

NSERC CREATE for BioZone Machine Learning Bootcamp

Natural Language Processing & GPT

What is Natural Language Processing (NLP)?

- Subfield of AI that focuses on reading, deciphering and producing human language
- Combines computational linguistics (e.g. rule-based modelling of language) with statistical, ML, and deep learning approaches
- Through NLP, machines can understand, analyze and generate language in ways that are meaningful and contextually appropriate
- While LLMs have driven an explosion in interest, there are many older technologies which paved the way

Common Applications

- **Text Classification:** e.g. spam detection, news classification
- **Question Answering:** processing input and finding relevant information
- **Machine Translation**
- **Sentiment Analysis**
- **Code Generation**

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

Quick history of NLP

- **1950s:**

- Alan Turing publishes “Computing Machinery and Intelligence”, in which he proposes the Turing test
- Noam Chomsky publishes “Syntactic Structures”, an attempt to construct a formal theory of linguistic structure

- **1960s:**

- Georgetown University develops a machine translation system, which automatically translates 60 Russian sentences into English using an extremely complex flowchart and a limited vocabulary
- The authors claim machine translation could be a solved problem in five years

Quick history of NLP

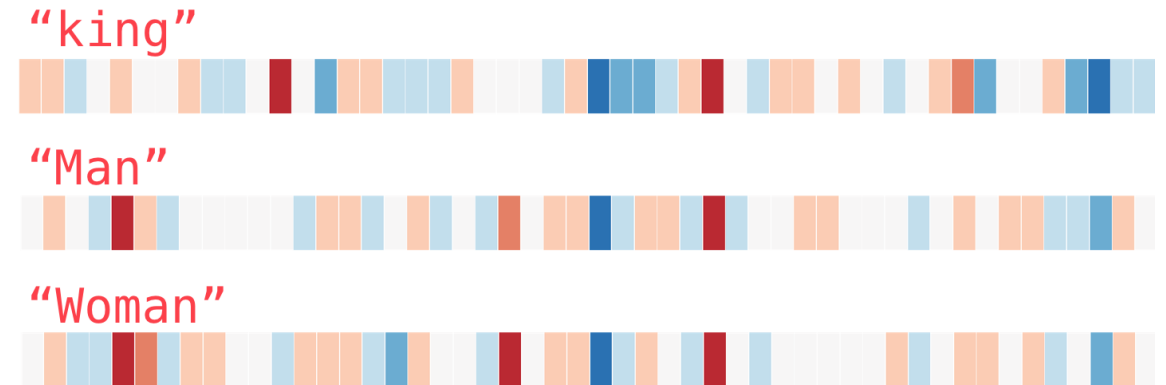
- **1980s – 1990s:**
 - Transition from rule-based to statistical approaches
 - Idea of using existing text to train a model begins to appear (e.g. bilingual documents from the Canadian parliament)
- **2000s – 2010s:**
 - Deep Learning takes over
 - Tools such as convolutional networks, and later RNNs, transform the field
- **2020s: The era of the Large Language Model**

Foundations of LLMs

- Fundamental goal of language modelling: next word prediction
- $P(\textit{cat} \mid \textit{the dog and the})$
- 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!

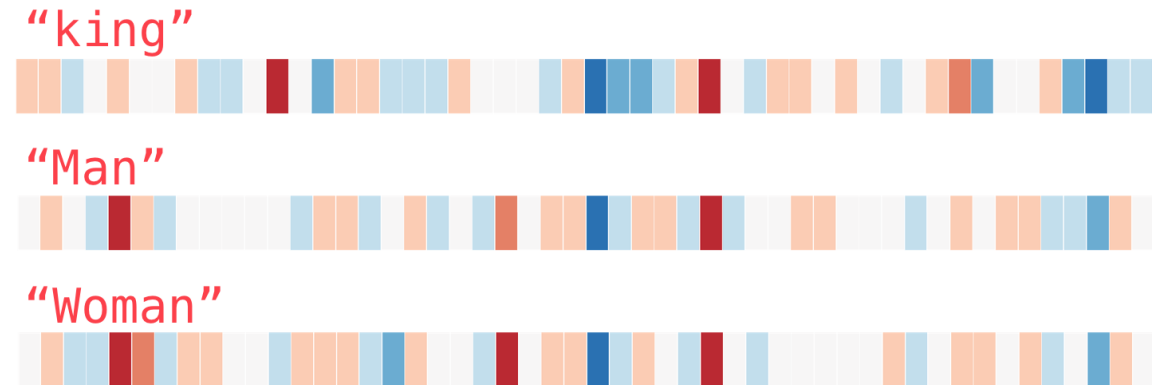
How does an LM “understand” word meaning?

- 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

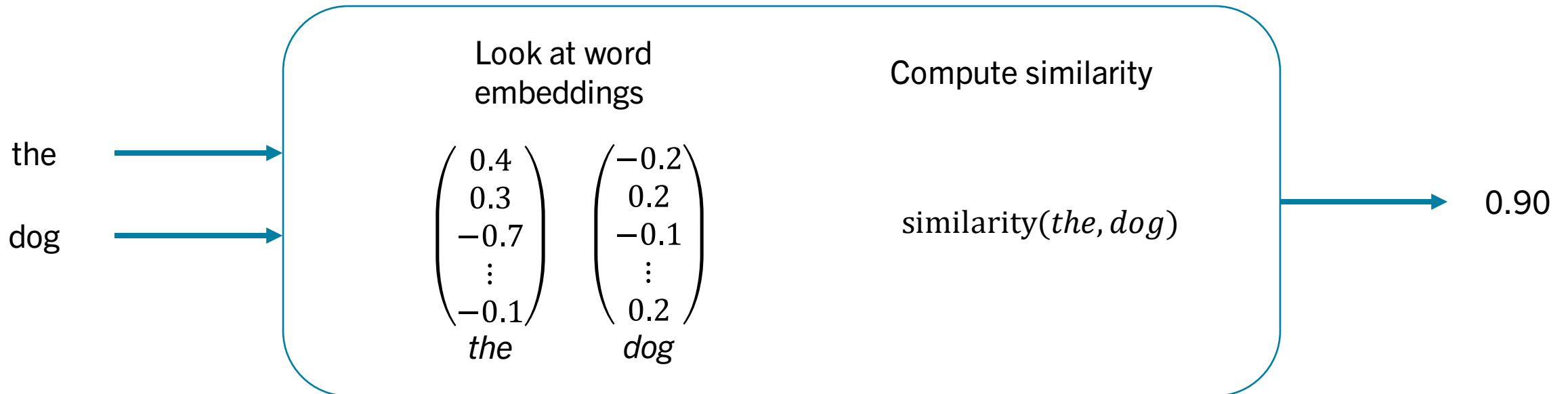


How does an LM “understand” word meaning?

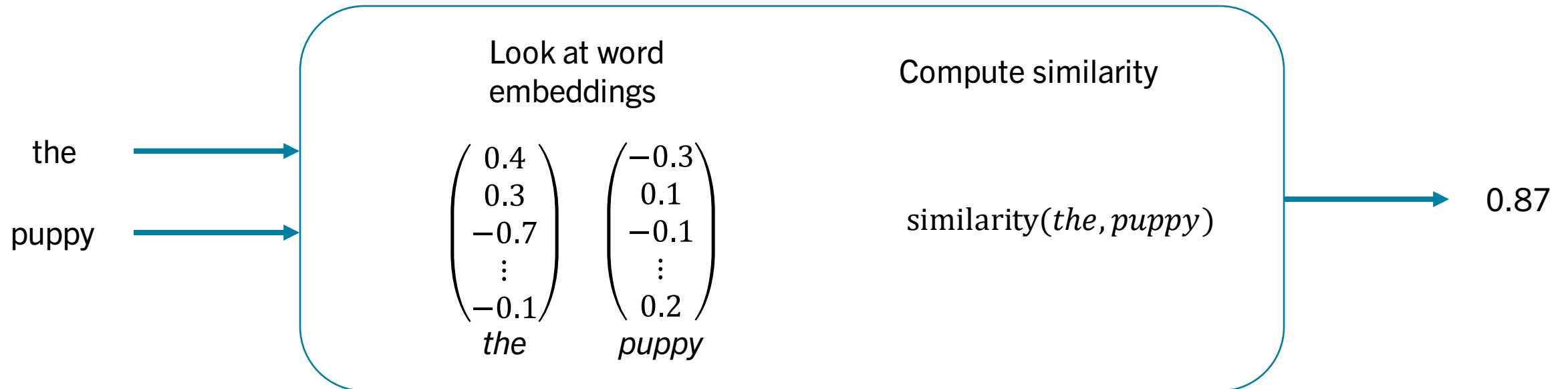
- Key concept of word embeddings: similar words should have similar vectors
- How do we accomplish this?



