CARTE ML Workshop

GPT and Attention

Challenges in NLP

- **Ambiguity:** Words can mean different things depending on context
- **Nuances:** Languages are full of idioms, slang, cultural references, sarcasm…
- **Syntax vs Semantics:** A grammatically correct sentence might not make sense, or a grammatically incorrect one might be easy to understand

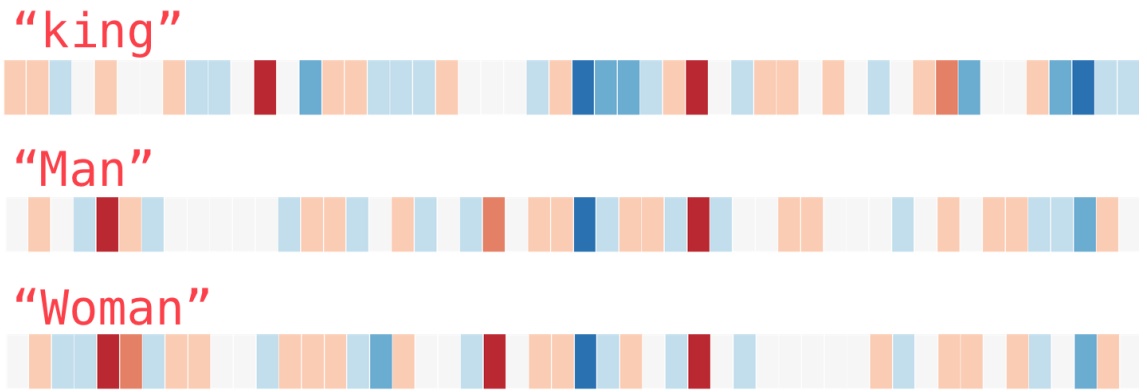
- I saw a man on the hill with the telescope
- That’s a cool cat
- Colourless green ideas sleep furiously
- Me went store

Foundations of LLMs

- Fundamental goal of language modelling: next word prediction
- $P(cat \mid the\ dog\ and\ the)$
- To generate, pick the word with highest likelihood
- Early models could handle one, two words of context
- Locally coherent, but longer texts quickly lose meaning
- More context requires more complexity!

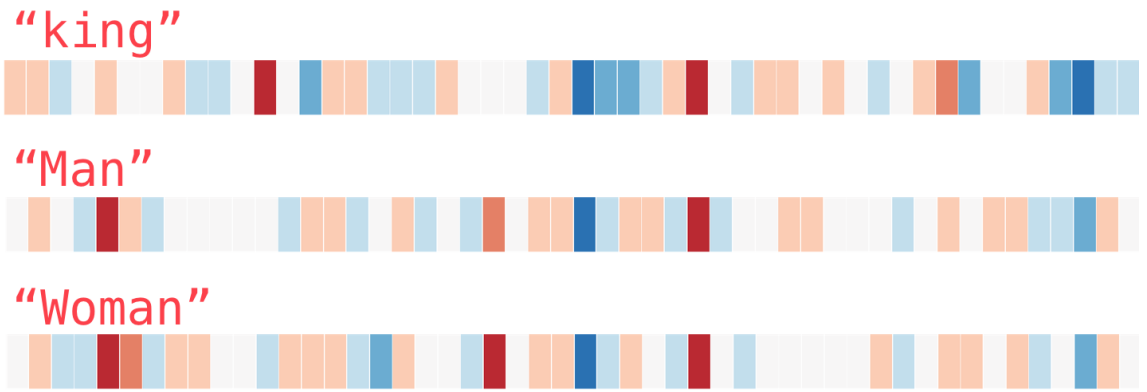
How does an LM “understand” word meaning?

- In order to predict the likelihood of a word, we must have some sense of its meaning
- Some words have similar meanings, and can easily fit in the same place
- In the same way CNNs convert an image into a set of feature maps, we can convert a word into a set of abstract linguistic features
- Word2Vec: 300 features
- GPT-3: 12,888

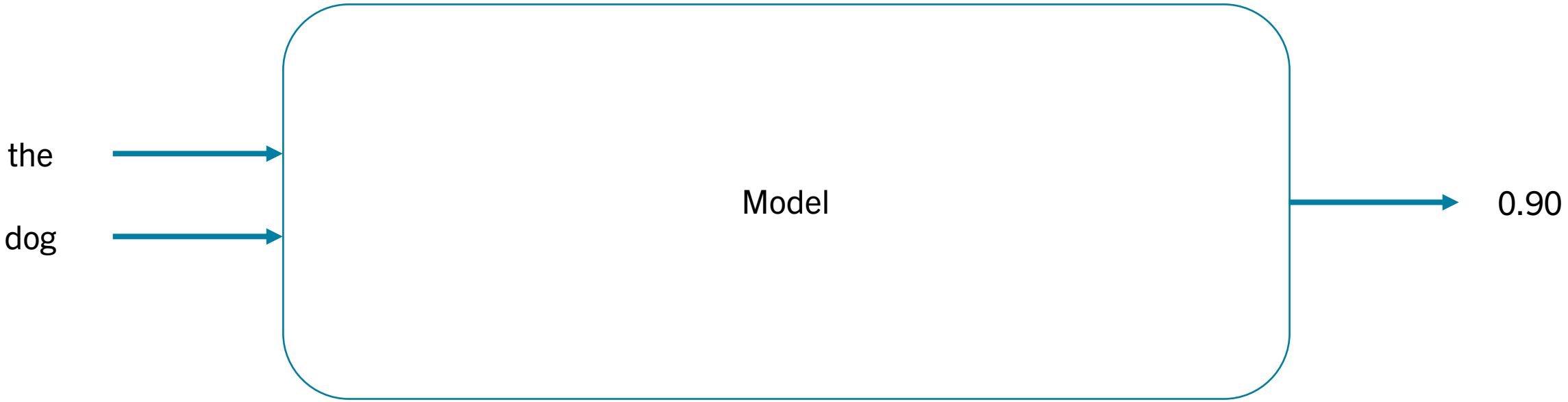


How does an LM “understand” word meaning?

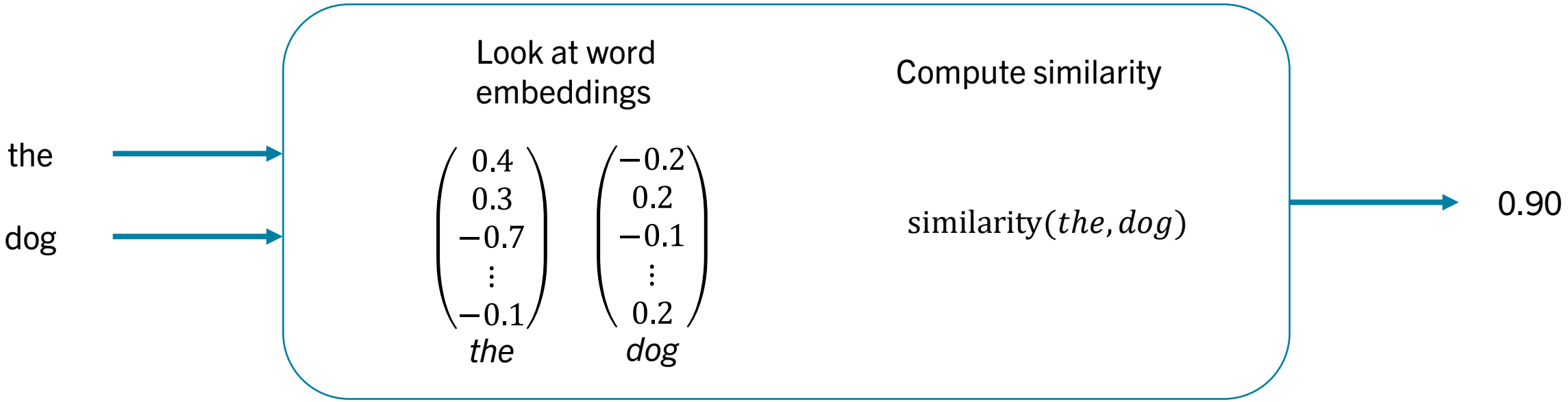
- Key concept of word embeddings: similar words should have similar vectors
- How do we accomplish this?



