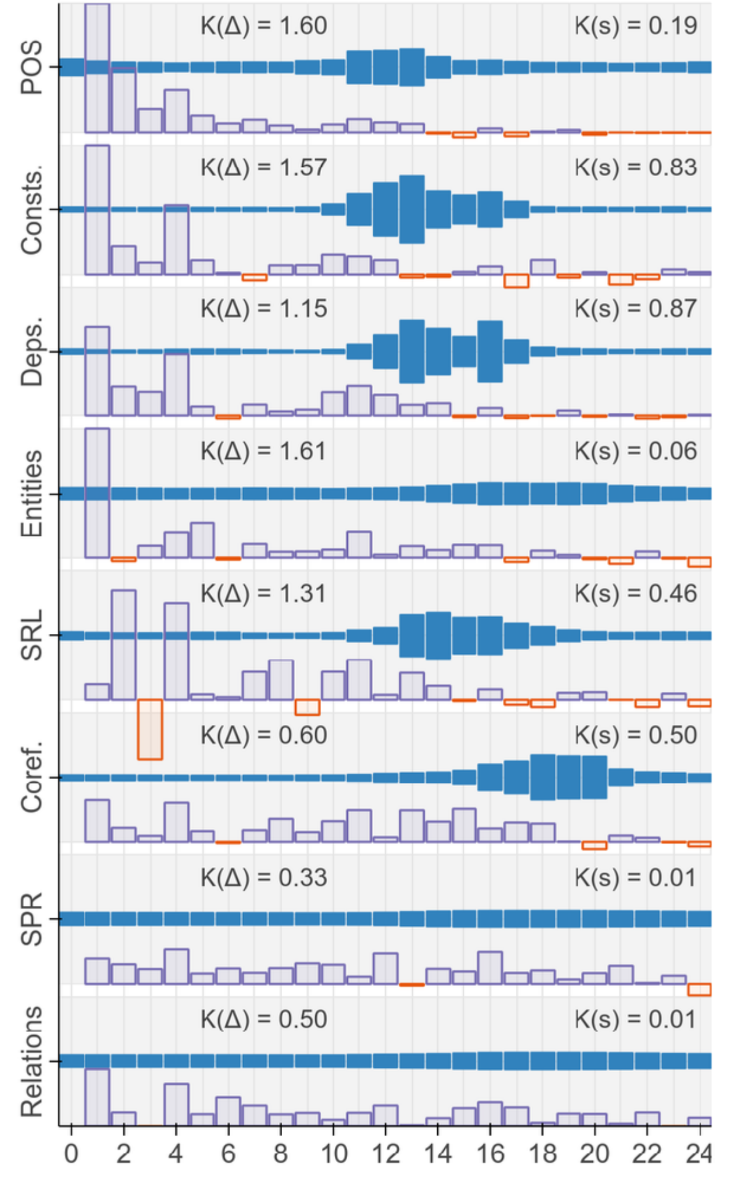


https://nlp.stanford.edu//~johnhew//structural-probe.html?utm_source=quora&utm_medium=referral#the-structural-probe

BERT Rediscovers the Classical NLP Pipeline

1. Part of Speech
2. Constituents
3. Dependencies
4. Entities
5. Semantic Role Labelling
6. Coreference
7. Semantic Proto-Roles
8. Relation Classification

https://arxiv.org/abs/1905.05950

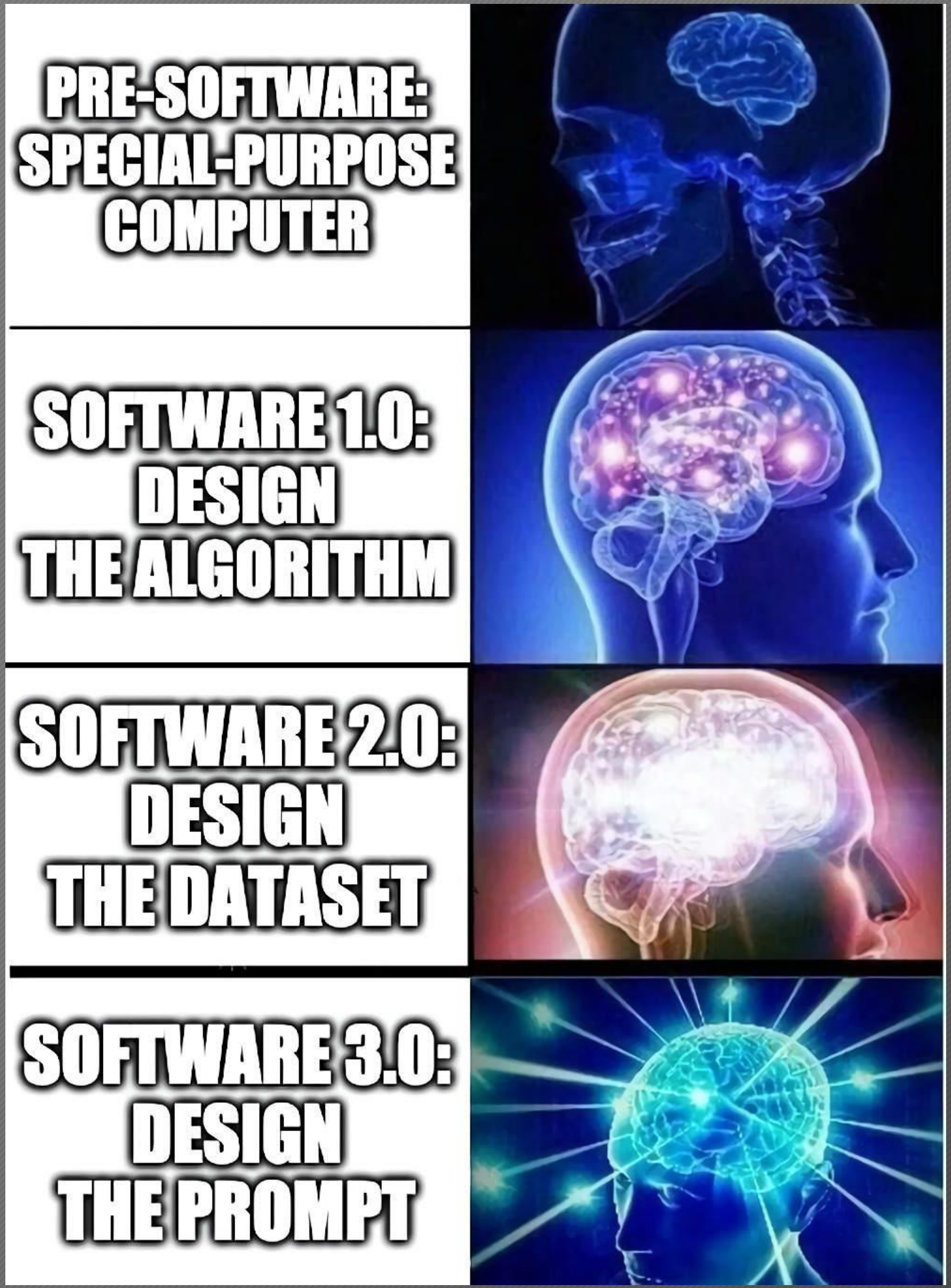


Distributional Semantics