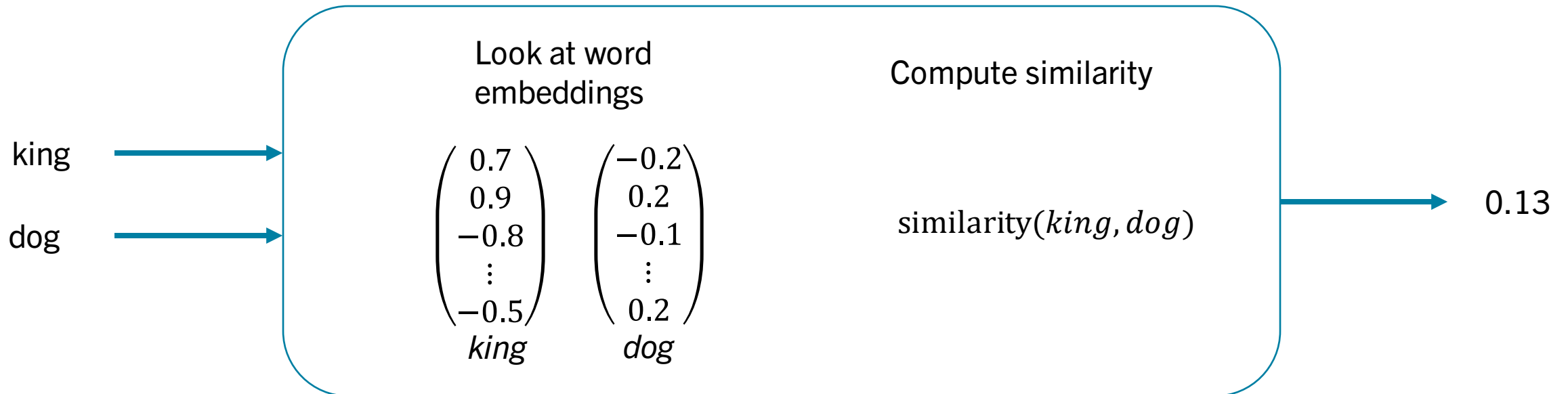
Building Word Embeddings



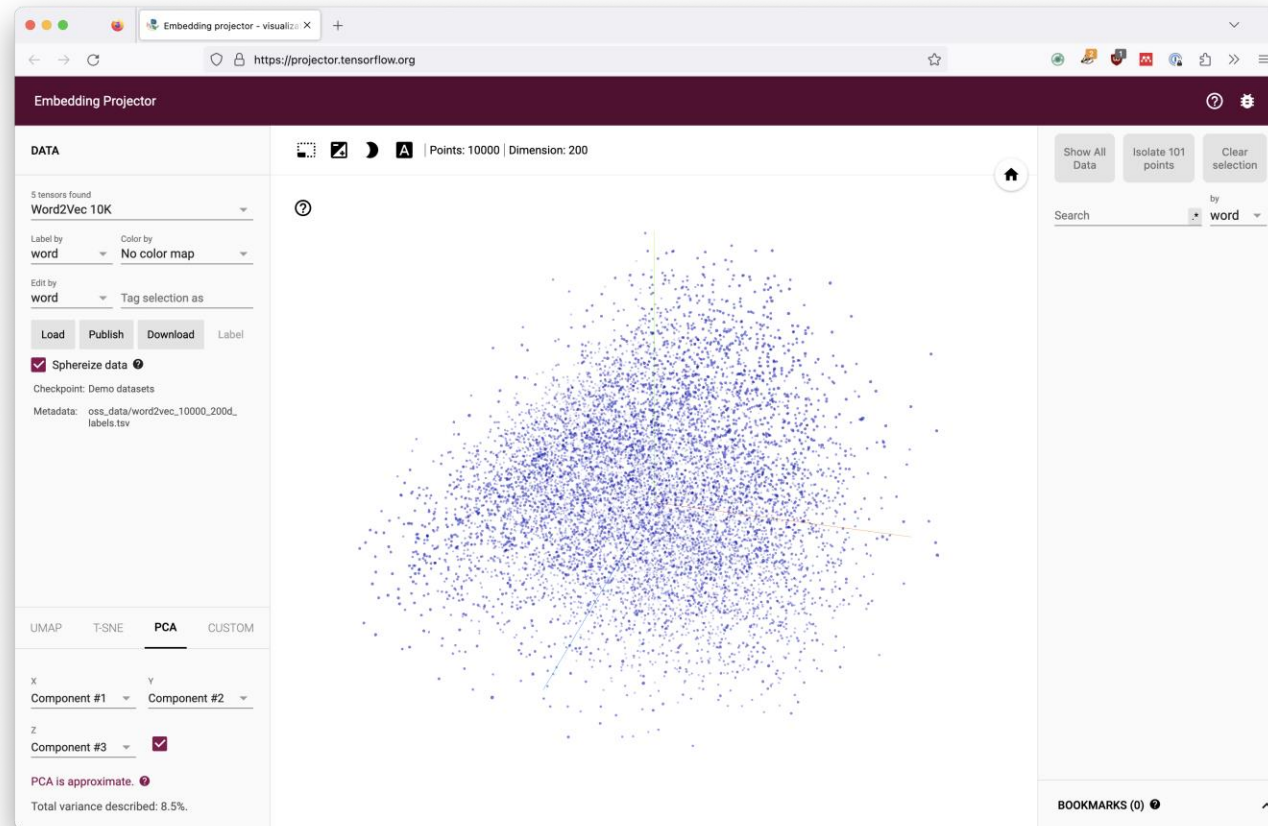
Building Word Embeddings



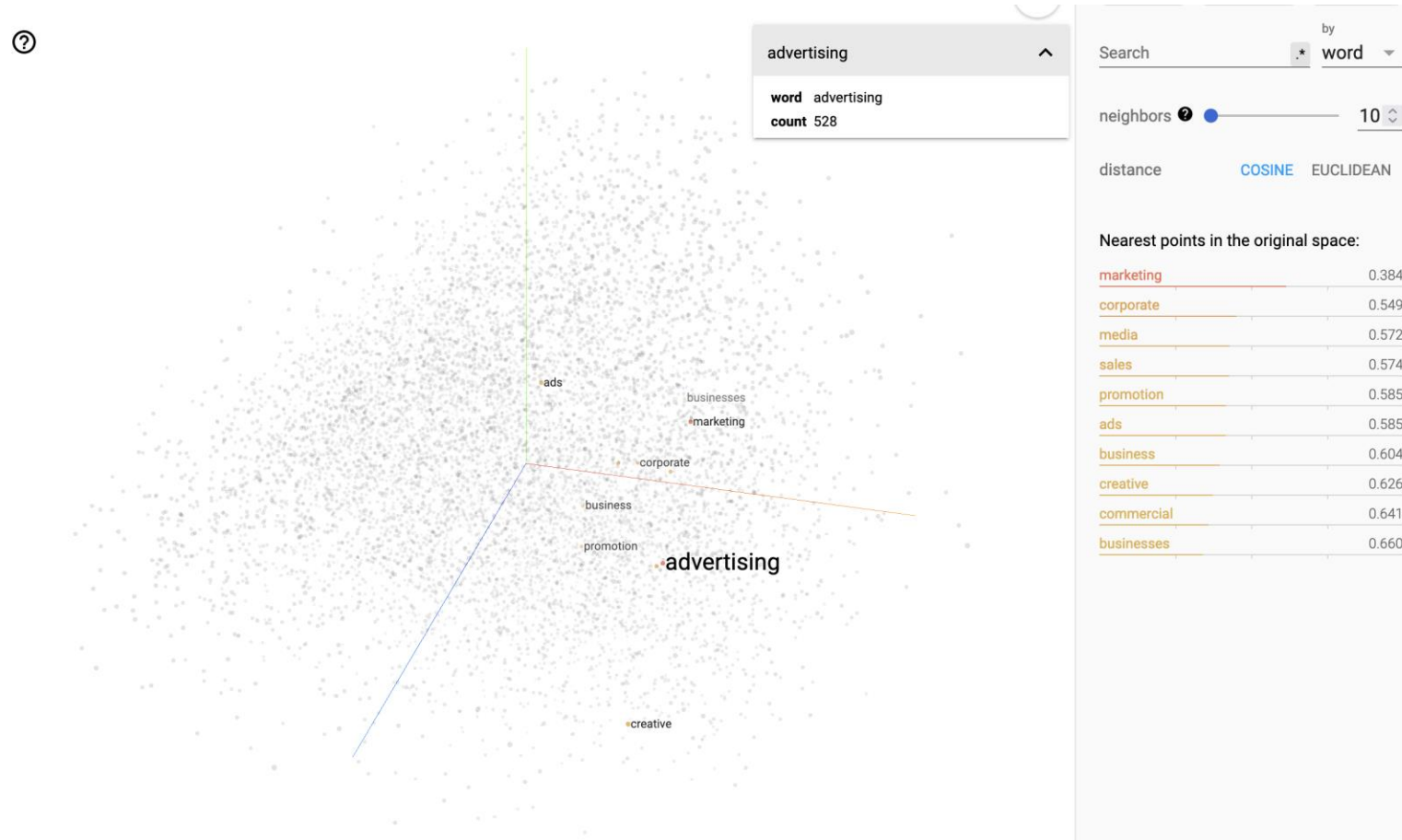
Building Word Embeddings