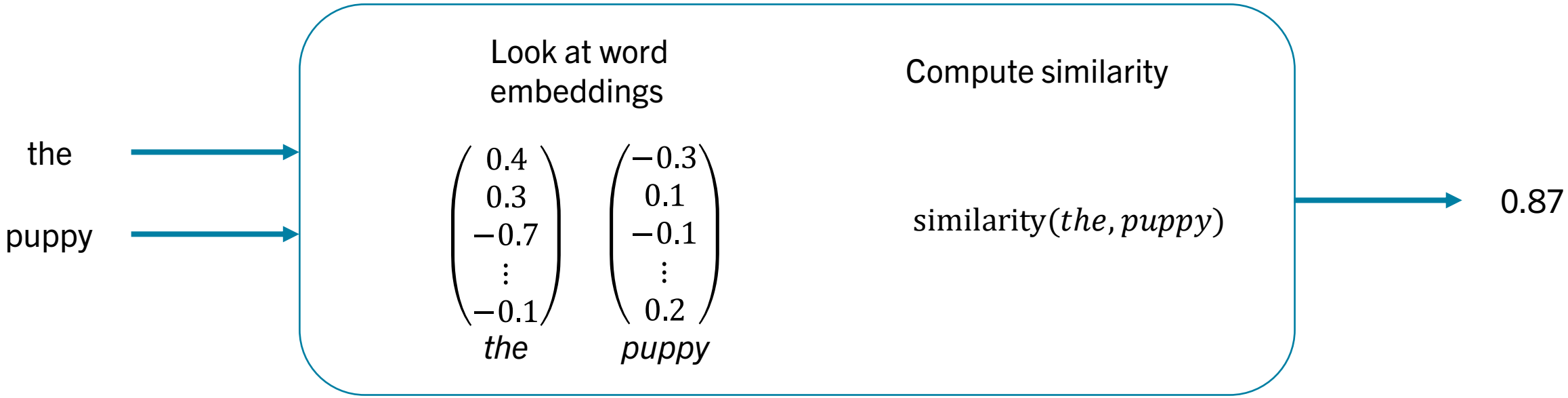
Building Word Embeddings



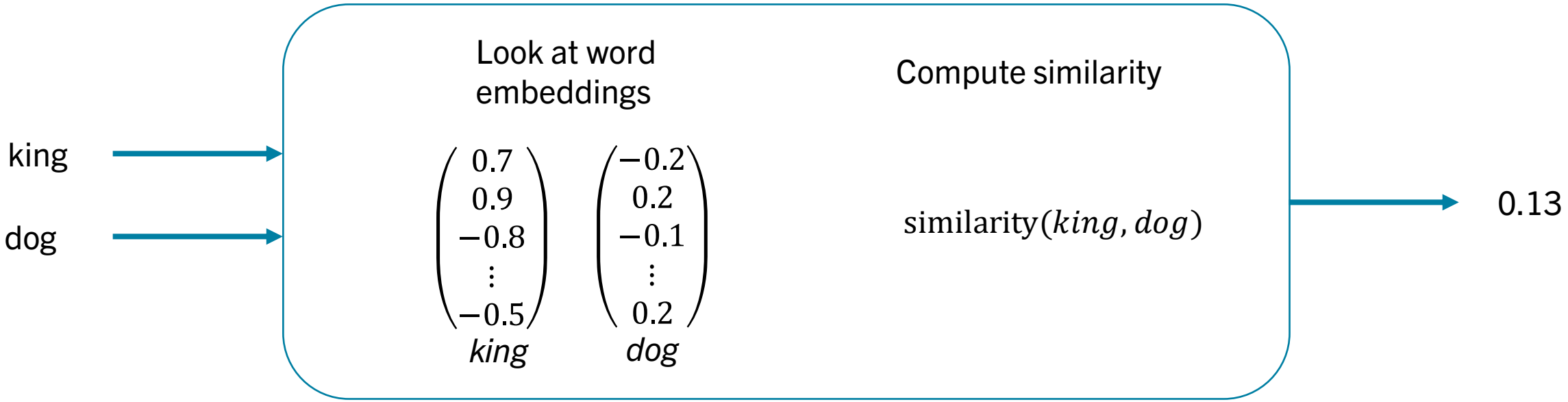
Building Word Embeddings



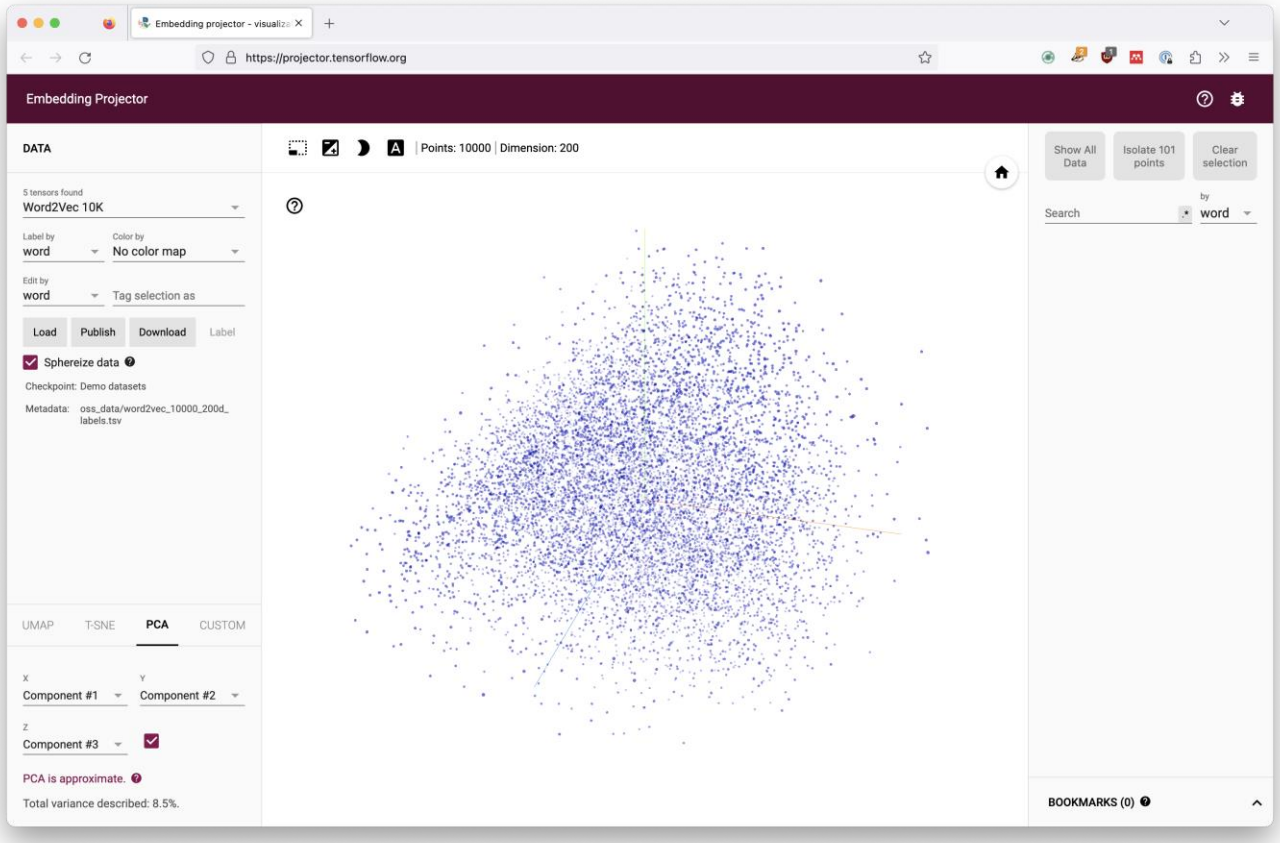
Building Word Embeddings



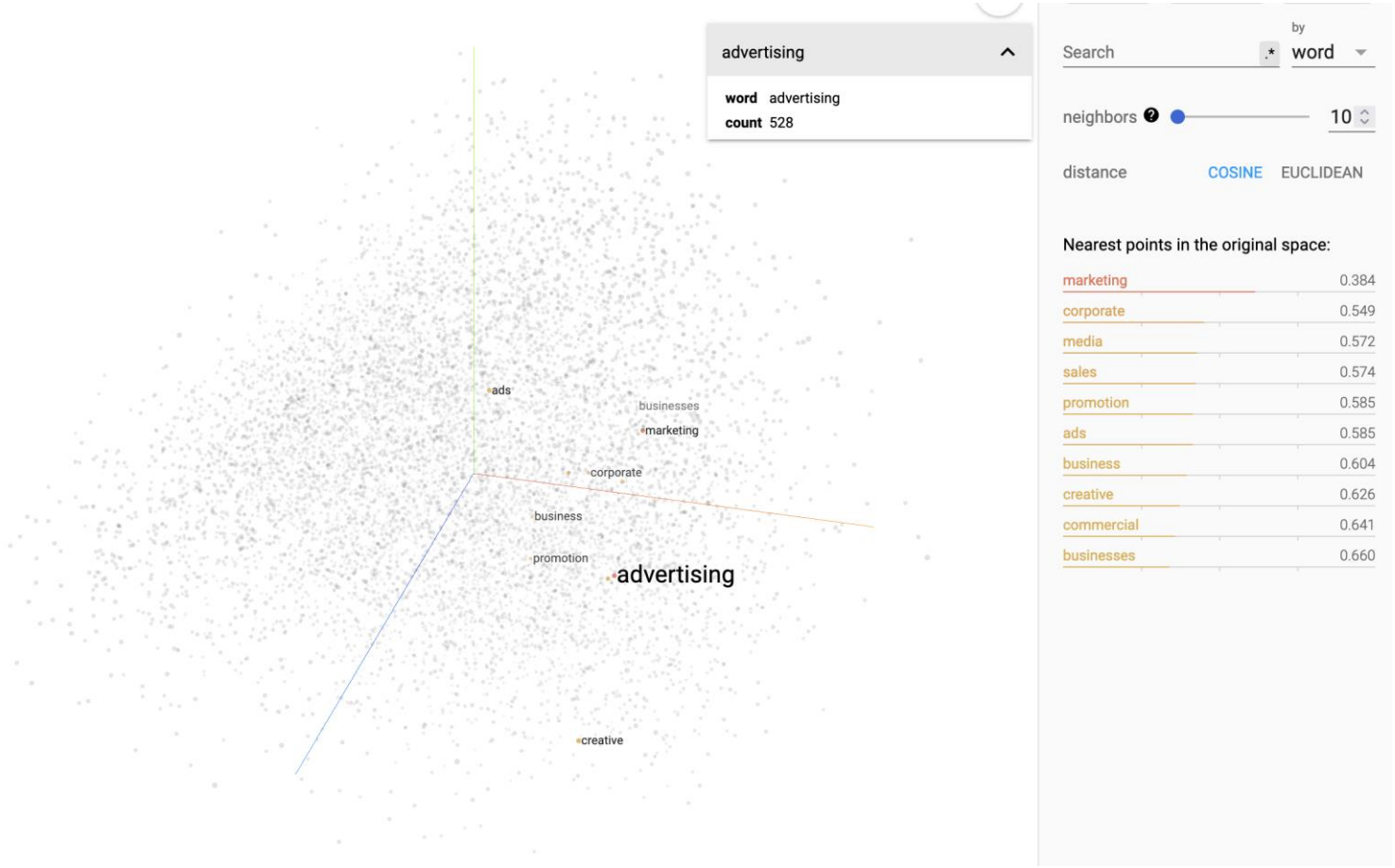
Building Word Embeddings



Results

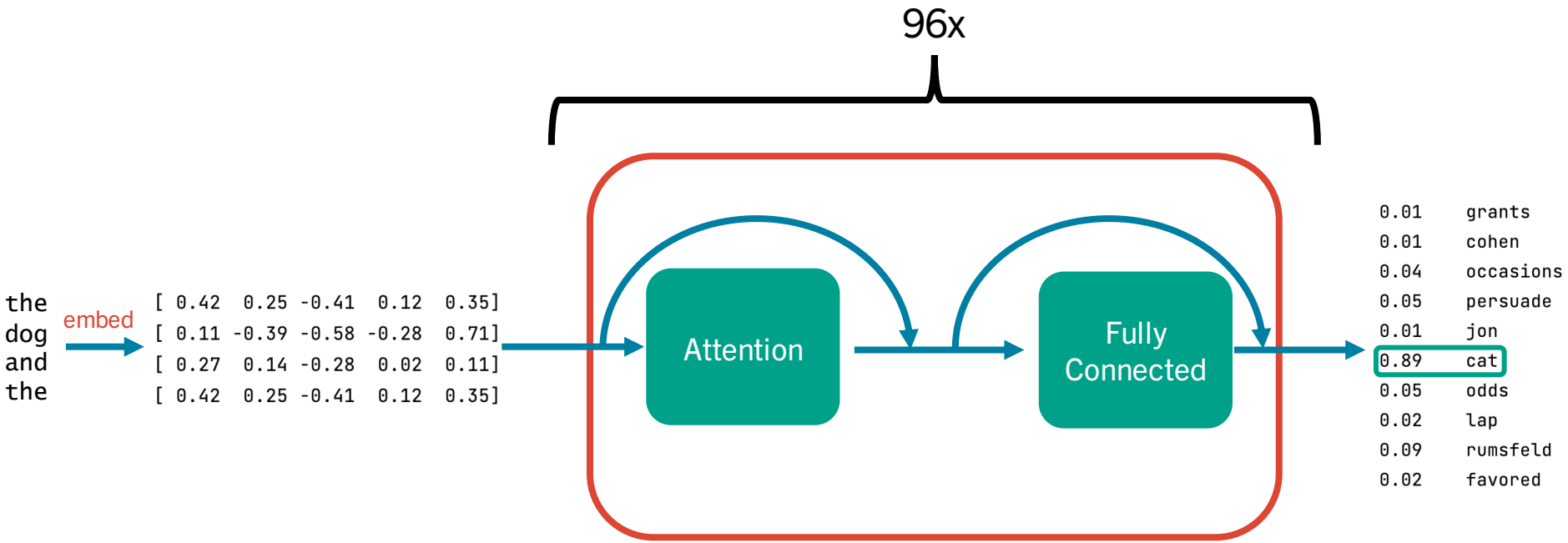

Embedding projector - visualize
https://projector.tensorflow.org
Embedding Projector
DATA
5 tensors found
Word2Vec 10K
Label by
word
Color by
No color map
Edit by
word
Tag selection as
Load
Publish
Download
Label
Sphereize data
Checkpoint: Demo datasets
Metadata: oss_data/word2vec_10000_200d_labels.tsv
UMAP
T-SNE
PCA
CUSTOM
X
Component #1
Y
Component #2
Z
Component #3
PCA is approximate.
Total variance described: 8.5%.
Points: 10000 | Dimension: 200
Show All Data
Isolate 101 points
Clear selection
Search
by
word
BOOKMARKS (0)