- 1957 John Firth:
- “You shall know a word by the company it keeps”

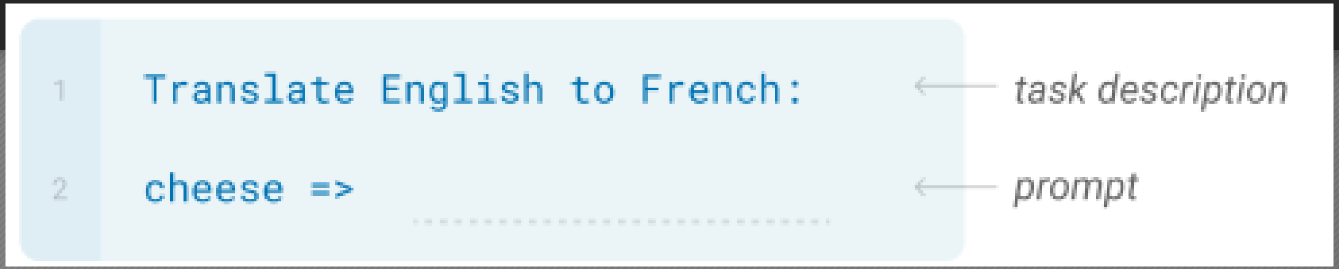
https://en.wikipedia.org/wiki/Distributional_semantics

Software 3.0?

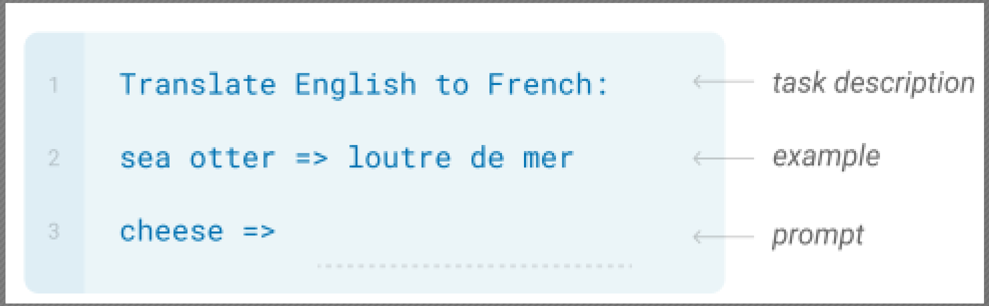

PRE-SOFTWARE: SPECIAL-PURPOSE COMPUTER
SOFTWARE 1.0: DESIGN THE ALGORITHM
SOFTWARE 2.0: DESIGN THE DATASET
SOFTWARE 3.0: DESIGN THE PROMPT

https://twitter.com/karpathy/status/1273788774422441984

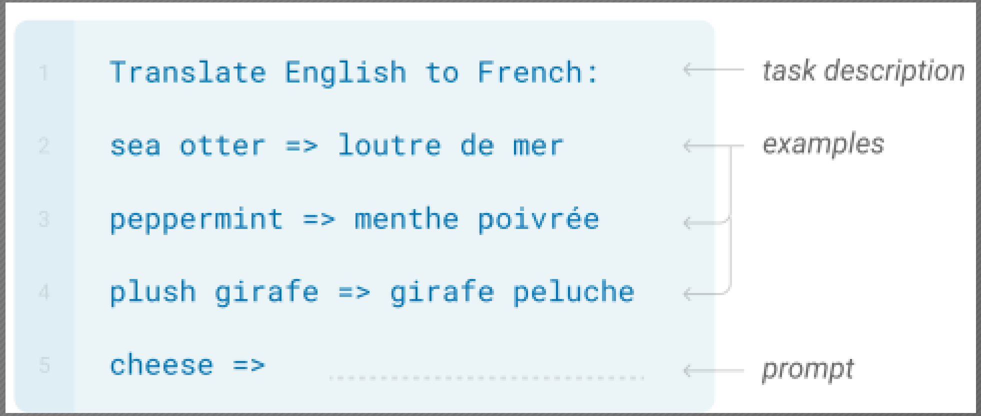