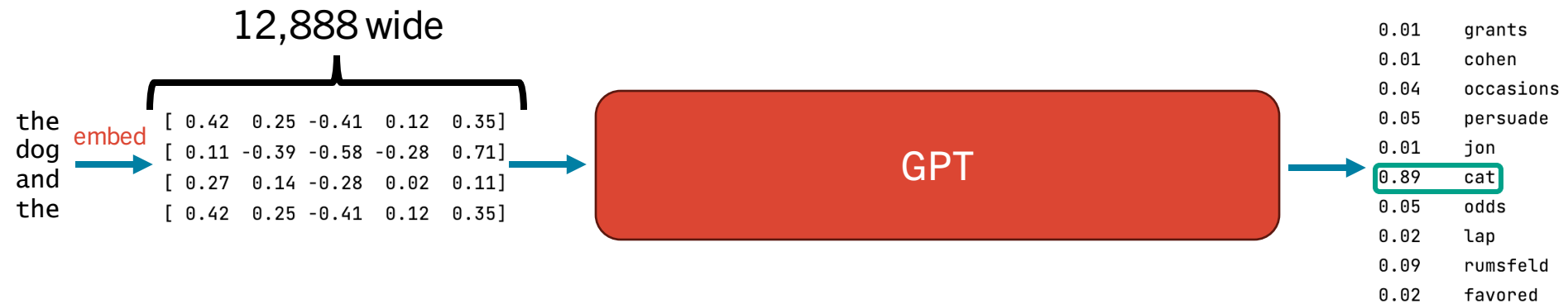
Results



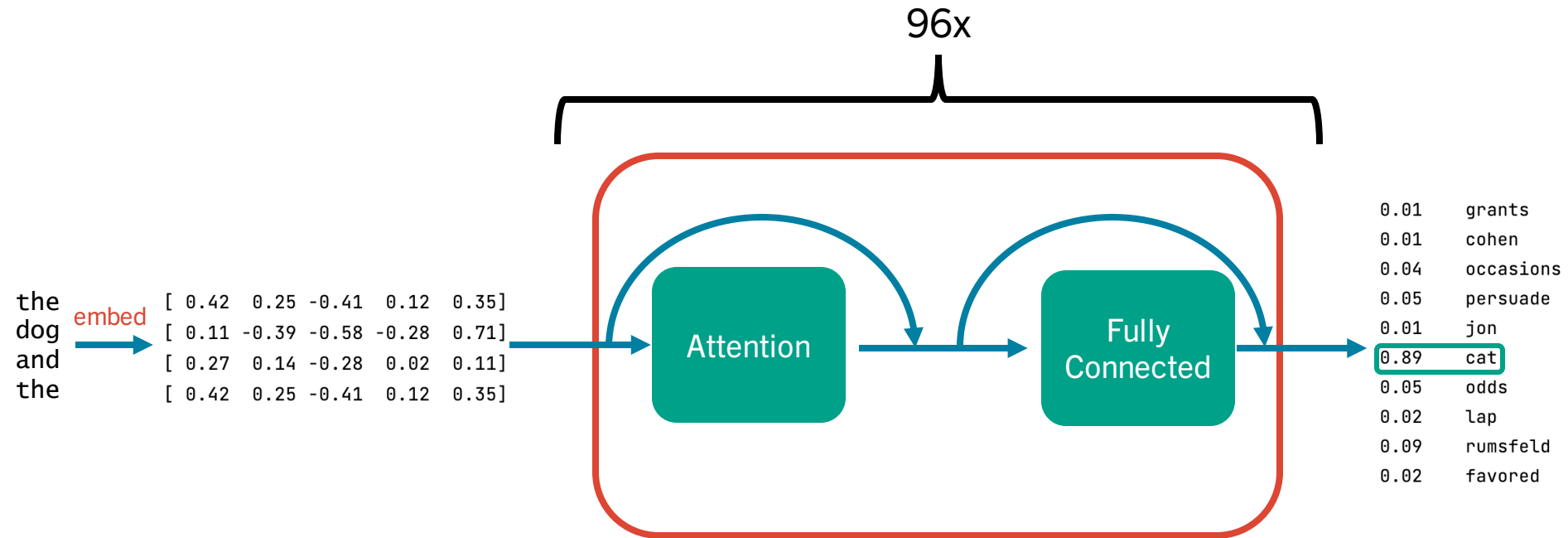
Results



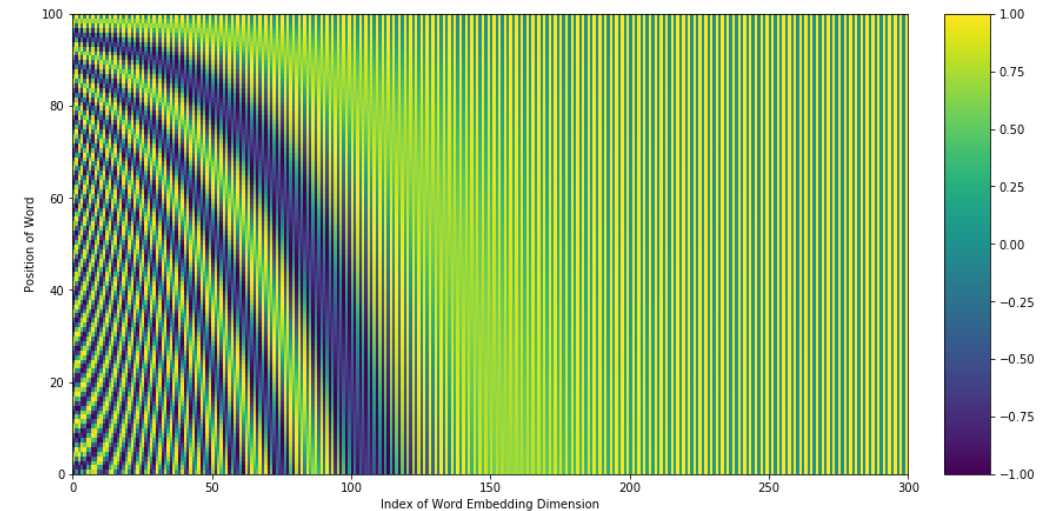
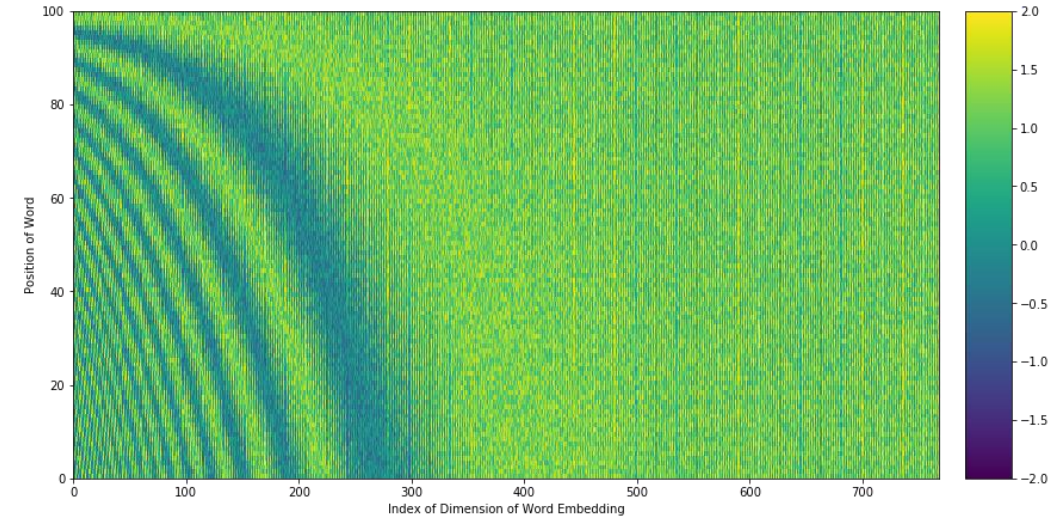
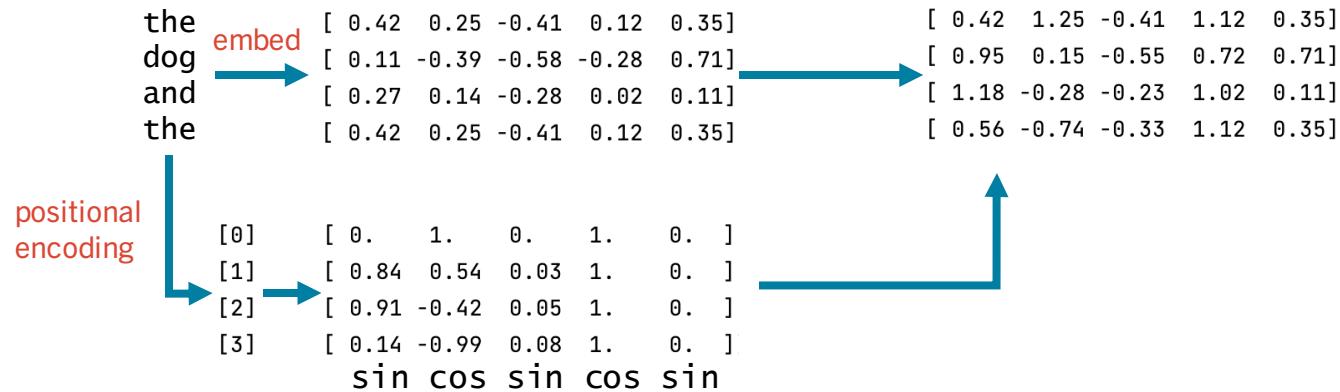
Building GPT