Results

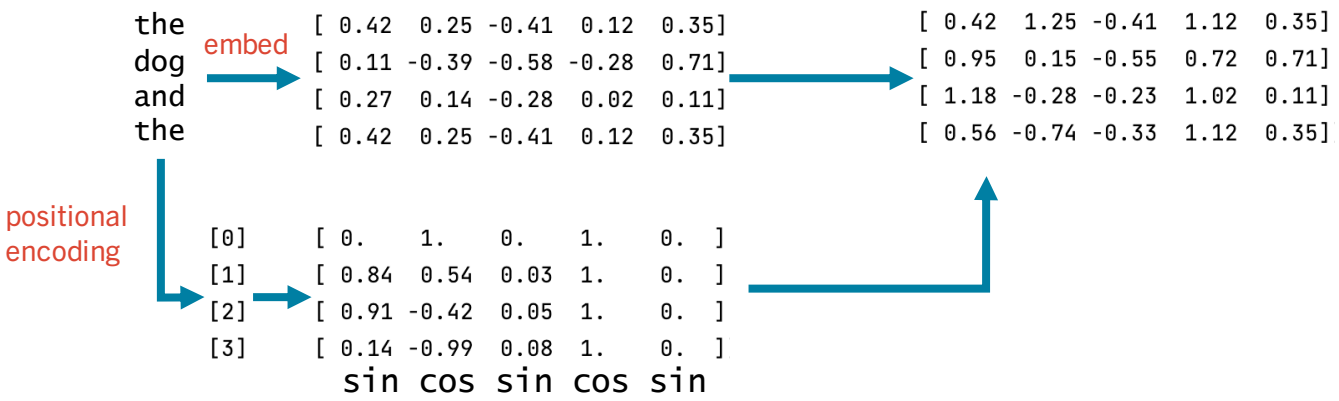

advertising
word advertising
count 528
Search
by
word
neighbors
10
distance
COSINE
EUCLIDEAN
Nearest points in the original space:
marketing 0.384
corporate 0.549
media 0.572
sales 0.574
promotion 0.585
ads 0.585
business 0.604
creative 0.626
commercial 0.641
businesses 0.660
ads
businesses
marketing
corporate
business
promotion
advertising
creative