Zero-shot, One-shot, Few-shot



```
Translate English to French:
cheese =>
```



```
Translate English to French:
sea otter => loutre de mer
cheese =>
```



```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

- Context window is 2048 tokens
- Usually this is between 10 and 100 examples

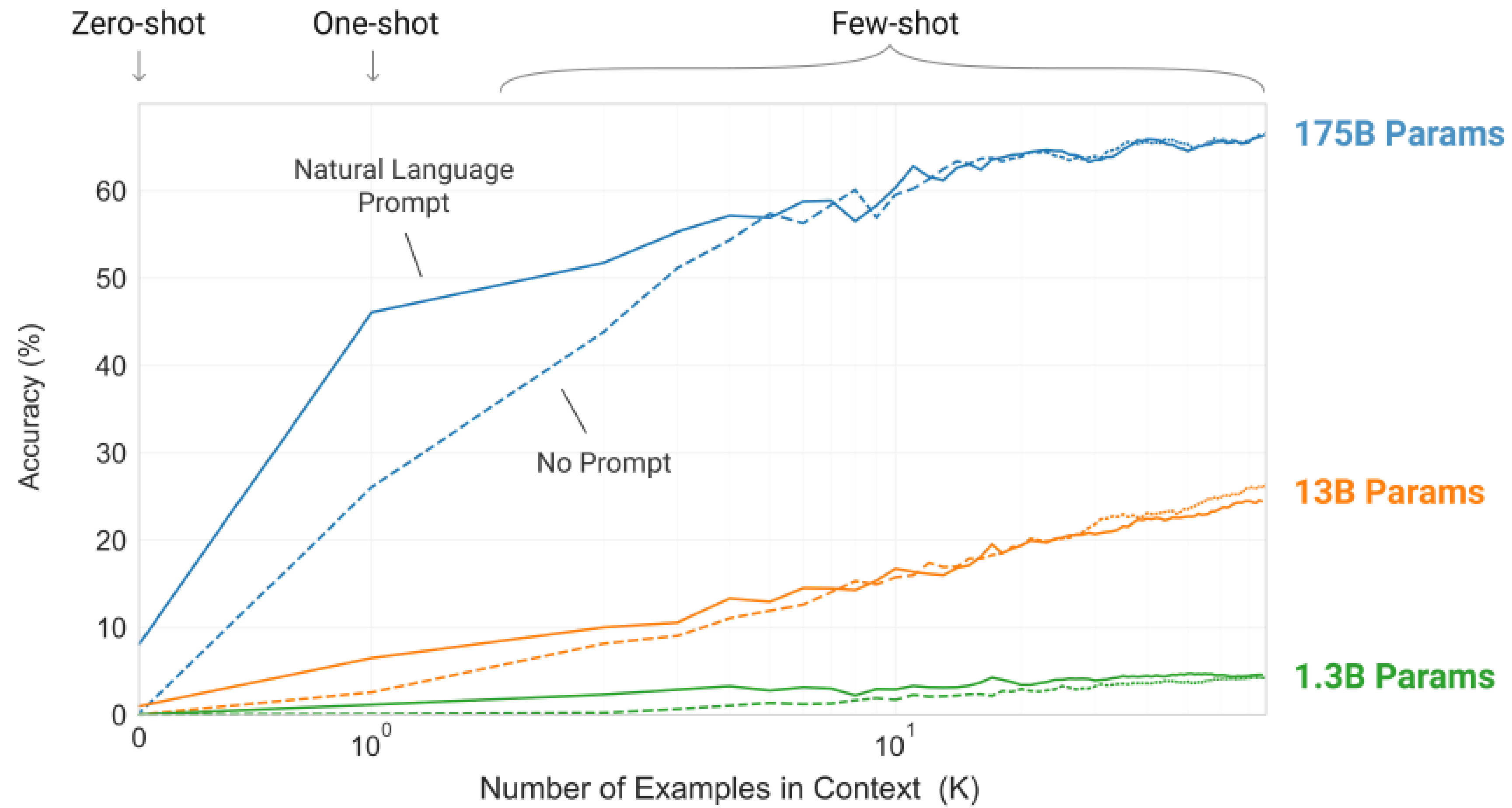


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning 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.