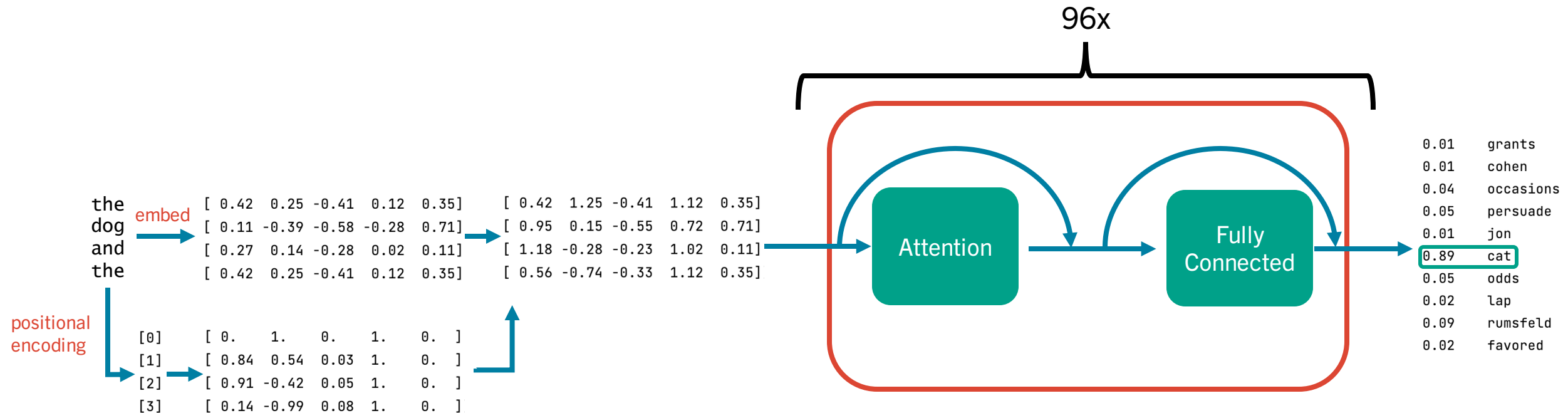
Building GPT: The Transformer



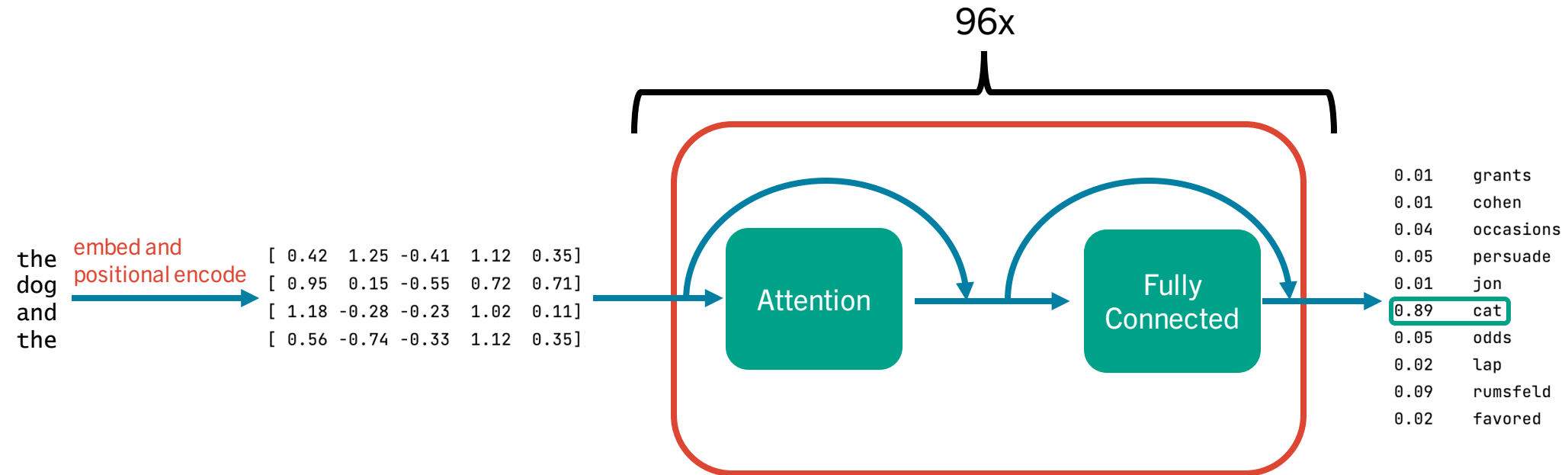
Building GPT: Positional Embedding



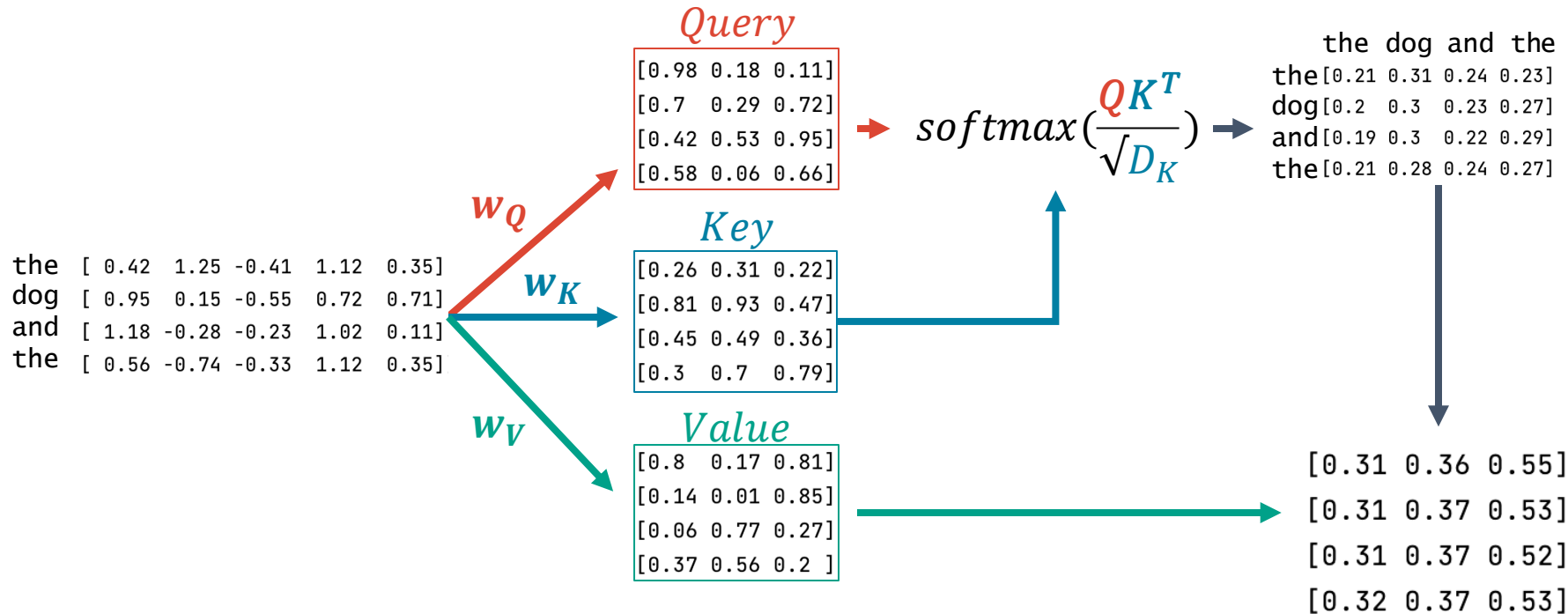
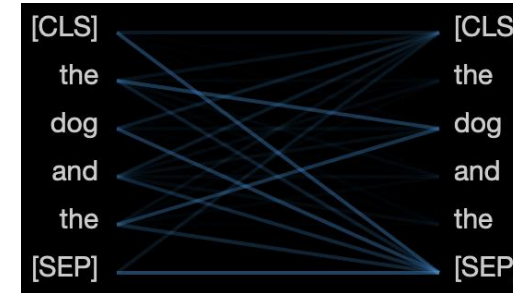
Building GPT