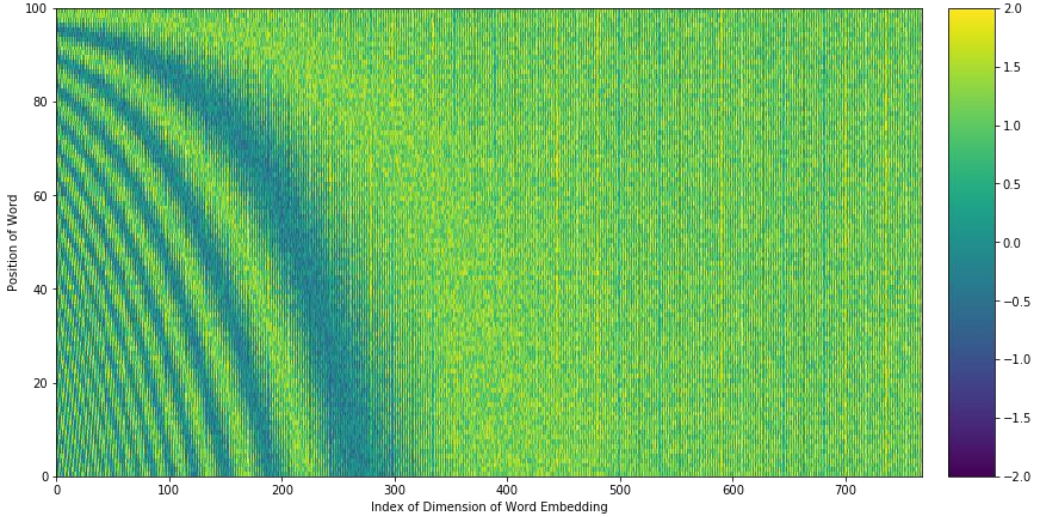
Building GPT

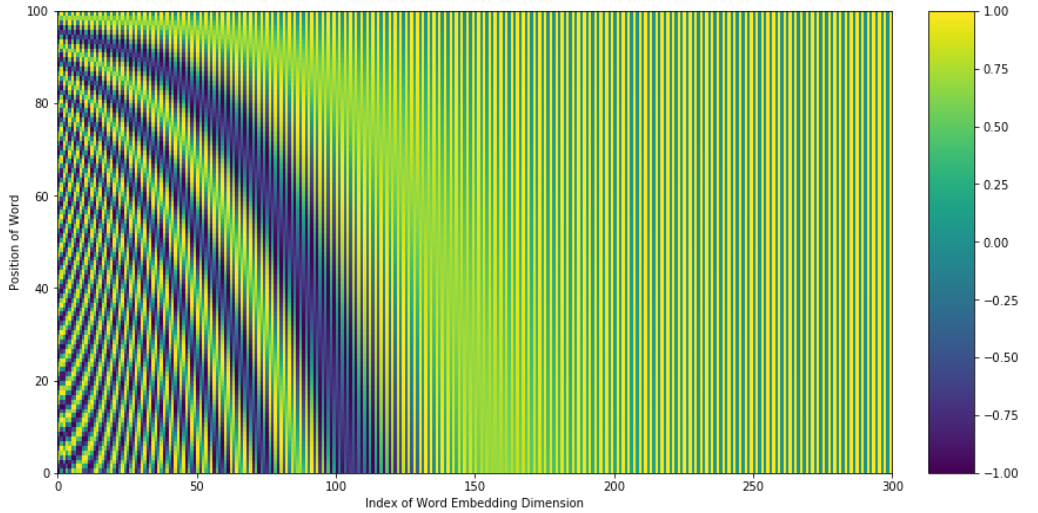
Building GPT: The Transfomer



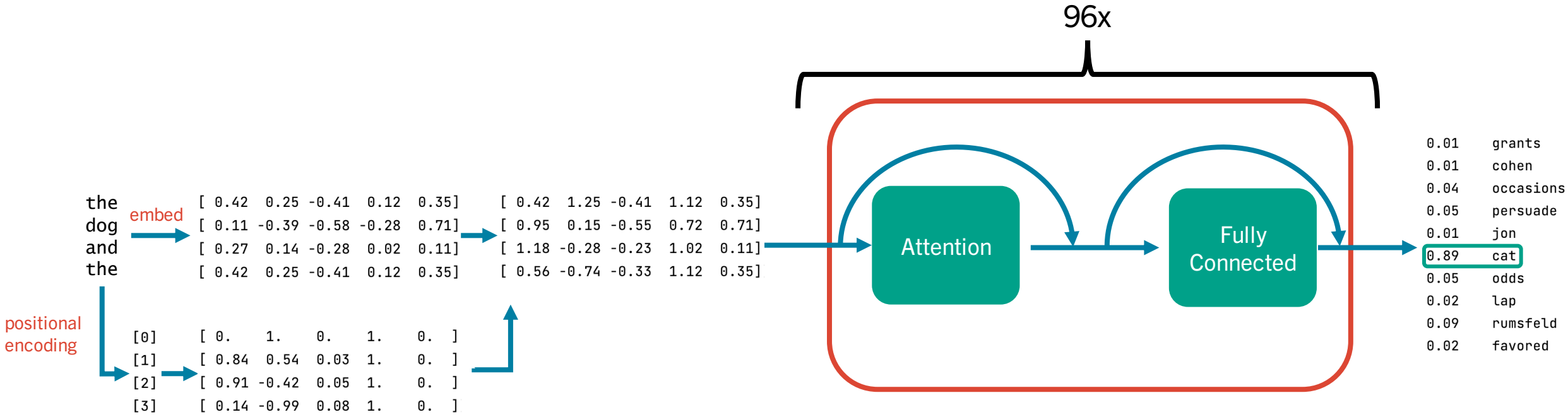
Building GPT: Positional Embedding



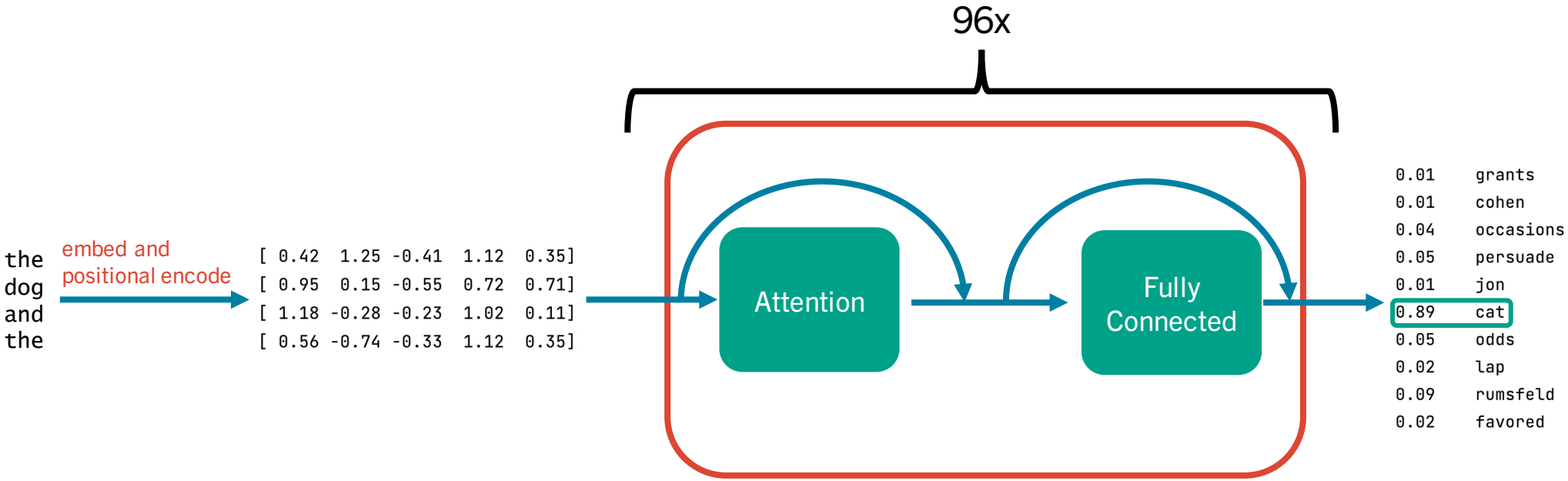




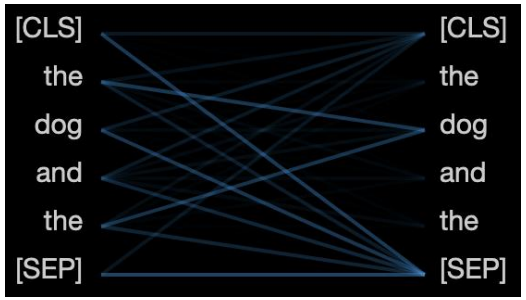
Building GPT

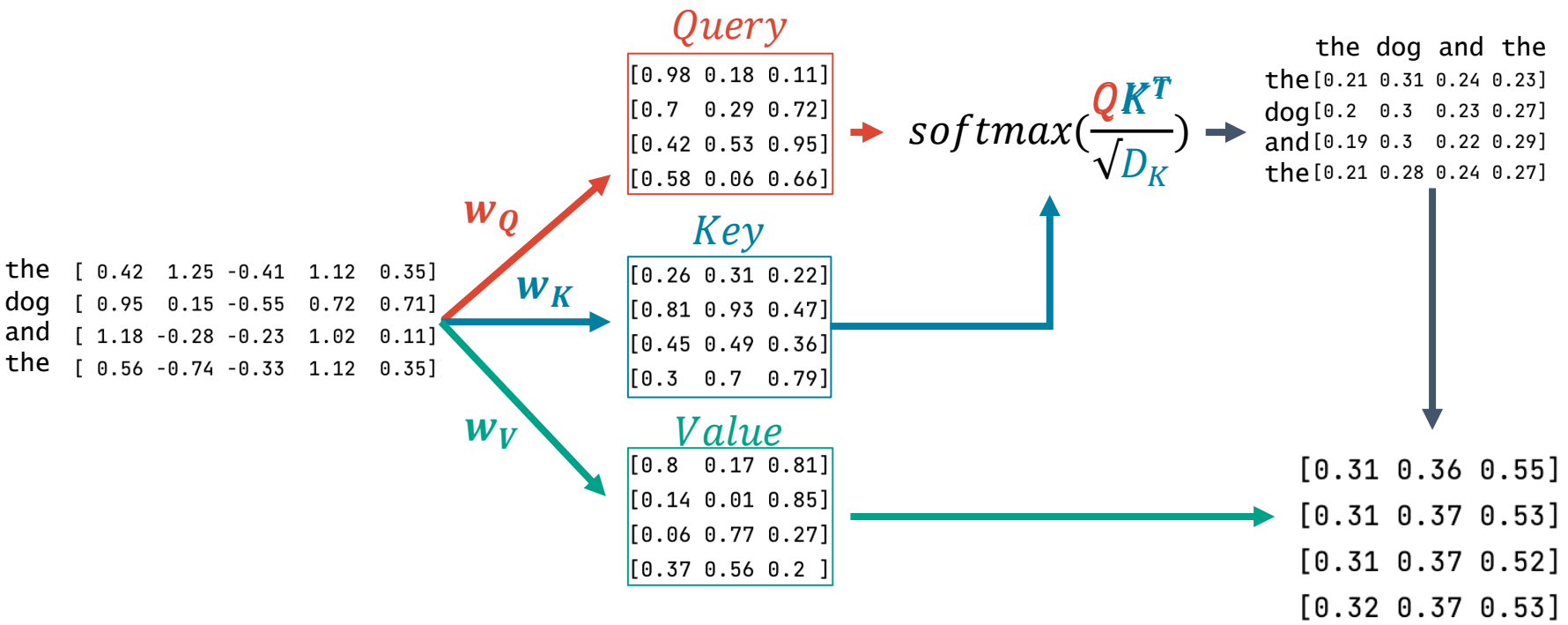


Building GPT

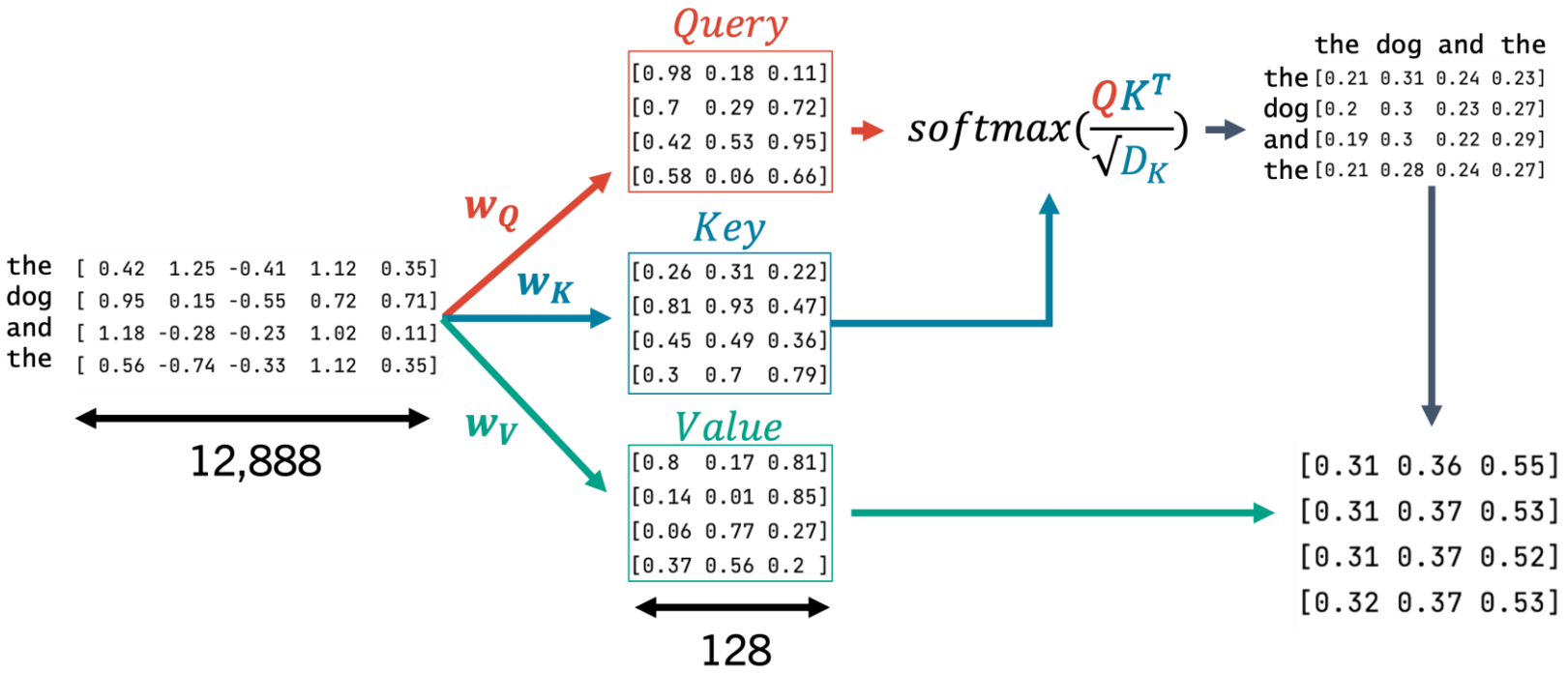


Building GPT: Attention