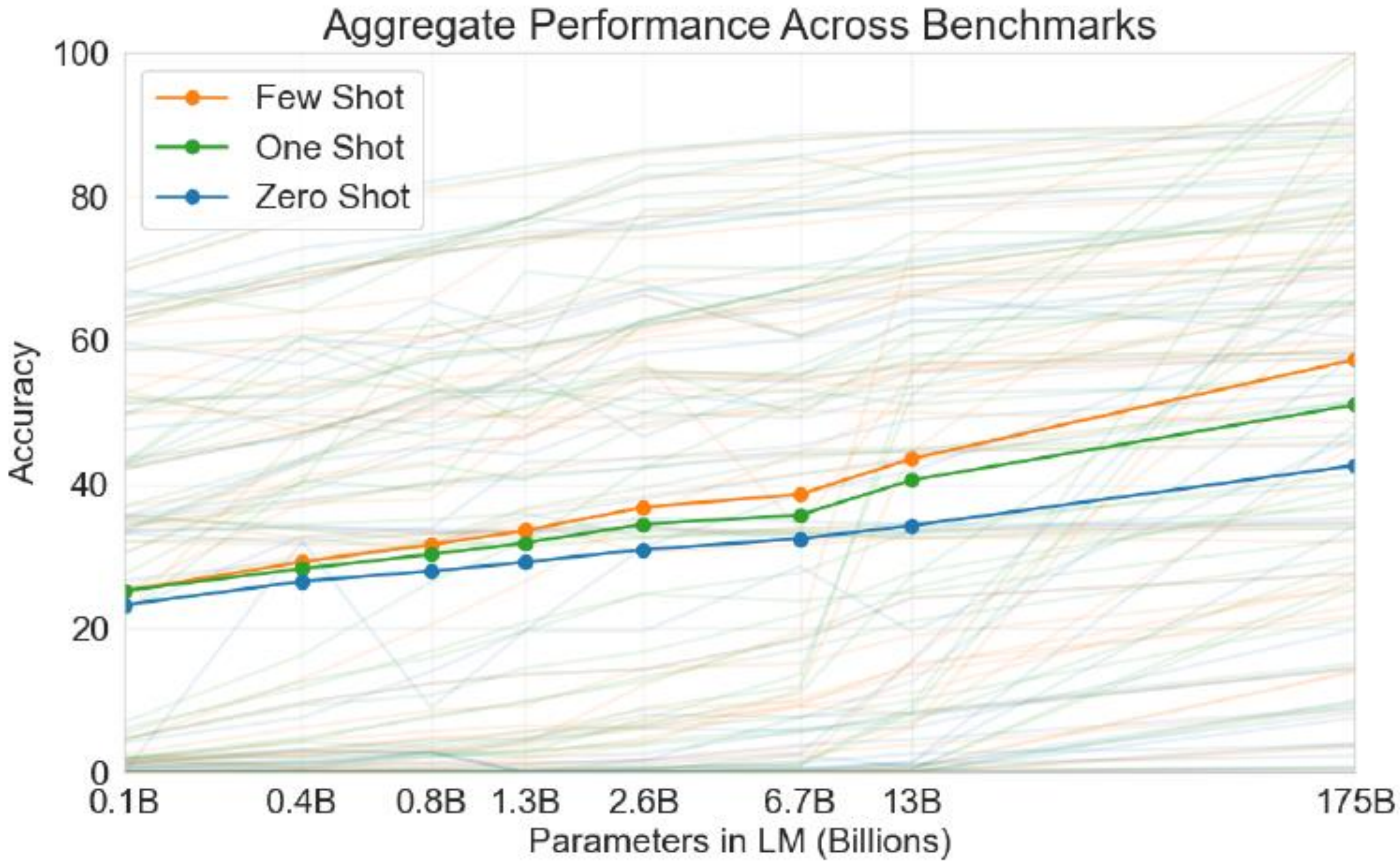


Figure 1.3: Aggregate performance for all 42 accuracy-denominated benchmarks While zero-shot performance improves steadily with model size, few-shot performance increases more rapidly, demonstrating that larger models are more proficient at in-context learning. See Figure 3.8 for a more detailed analysis on SuperGLUE, a standard NLP benchmark suite.

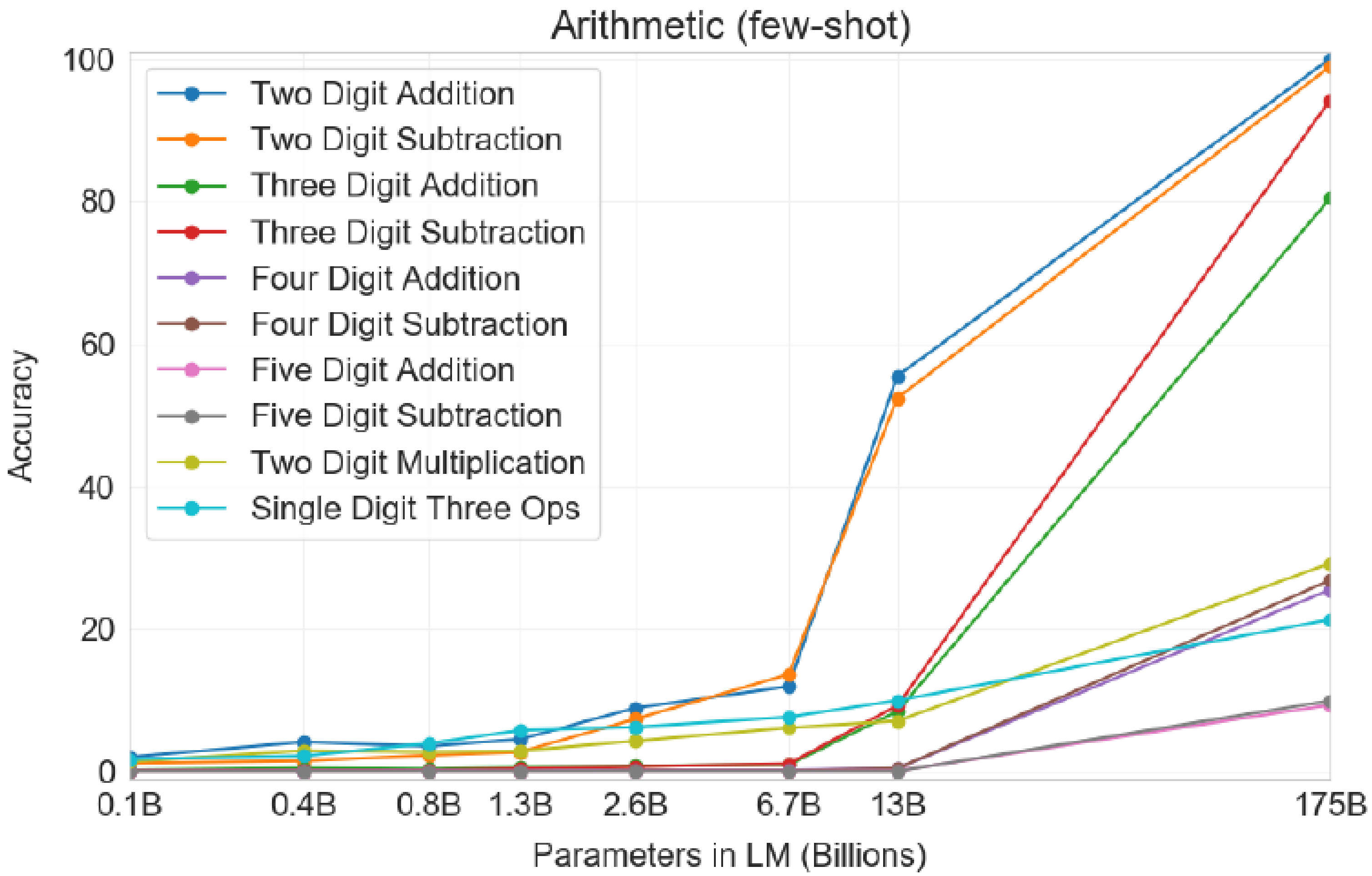


Figure 3.10: Results on all 10 arithmetic tasks in the few-shot settings for models of different sizes. There is a significant jump from the second largest model (GPT-3 13B) to the largest model (GPT-3 175), with the latter being able to reliably accurate 2 digit arithmetic, usually accurate 3 digit arithmetic, and correct answers a significant fraction of the time on 