# Transformers and LLMs

Alex Olson

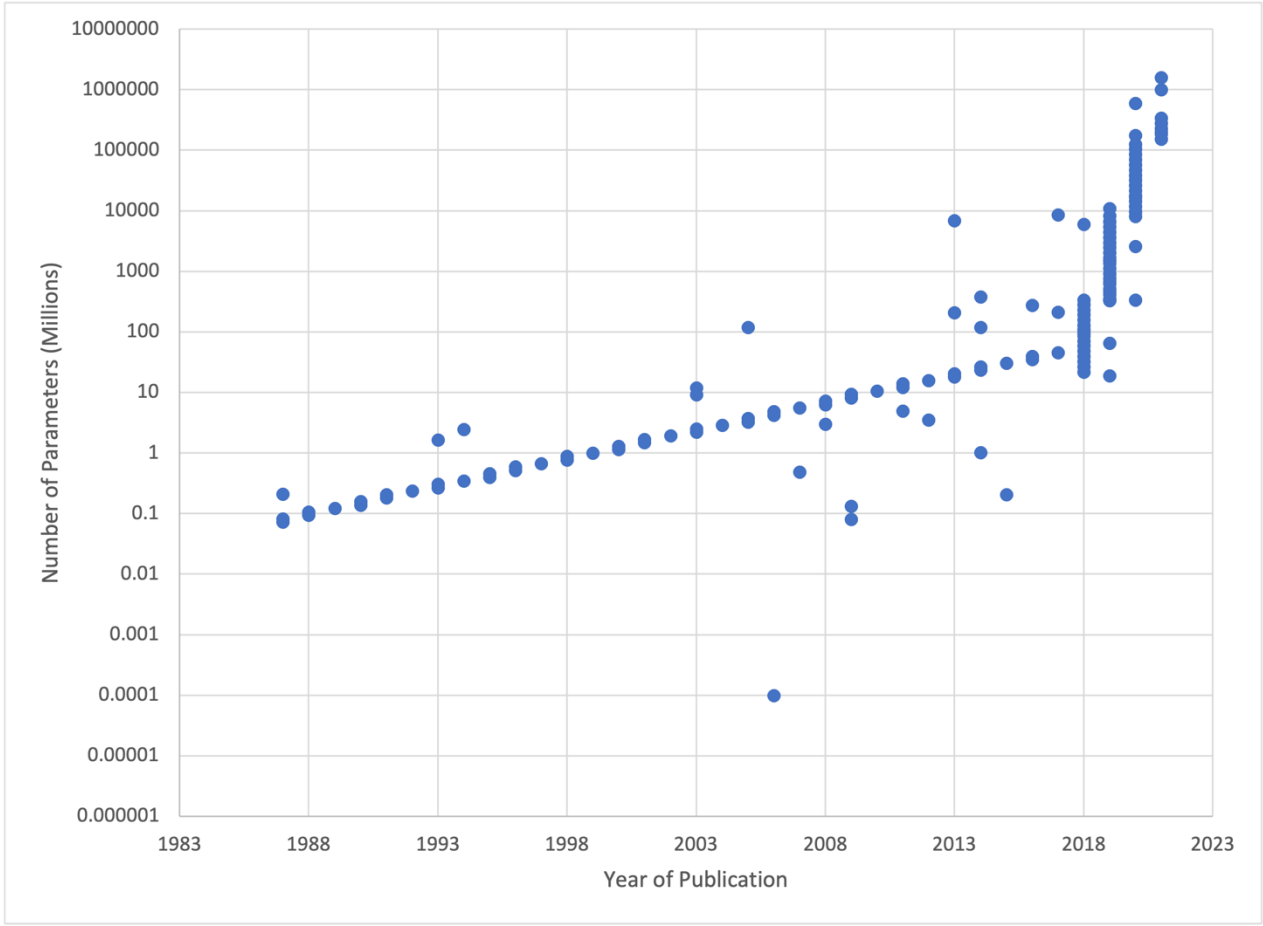
# Language Models

- Estimates the likelihood of a sequence of words occurring
- To generate text, select the word most likely to appear next
- How do we estimate likelihood? By looking at lots of text
- Simple approach: look up the number of times a sequence occurs
- More sophisticated: Neural Networks

$$P(The, dog, and, the, cat) > P(The, dog, and, the, ostrich)$$

# Large Language Models

- Latest models are capable of learning from much more data
- Both thanks to technological improvements, and a willingness to spend more money