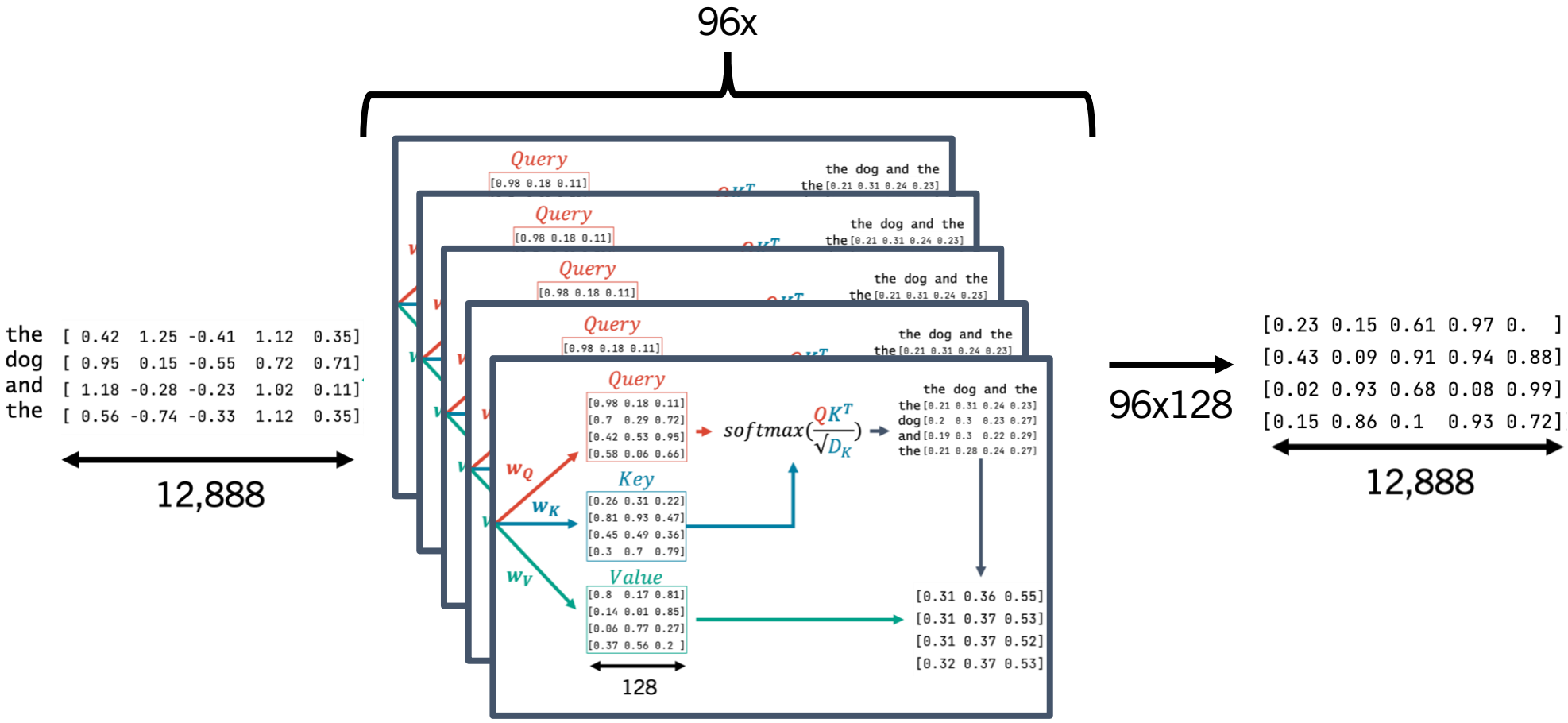




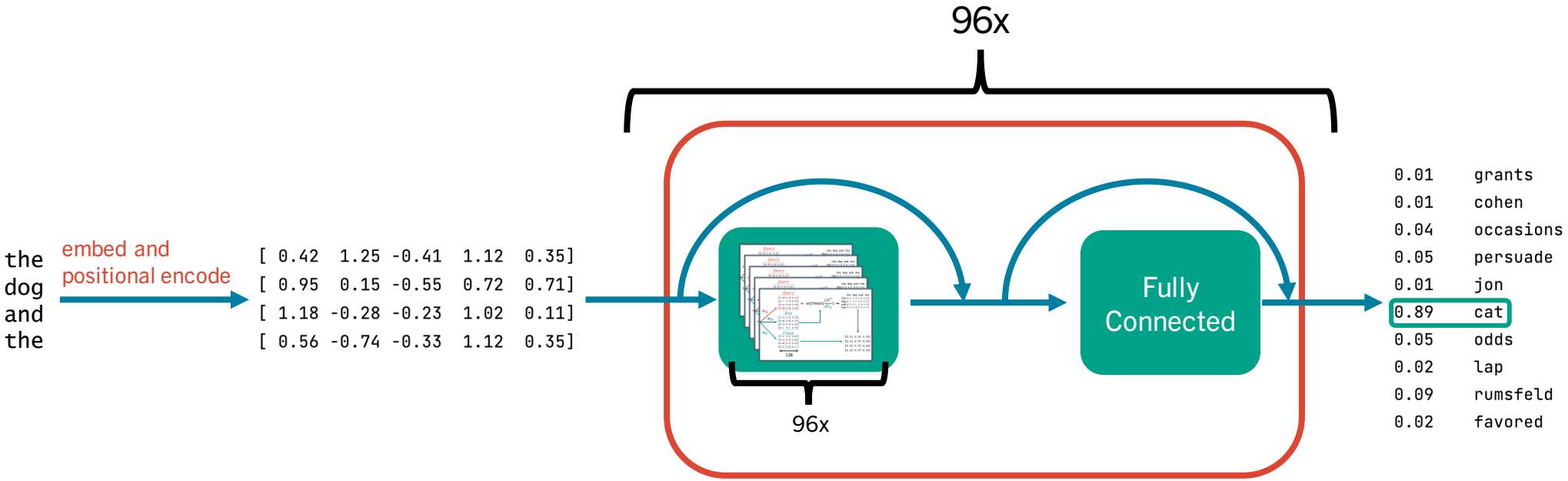
Building GPT: Attention



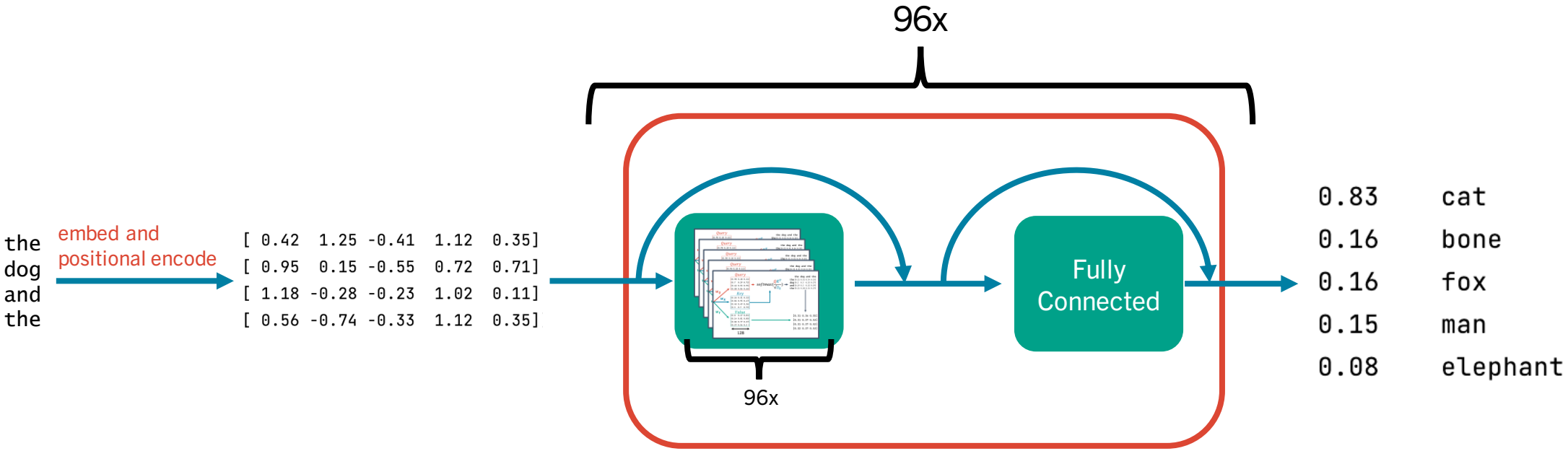
Building GPT: Attention



Building GPT



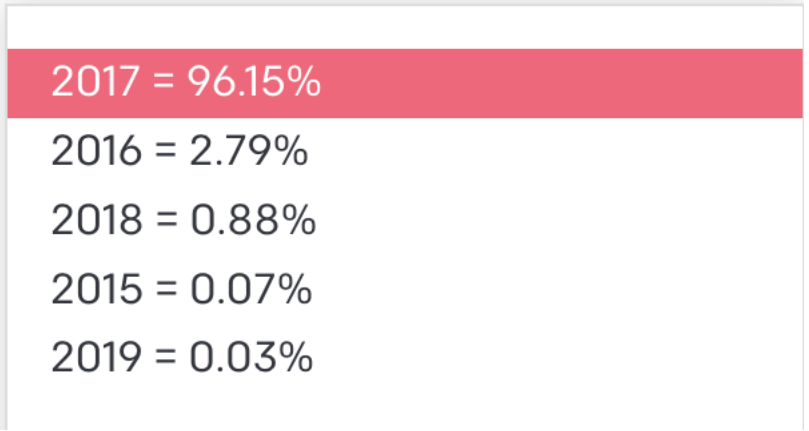
Building GPT: Top-P



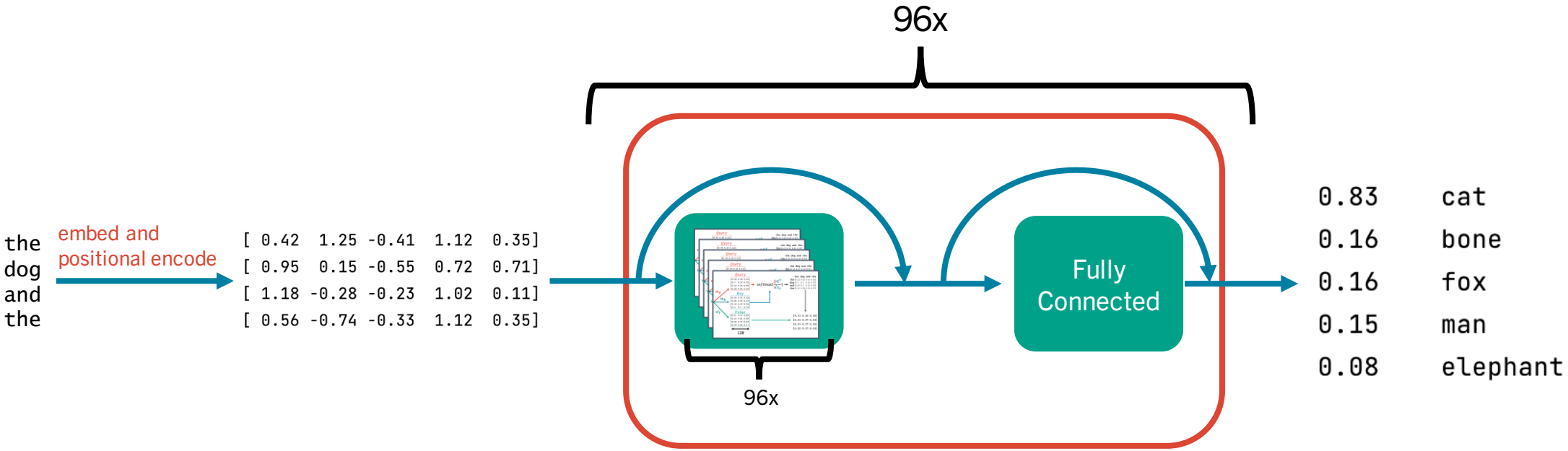
Building GPT: Top-P

Top 10 documentaries about artificial intelligence:

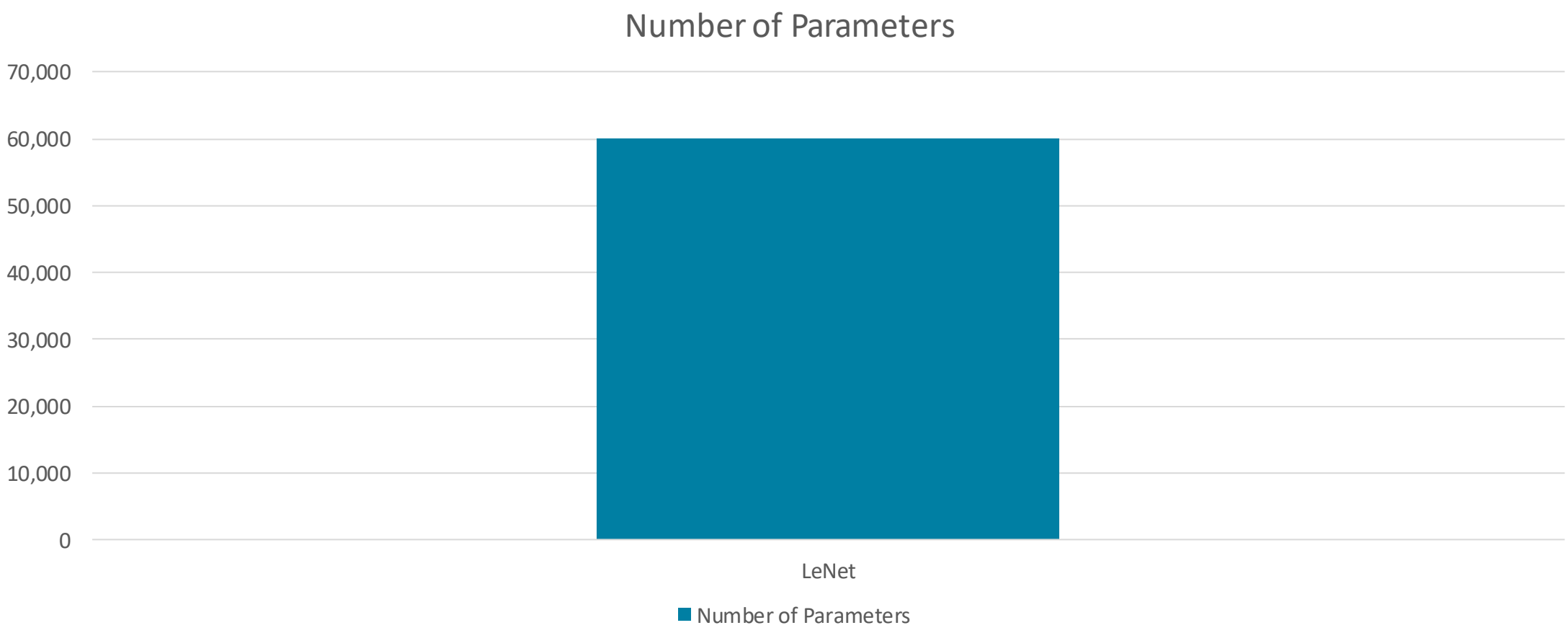
1. AlphaGo (2017)