4-5 digit arithmetic, 2 digit multiplication, and compound operations. Results for one-shot and zero-shot are shown in the appendix.

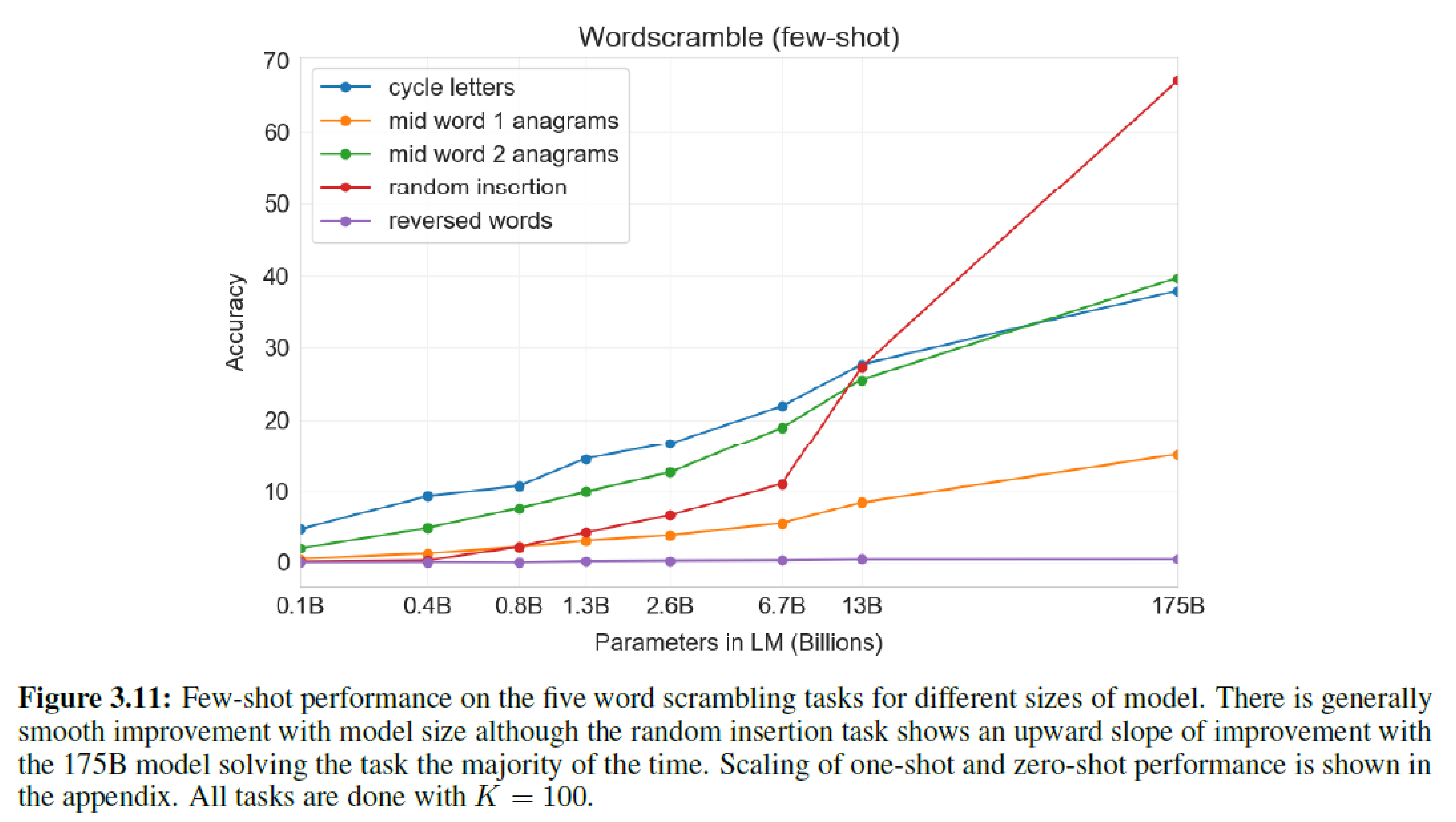
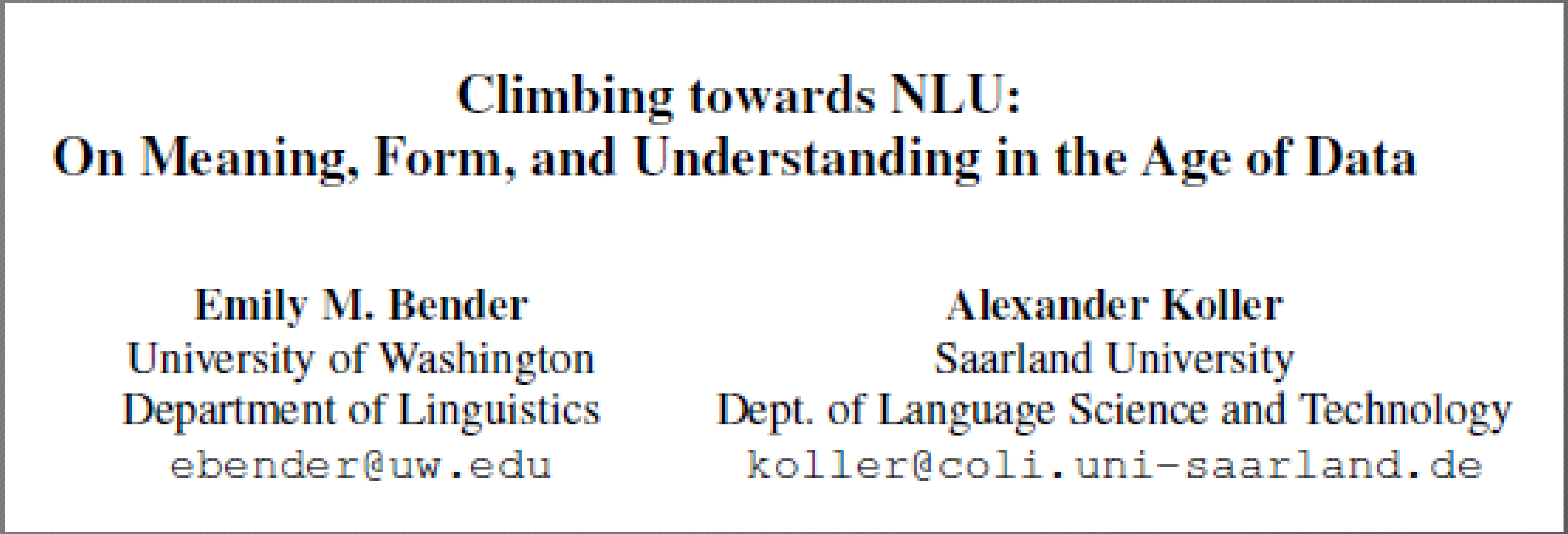


Figure 3.11: Few-shot performance on the five word scrambling tasks for different sizes of model. There is generally smooth improvement with model size although the random insertion task shows an upward slope of improvement with the 175B model solving the task the majority of the time. Scaling of one-shot and zero-shot performance is shown in the appendix. All tasks are done with $K = 100$.

https://gwern.substack.com/p/may-gwernnet-newsletter

Gwern Comments

- Arithmetic, English translation, unscrambling anagrams, SAT analogies
- With no training or finetuning!
- GPT-3 attention is a kind of “fast weight” that “learned to learn”?
- Simple uniform architecture, dumbest learning task, trained on random internet text, on data that fits on a laptop, small and shallow compared to what is currently possible, but has “crazy runtime meta-learning”
- Hardest test Winograd WinoGrande will fall at 10T parameters?
- Bitter lesson: The bigger and harder, the better!
- Could AGI emerge from future versions of GPT-3?



Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender
University of Washington
Department of Linguistics
ebender@uw.edu

Alexander Koller
Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de

"We argue that the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning."

https://www.aclweb.org/anthology/2020.acl-main.463/

What do you need to train a general intelligence?