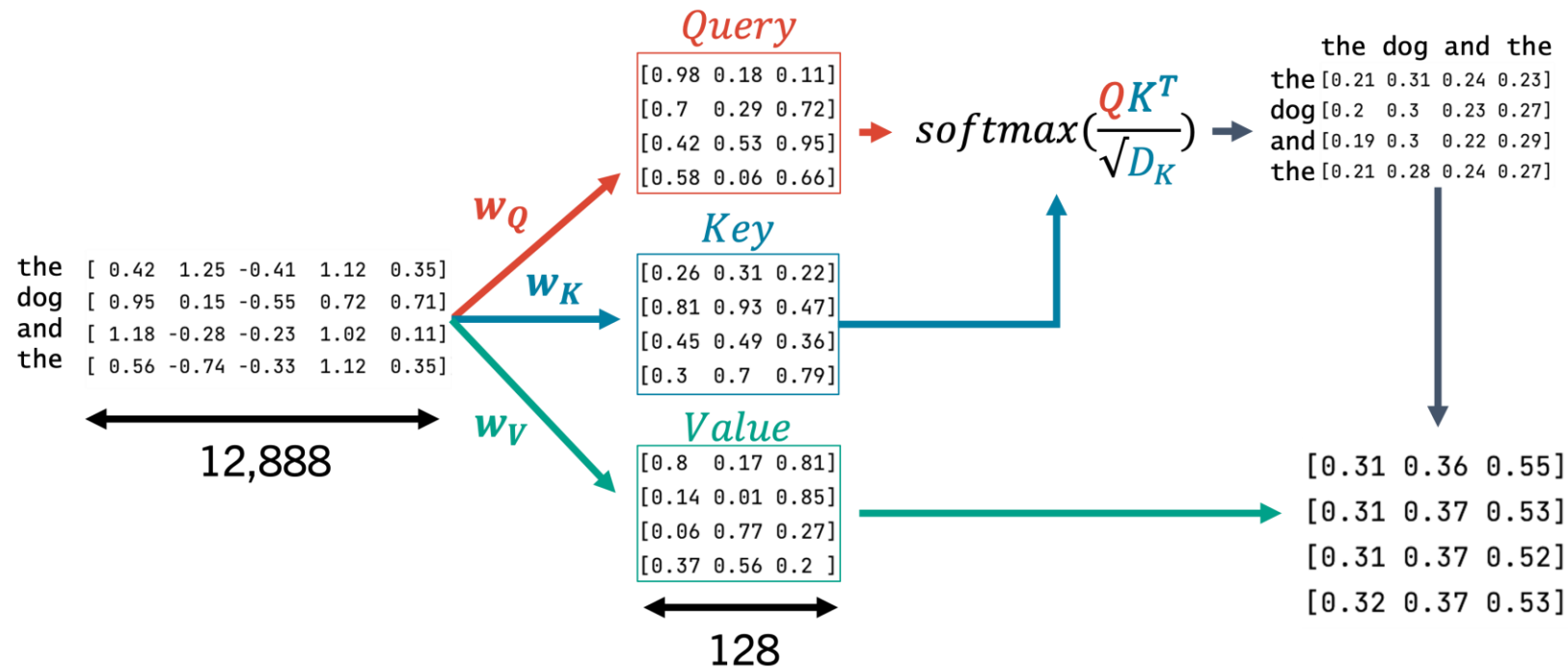
Building GPT



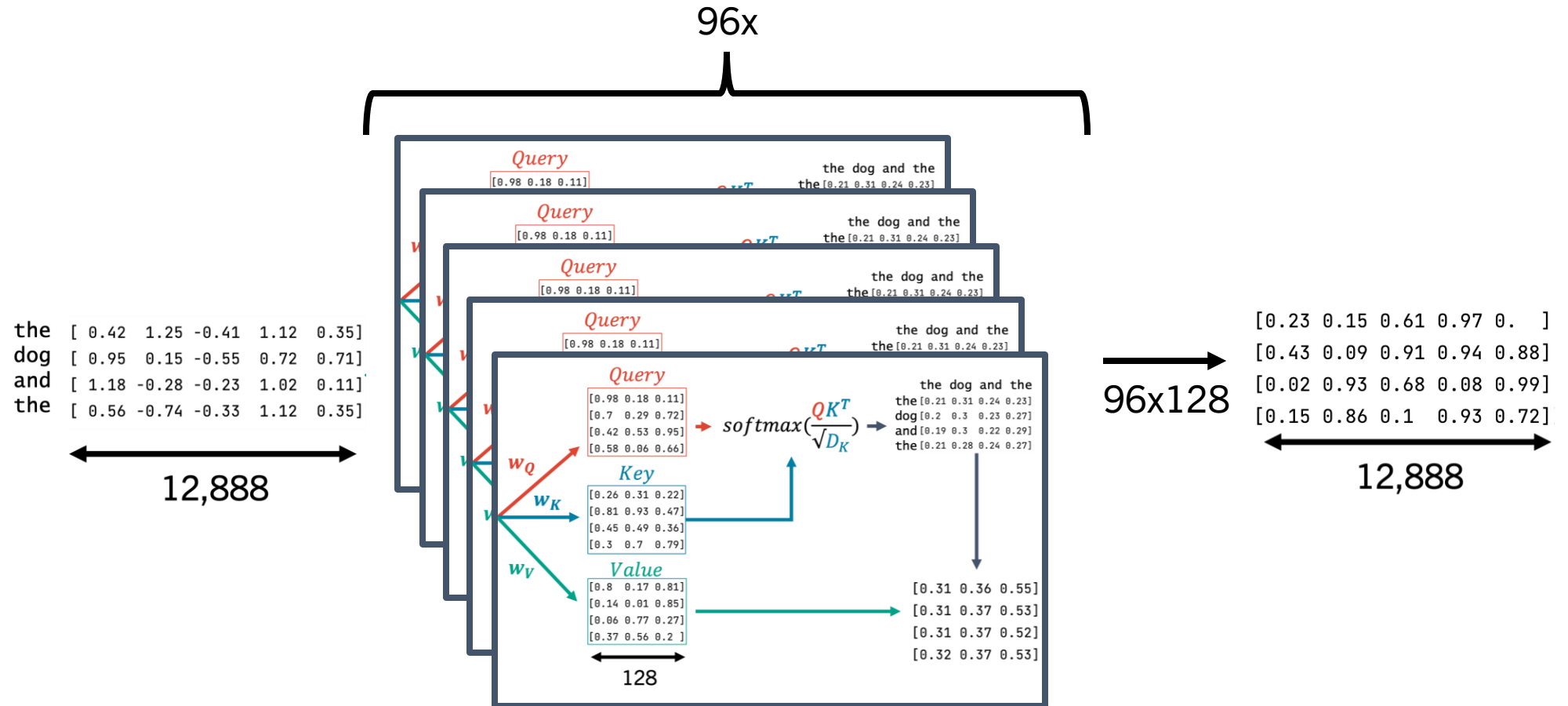
Building GPT: Attention



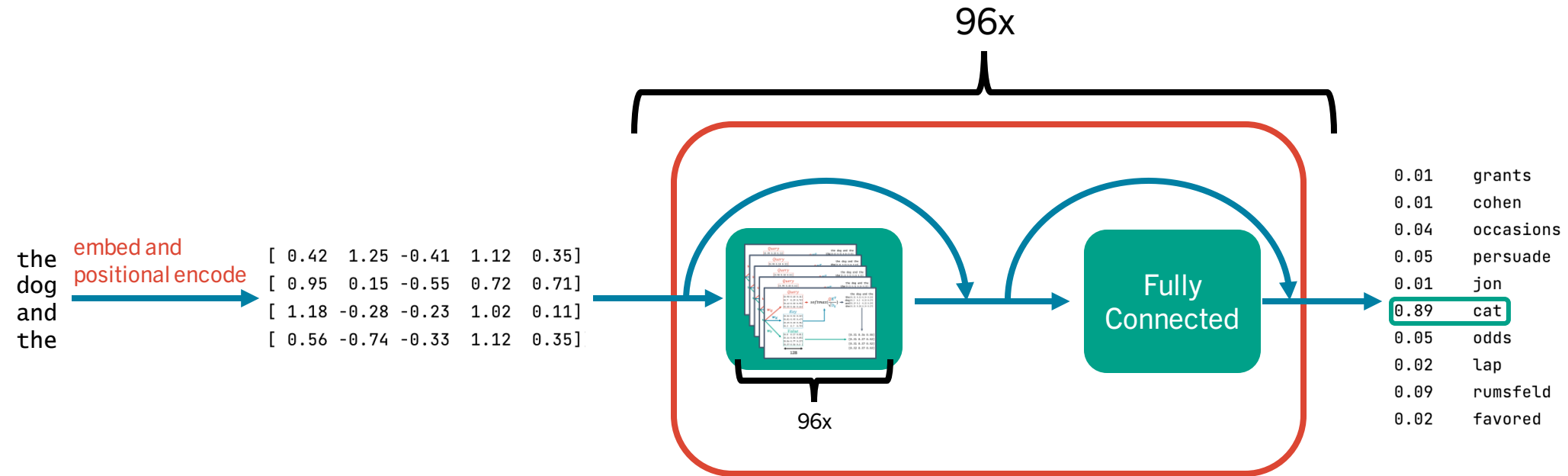
Building GPT: Attention