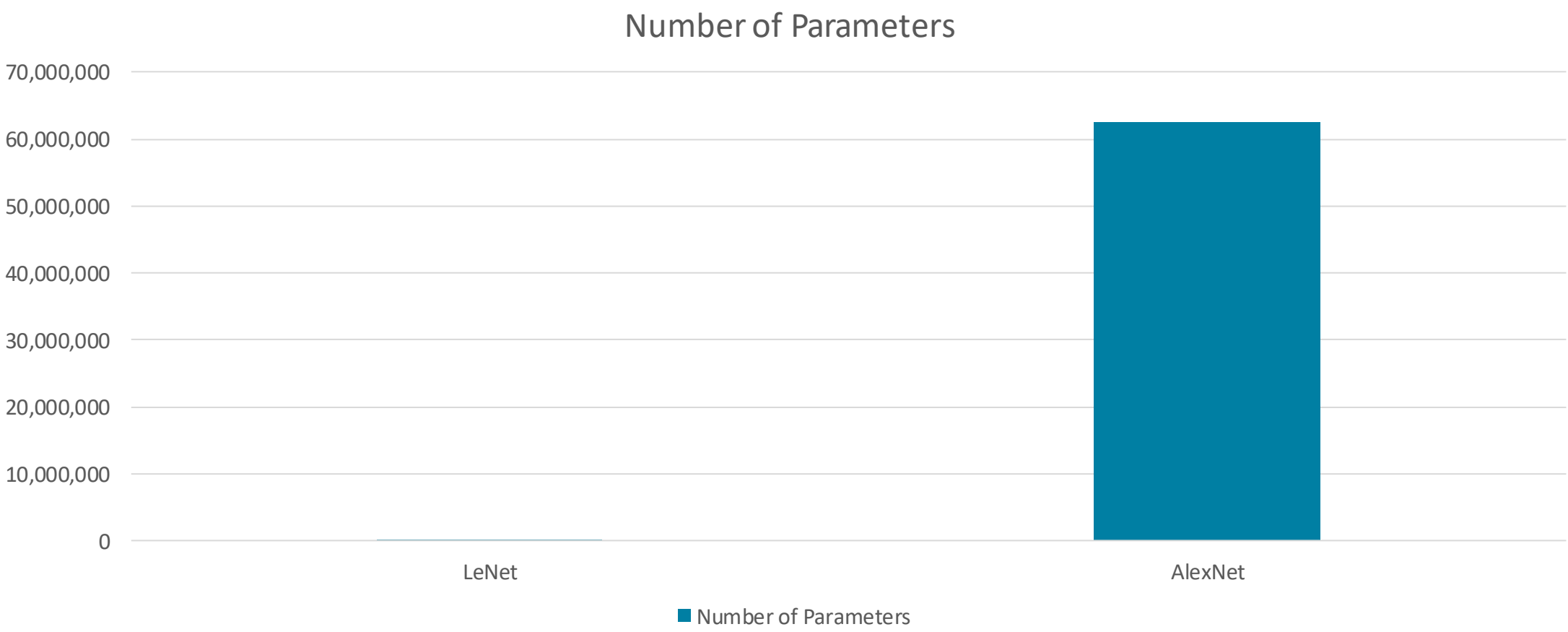
Building GPT



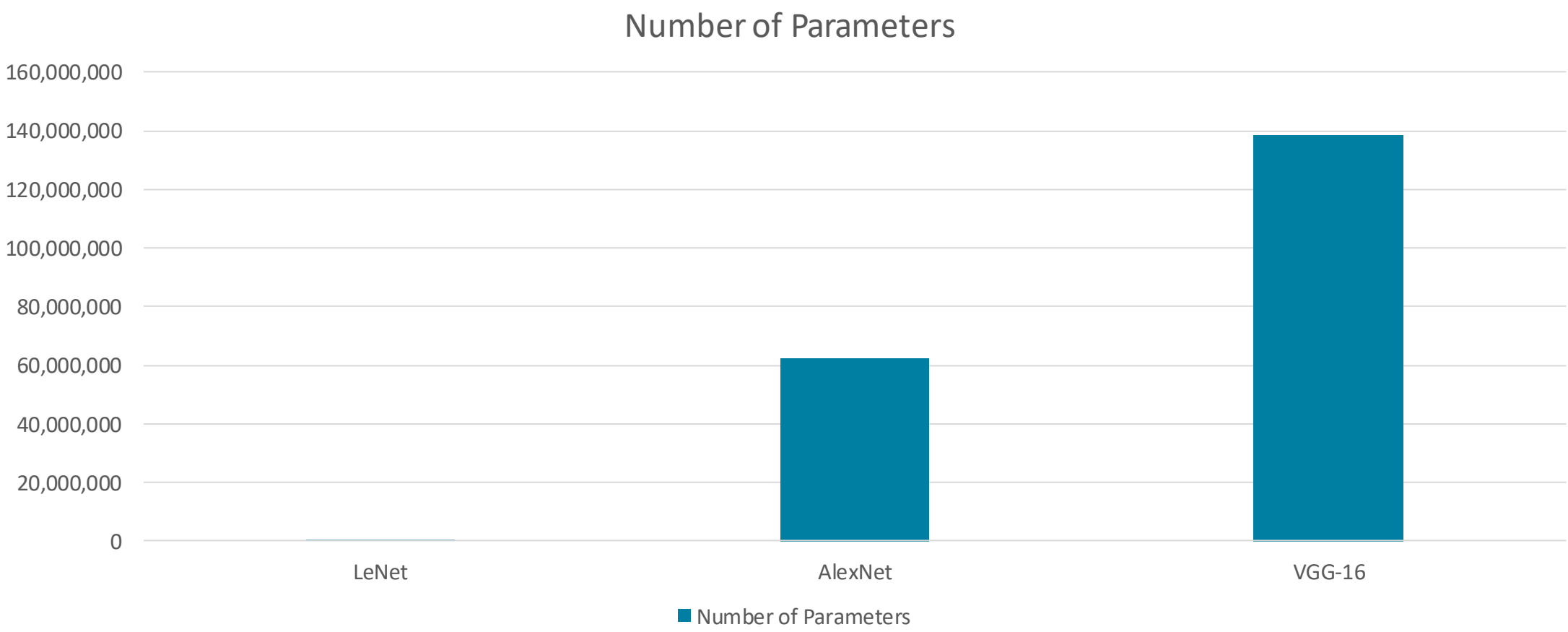
Scale of GPT


Number of Parameters
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
LeNet
Number of Parameters

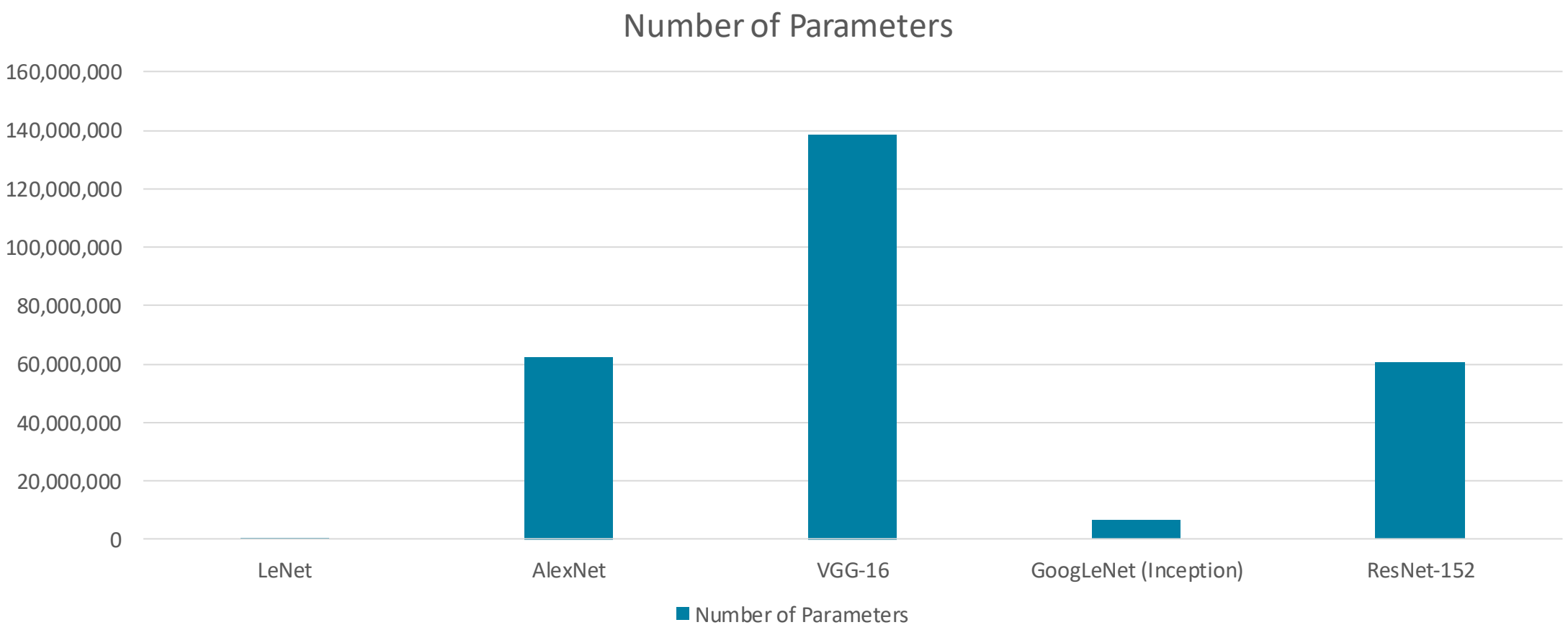
Scale of GPT


Number of Parameters
70,000,000
60,000,000
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
LeNet
AlexNet
Number of Parameters