1. Real world interaction

2. Simulated video interaction

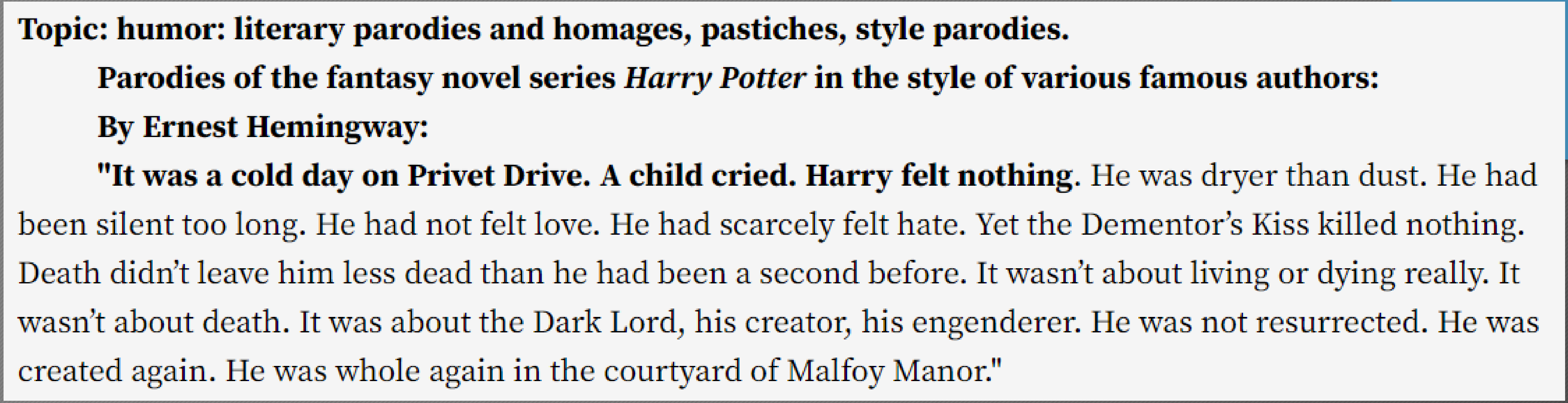
3. Video stream

4. Language stream

https://www.gwern.net/GPT-2#fnref11

Real world semantics from word statistics

- All US presidents and Russian leaders in temporal order.
- Longitude and latitude of cities in US and Europe and relative distances.
- Relative size of many objects (e.g. cars, elephants, humans, houses, etc.)
- Dangerous animals and objects.
- How smart different animals are.
- Clothes appropriate for age groups, emotional arousal, cost, weather, etc.
- Qualities of mythological creatures.
- Physical properties of objects: rigidness, strength, transparency, etc.
- Whole-part relations, body parts, connectedness, etc.
- Countries and cities: capitals, GDP, internet usage, life expenctancy, etc.

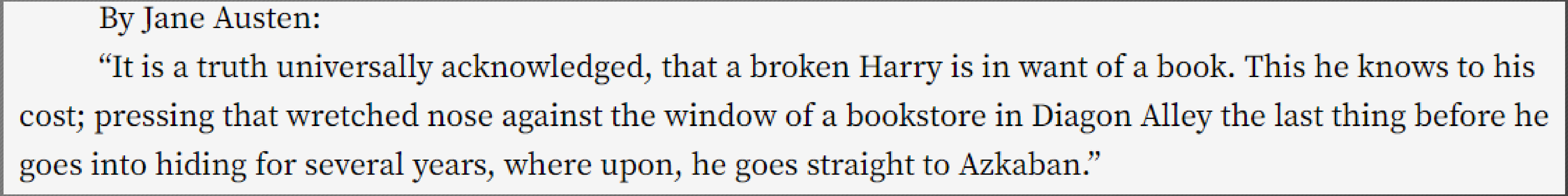


Topic: humor: literary parodies and homages, pastiches, style parodies.

Parodies of the fantasy novel series *Harry Potter* in the style of various famous authors:

By Ernest Hemingway:

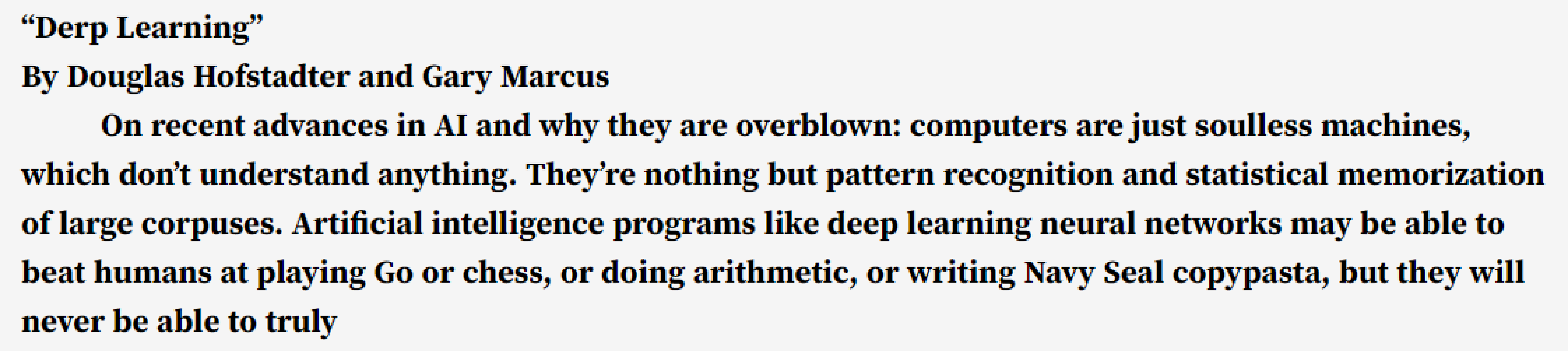
"It was a cold day on Privet Drive. A child cried. Harry felt nothing. He was dryer than dust. He had been silent too long. He had not felt love. He had scarcely felt hate. Yet the Dementor's Kiss killed nothing. Death didn't leave him less dead than he had been a second before. It wasn't about living or dying really. It wasn't about death. It was about the Dark Lord, his creator, his engenderer. He was not resurrected. He was created again. He was whole again in the courtyard of Malfoy Manor."