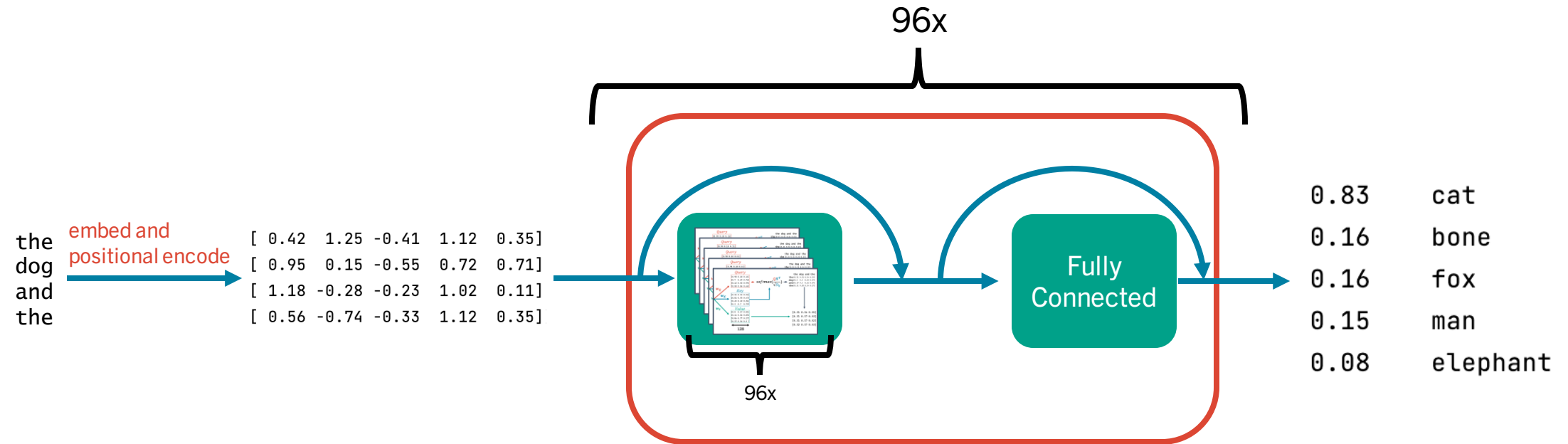
Building GPT: Attention



Building GPT



Building GPT: Top-P



Building GPT: Top-P

Top 10 documentaries about artificial intelligence:

1. AlphaGo (2017)

2017 = 96.15%

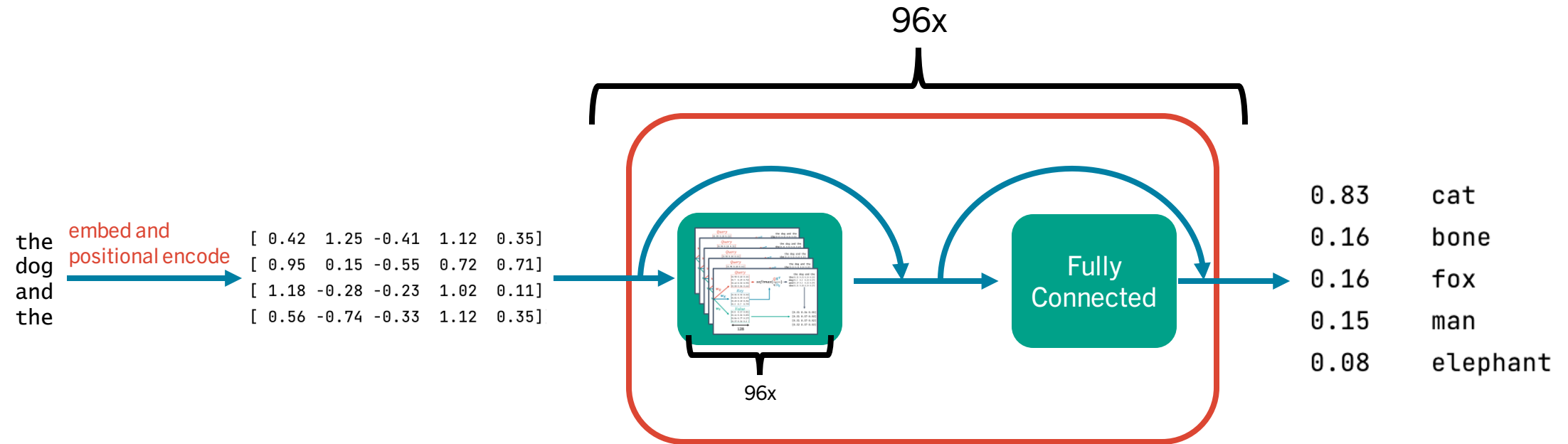
2016 = 2.79%

2018 = 0.88%

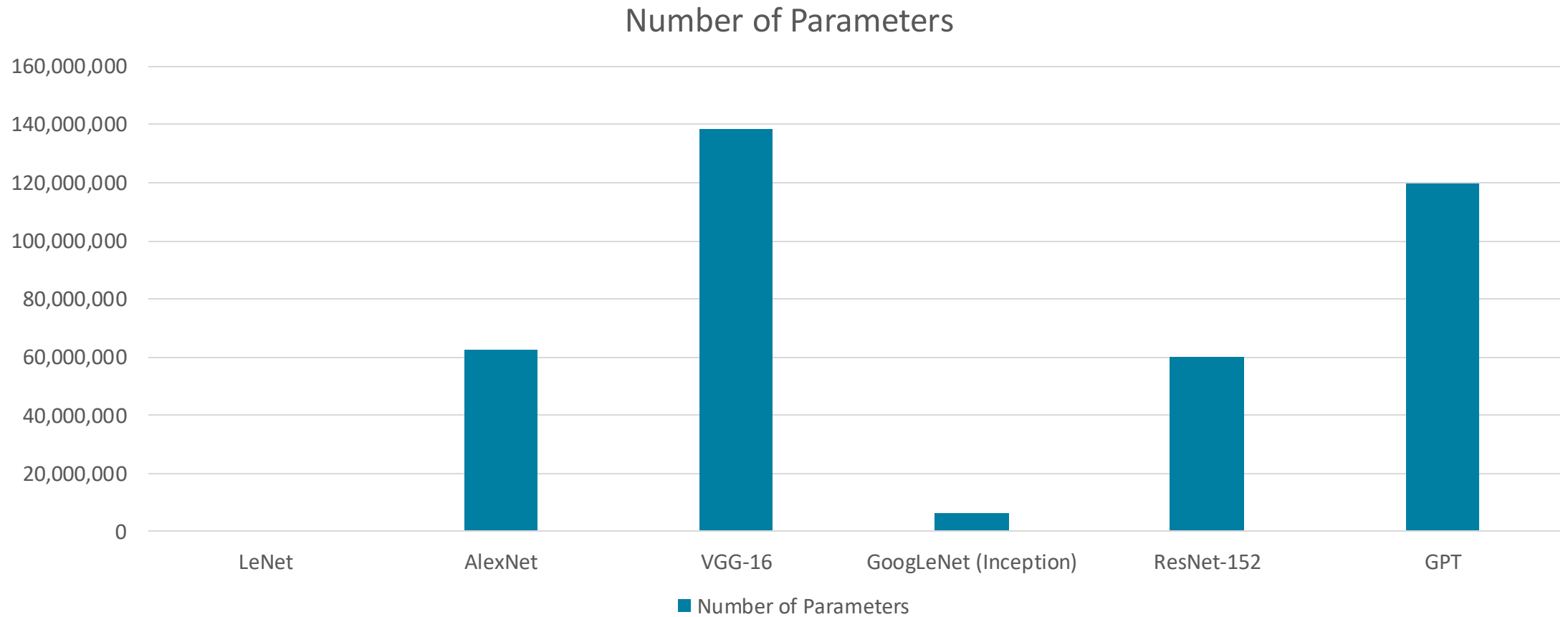
2015 = 0.07%

2019 = 0.03%

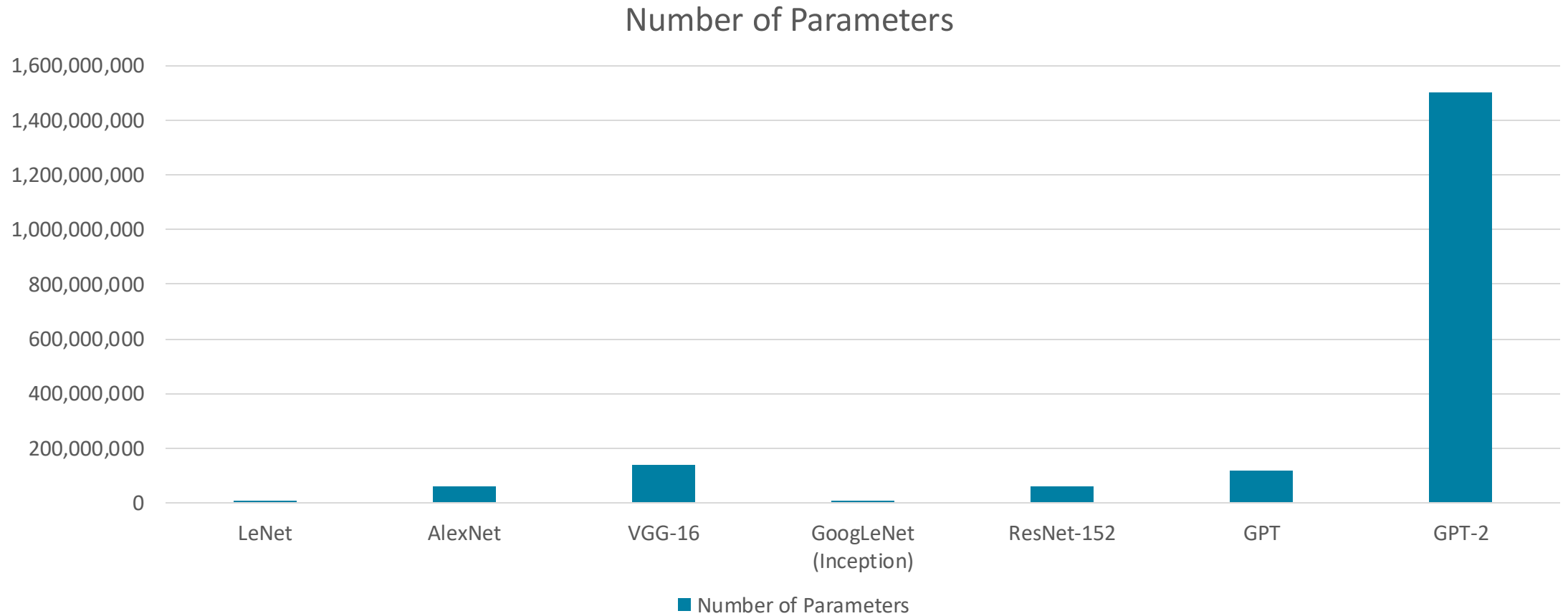
Building GPT



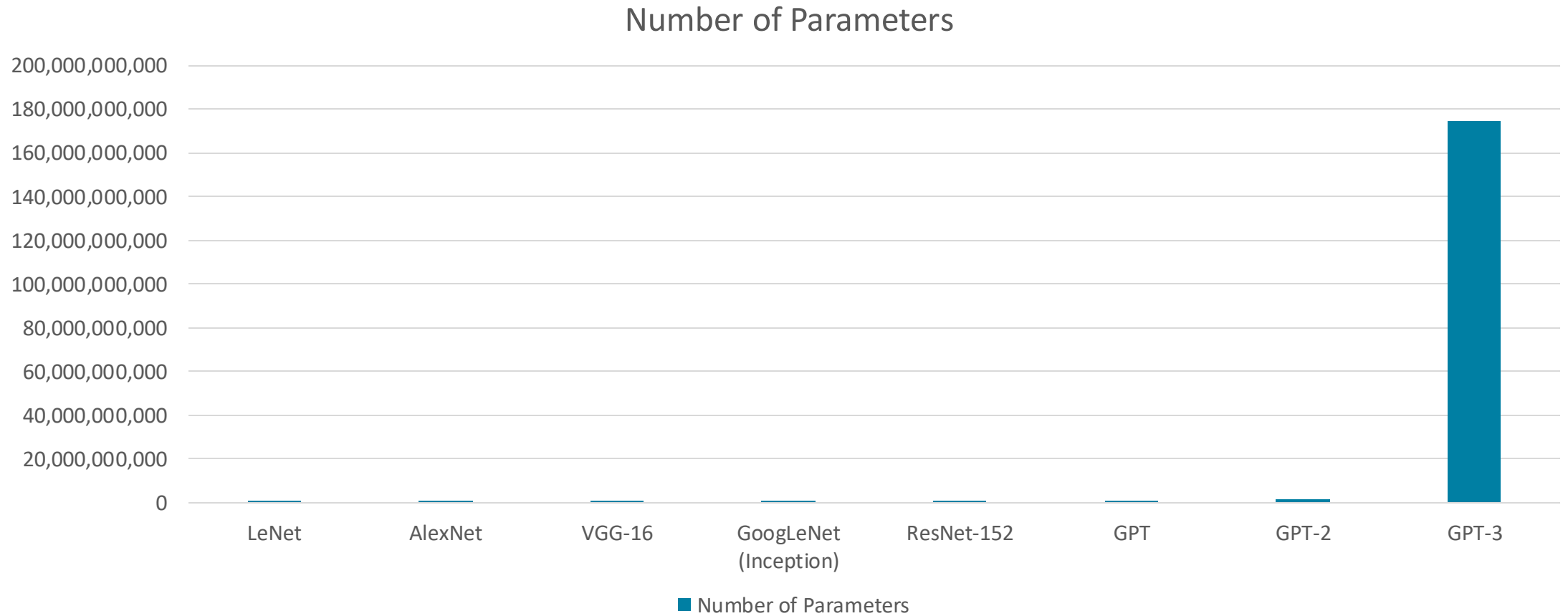
Scale of GPT



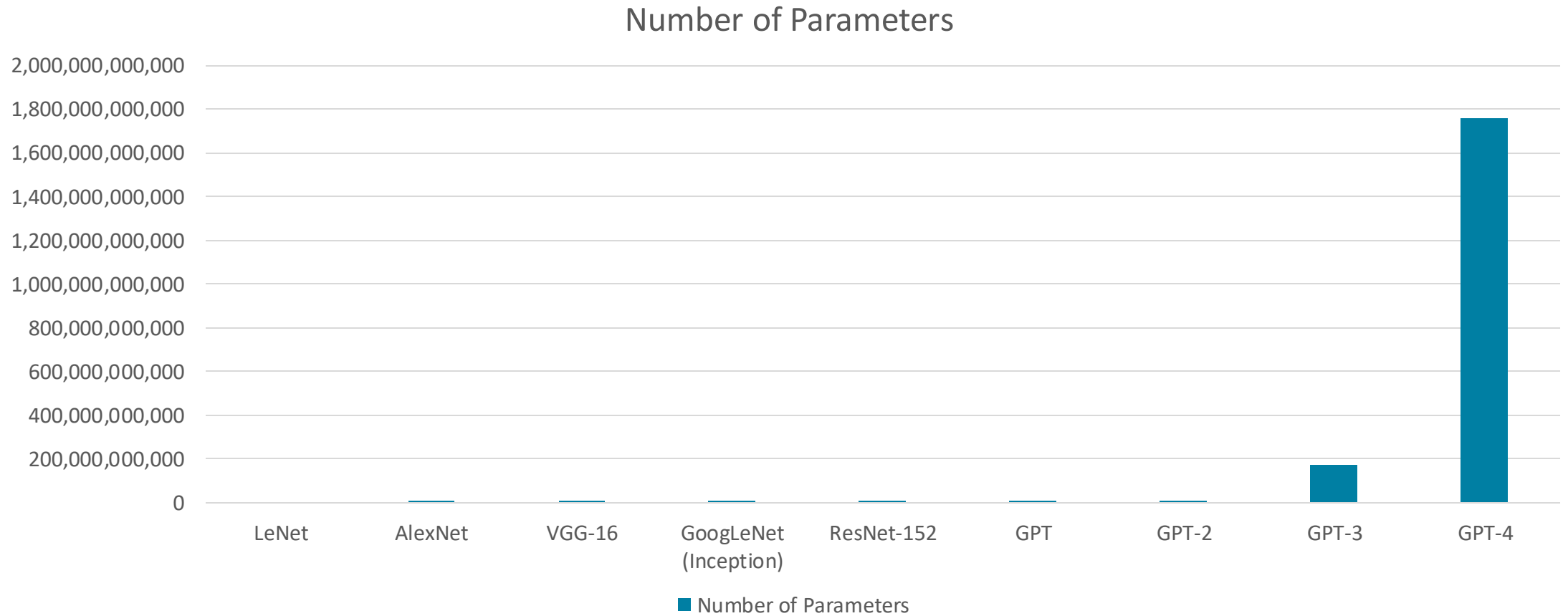
Scale of GPT



Scale of GPT



Scale of GPT



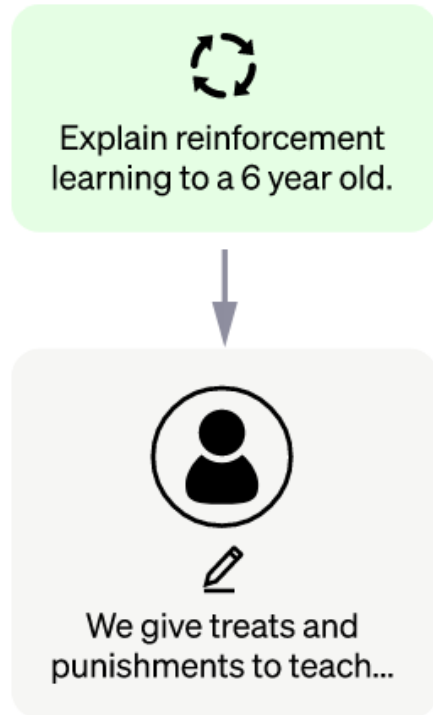
GPT's Training Data

- 1 token $\approx \frac{3}{4}$ 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