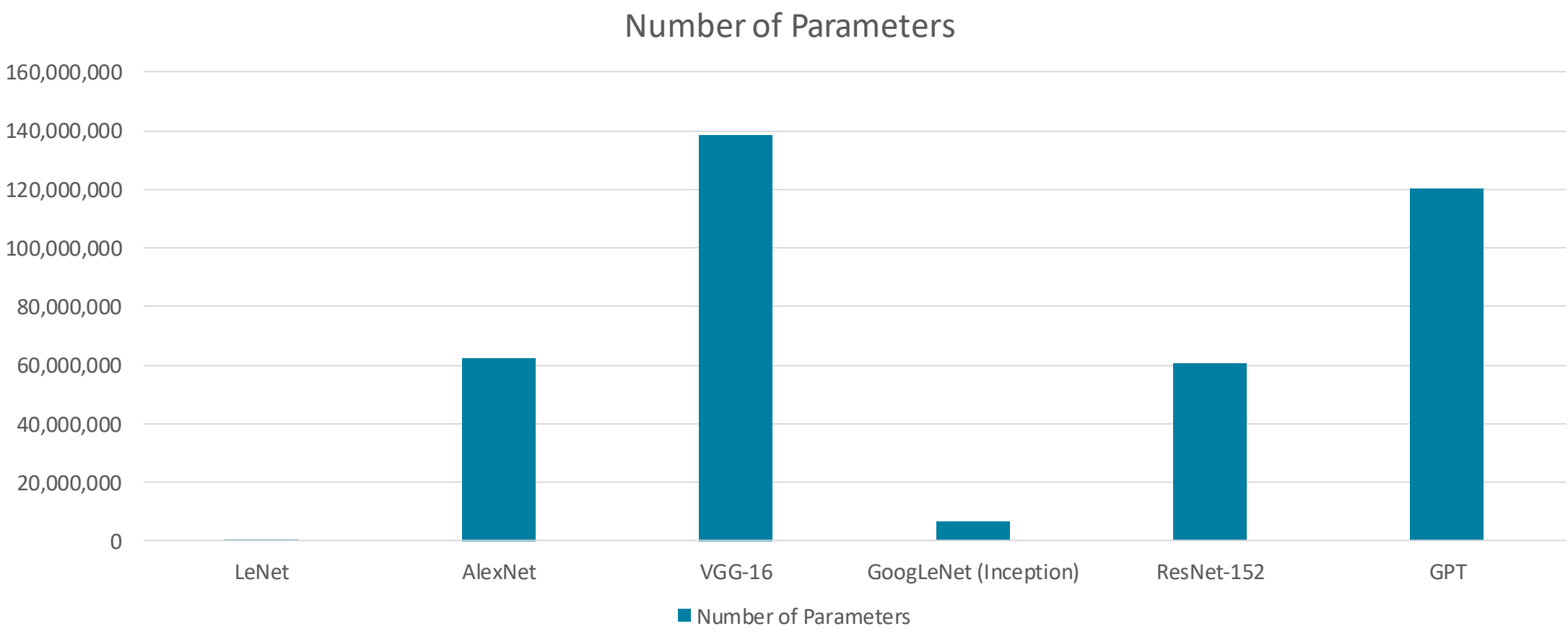
Scale of GPT


Number of Parameters
160,000,000
140,000,000
120,000,000
100,000,000
80,000,000
60,000,000
40,000,000
20,000,000
0
LeNet
AlexNet
VGG-16
Number of Parameters

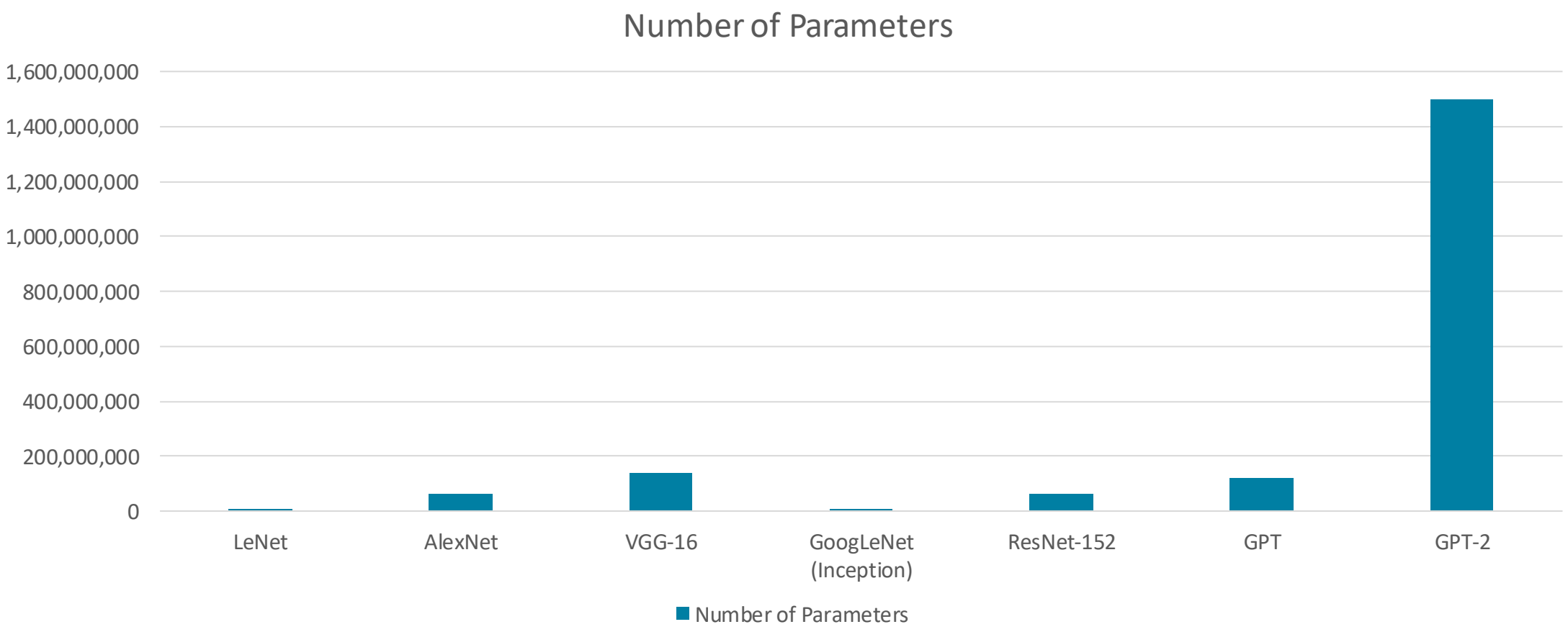
Scale of GPT


Number of Parameters
160,000,000
140,000,000
120,000,000
100,000,000
80,000,000
60,000,000
40,000,000
20,000,000
0
LeNet
AlexNet
VGG-16
GoogLeNet (Inception)
ResNet-152
Number of Parameters

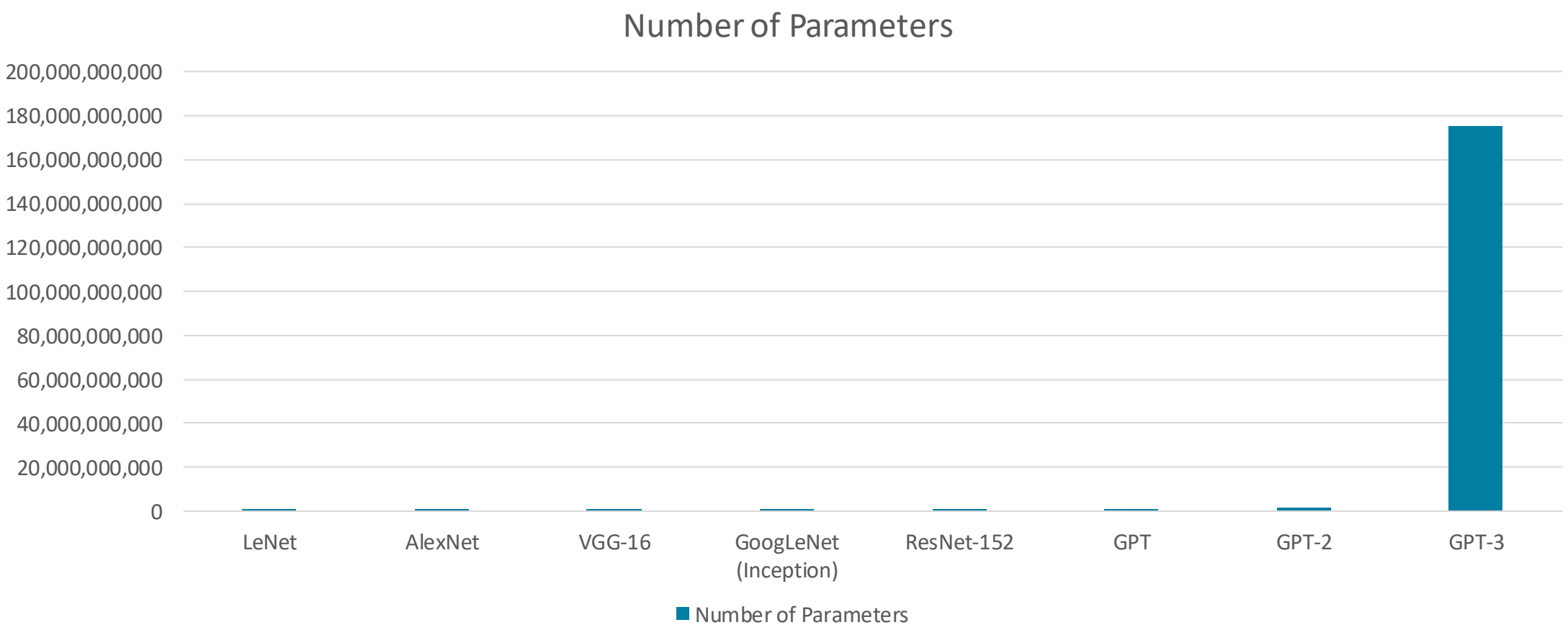
Scale of GPT


Number of Parameters
160,000,000
140,000,000
120,000,000
100,000,000
80,000,000
60,000,000
40,000,000
20,000,000
0
LeNet
AlexNet
VGG-16
GoogLeNet (Inception)
ResNet-152
GPT
Number of Parameters

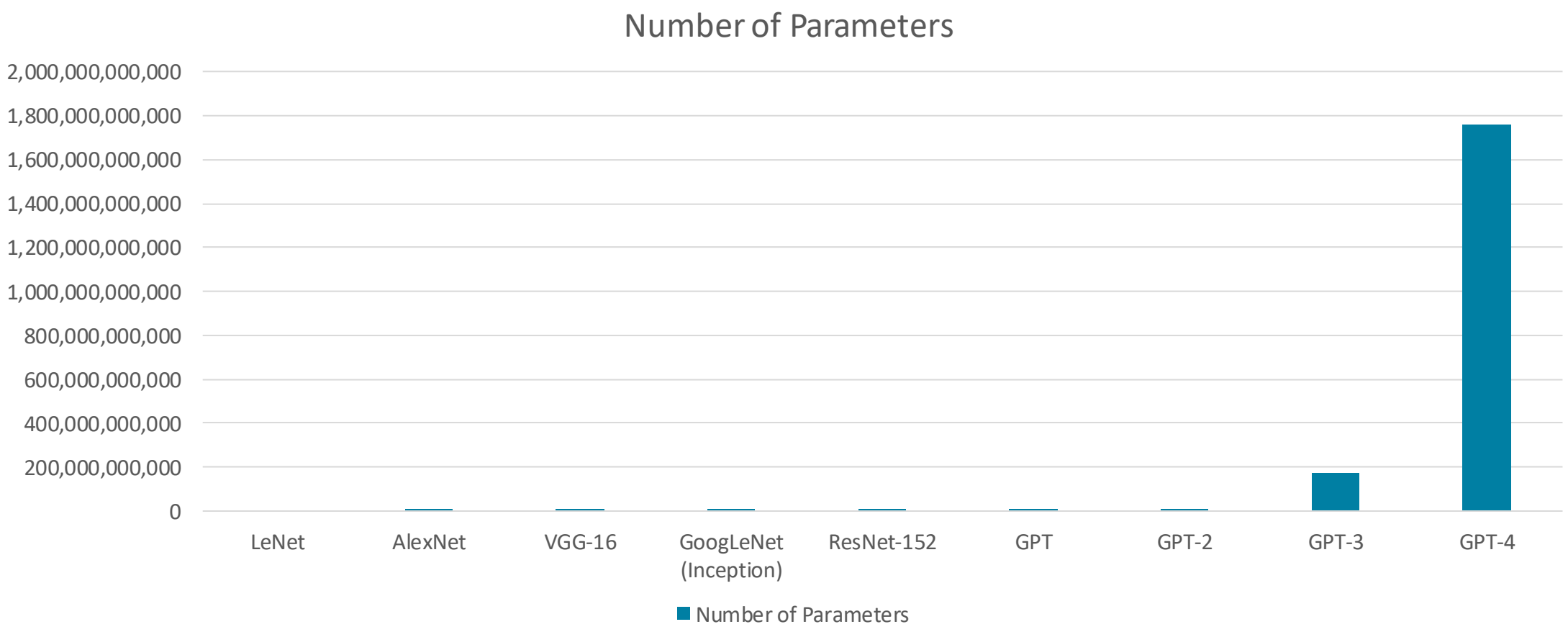
Scale of GPT


Number of Parameters
1,600,000,000
1,400,000,000
1,200,000,000
1,000,000,000
800,000,000
600,000,000
400,000,000
200,000,000
0
LeNet
AlexNet
VGG-16
GoogLeNet (Inception)
ResNet-152
GPT
GPT-2
Number of Parameters