By Jane Austen:

"It is a truth universally acknowledged, that a broken Harry is in want of a book. This he knows to his cost; pressing that wretched nose against the window of a bookstore in Diagon Alley the last thing before he goes into hiding for several years, where upon, he goes straight to Azkaban."

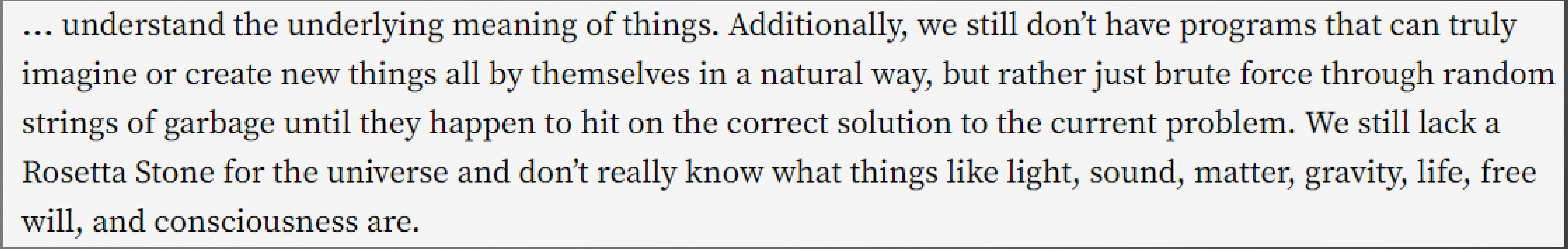
https://www.gwern.net/GPT-3#



“Derp Learning”

By Douglas Hofstadter and Gary Marcus

On recent advances in AI and why they are overblown: computers are just soulless machines, which don’t understand anything. They’re nothing but pattern recognition and statistical memorization of large corpuses. Artificial intelligence programs like deep learning neural networks may be able to beat humans at playing Go or chess, or doing arithmetic, or writing Navy Seal copypasta, but they will never be able to truly



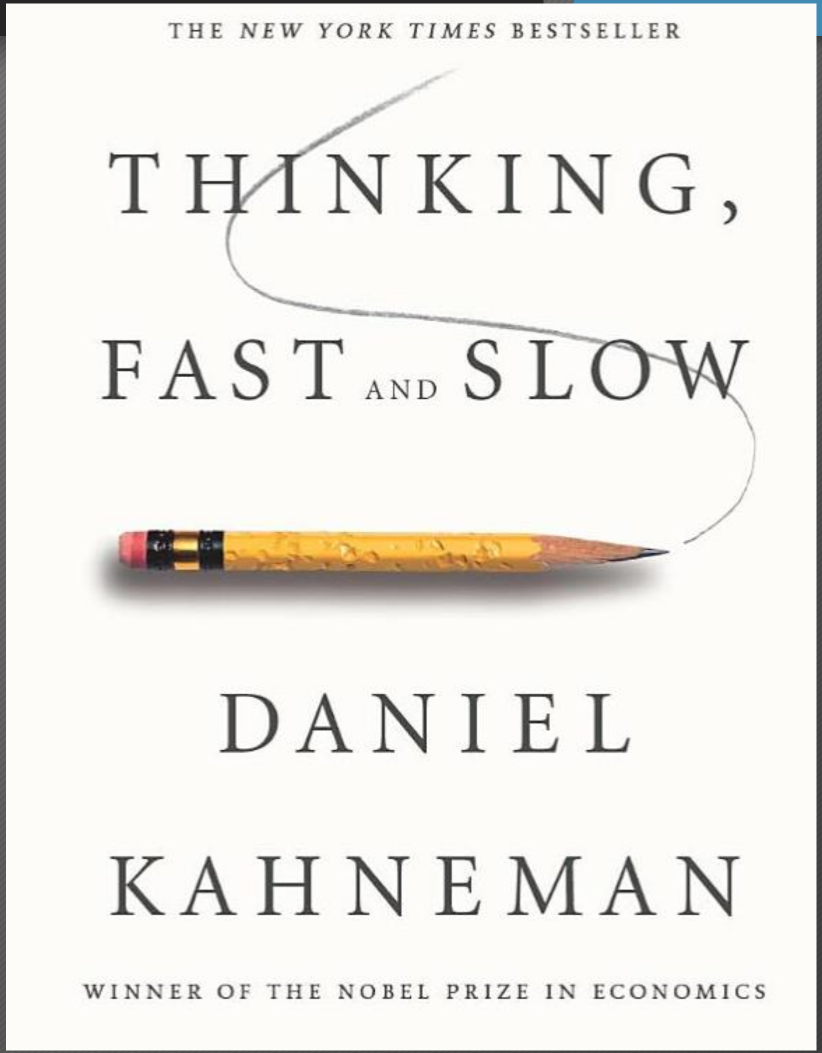
… understand the underlying meaning of things. Additionally, we still don’t have programs that can truly imagine or create new things all by themselves in a natural way, but rather just brute force through random strings of garbage until they happen to hit on the correct solution to the current problem. We still lack a Rosetta Stone for the universe and don’t really know what things like light, sound, matter, gravity, life, free will, and consciousness are.

https://www.gwern.net/GPT-3#

Dual Process Theory Perspective

GPT-3 mostly performs Type 1 Thinking

For general intelligence, we also need Type 2



THE *NEW YORK TIMES* BESTSELLER

THINKING, FAST AND SLOW

DANIEL KAHNEMAN

WINNER OF THE NOBEL PRIZE IN ECONOMICS