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

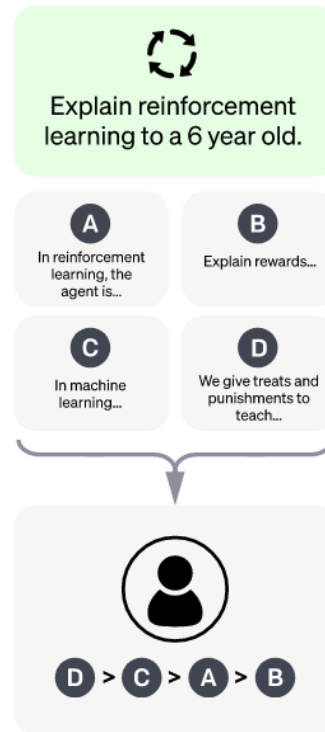
The training innovation of ChatGPT

Human annotators write answers to questions



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

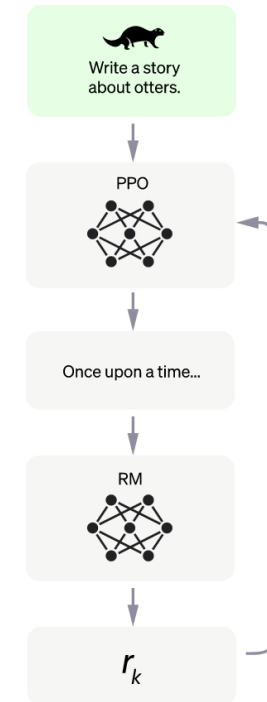
Human annotators write more answers, and someone else ranks them



A separate model learns to rate the quality of an answer

No more humans involved!

GPT writes answers to sampled questions



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