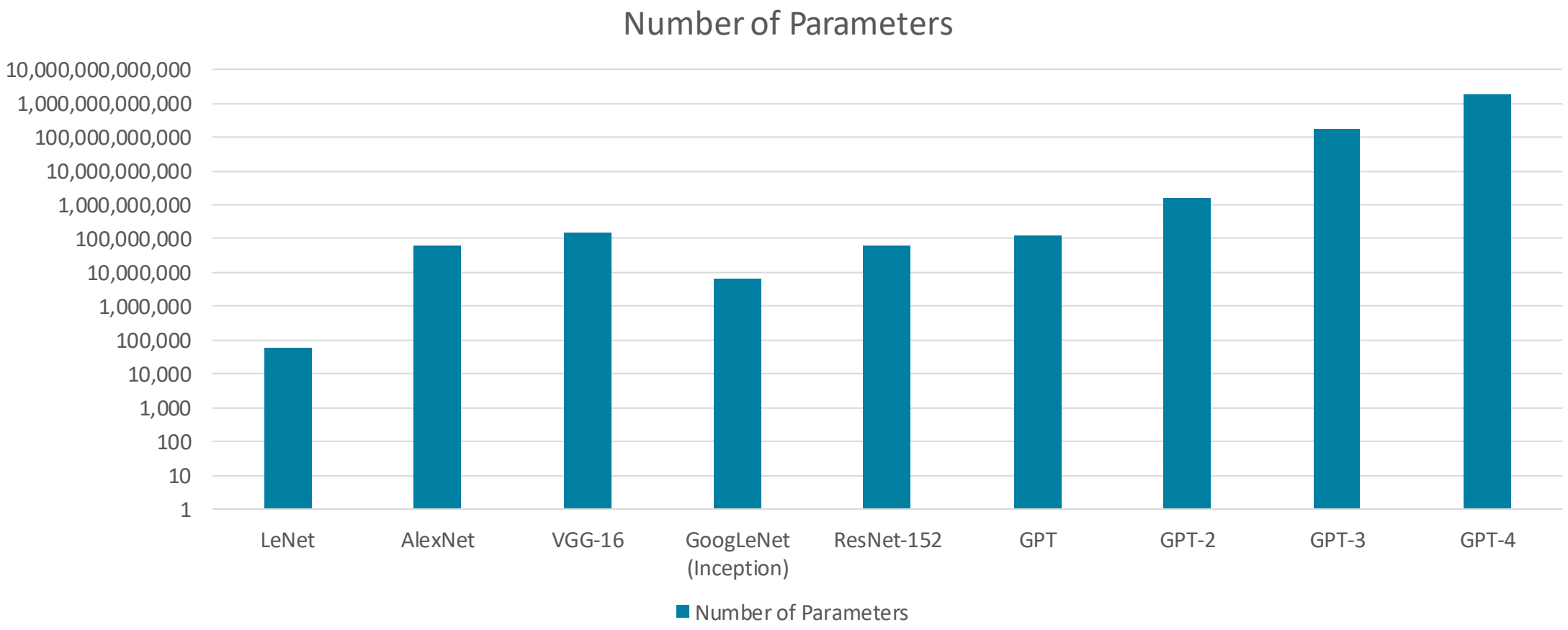
Scale of GPT


Number of Parameters
200,000,000,000
180,000,000,000
160,000,000,000
140,000,000,000
120,000,000,000
100,000,000,000
80,000,000,000
60,000,000,000
40,000,000,000
20,000,000,000
0
LeNet
AlexNet
VGG-16
GoogLeNet (Inception)
ResNet-152
GPT
GPT-2
GPT-3
Number of Parameters

Scale of GPT


Number of Parameters
2,000,000,000,000
1,800,000,000,000
1,600,000,000,000
1,400,000,000,000
1,200,000,000,000
1,000,000,000,000
800,000,000,000
600,000,000,000
400,000,000,000
200,000,000,000
0
LeNet
AlexNet
VGG-16
GoogLeNet (Inception)
ResNet-152
GPT
GPT-2
GPT-3
GPT-4
Number of Parameters

Scale of GPT


Number of Parameters
10,000,000,000,000
1,000,000,000,000
100,000,000,000
10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
LeNet
AlexNet
VGG-16
GoogLeNet (Inception)
ResNet-152
GPT
GPT-2
GPT-3
GPT-4
Number of Parameters

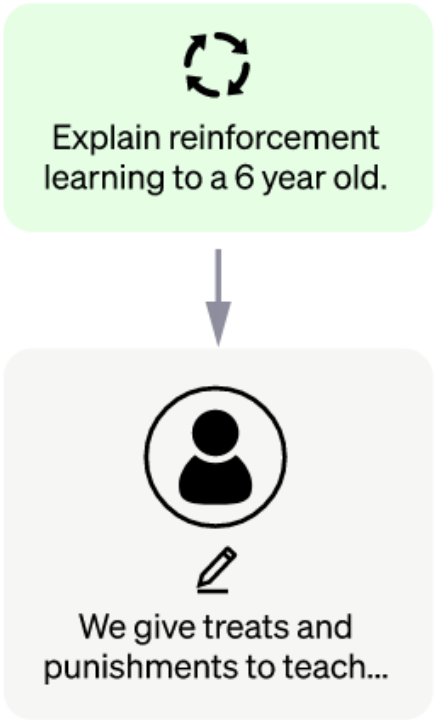
GPT's Training Data

- 1 token ≈ ¾ word
- Some datasets are sampled more times than others
- Common Crawl: billions of webpages collected over 7 years
- Webtext2: Dataset of webpages that have been shared on Reddit
- Books1: Free ebooks (?)
- Books2: Secret!
- English Wikipedia

Dataset	Quantity (tokens)	Weight in training mix

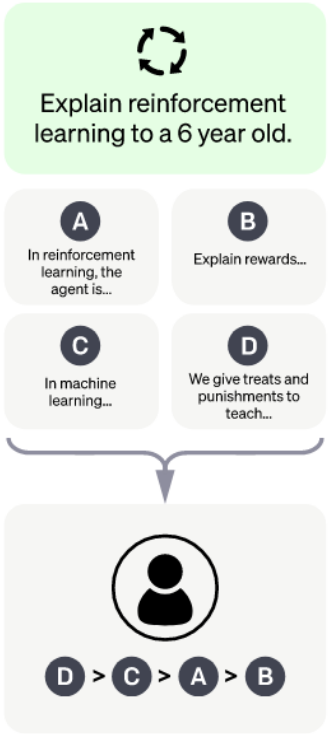
The training innovation of ChatGPT

Human annotators write answers to questions



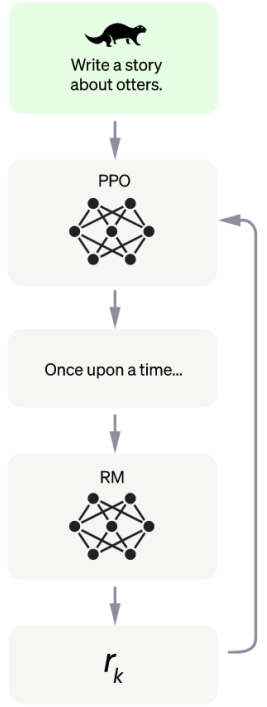
The generalist GPT model is taught from these Q&A pairs

Human annotators write more answers, and someone else ranks them



A separate model learns to rate the quality of an answer

GPT writes answers to sampled questions



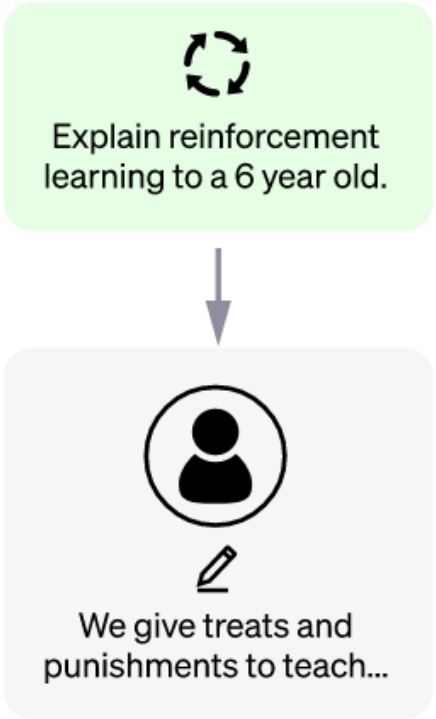
The reward model rates each answer, allowing GPT to keep learning

Figures: https://openai.com/blog/chatgpt/

The training innovation of ChatGPT

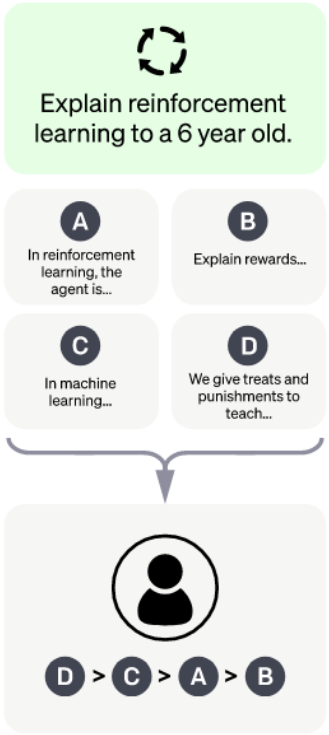
No more humans involved!

Human annotators write answers to questions



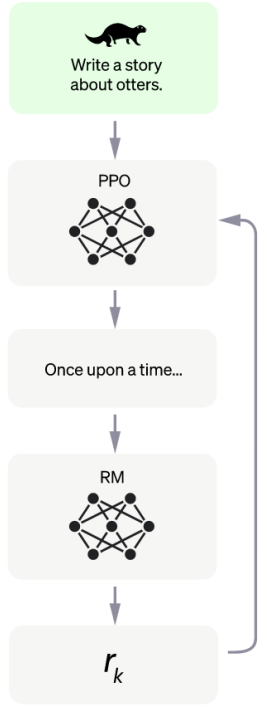
The generalist GPT model is taught from these Q&A pairs

Human annotators write more answers, and someone else ranks them



A separate model learns to rate the quality of an answer

GPT writes answers to sampled questions



The reward model rates each answer, allowing GPT to keep learning

Figures: https://openai.com/blog/chatgpt/