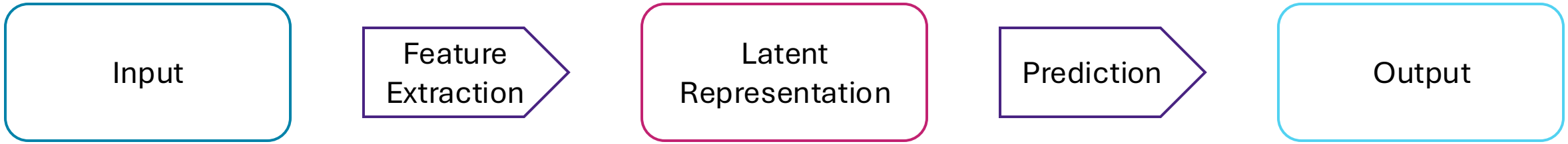


# Defining GPT

- 2018: Generative Pre-Trained Transformer
  - Key innovation in GPT was the *training*, not the model itself
- GPT-2 and GPT-3: Almost the same model, but with *(way)* more data
- GPT-4: Even larger, with an optional computer vision component
- Now: 4o, 4o-mini, 4o-turbo, o1…

# Deep Learning at a high level

- Now that we have a method to *extract features* from our symbols, we can use those representations to predict




- In many cases, feature extraction is the hard part, and prediction is comparatively easy (e.g. many vision problems)
- This is not really true for language, however

# The Transformer, and Attention

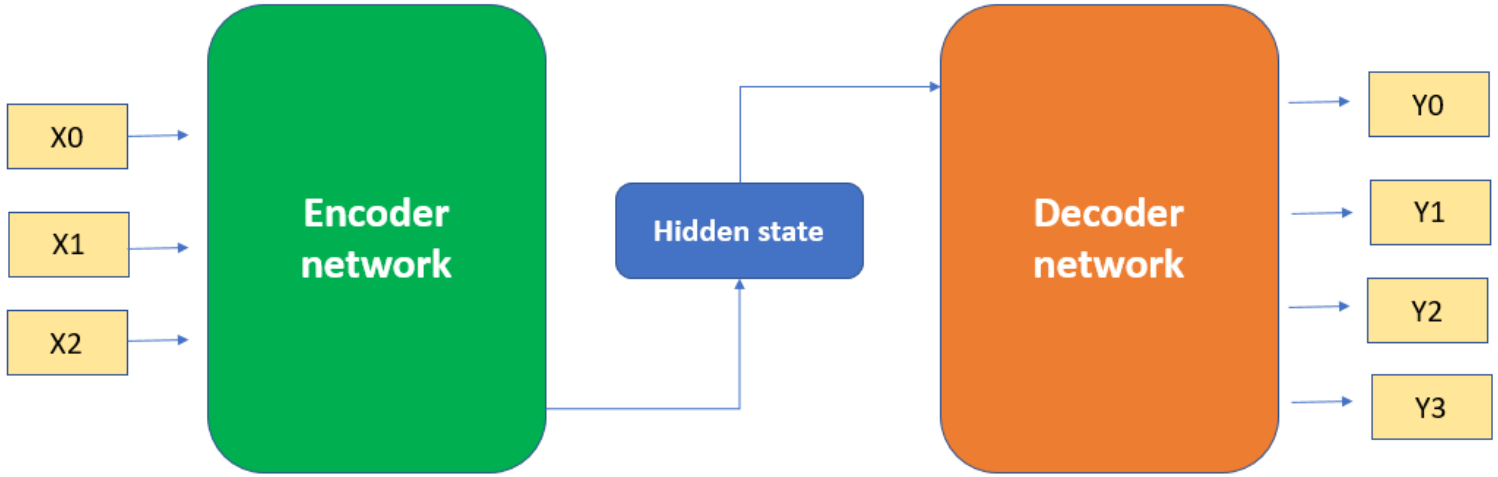
- The current revolution in large language models is driven by a key innovation in deep learning first published in 2017: the *transformer*
- Transformers introduce a new concept in deep learning called attention
- Understanding these two concepts is critical to understanding why these models work so much better today

# Encoder-Decoder Networks

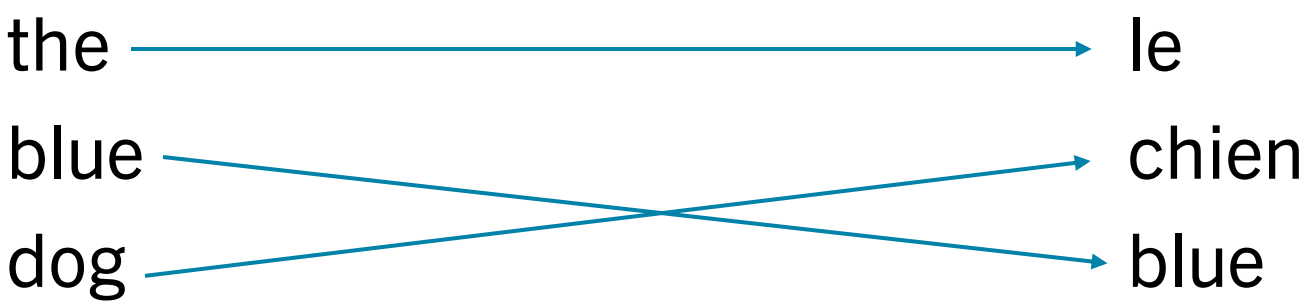
“the blue dog”


“le chien bleu”

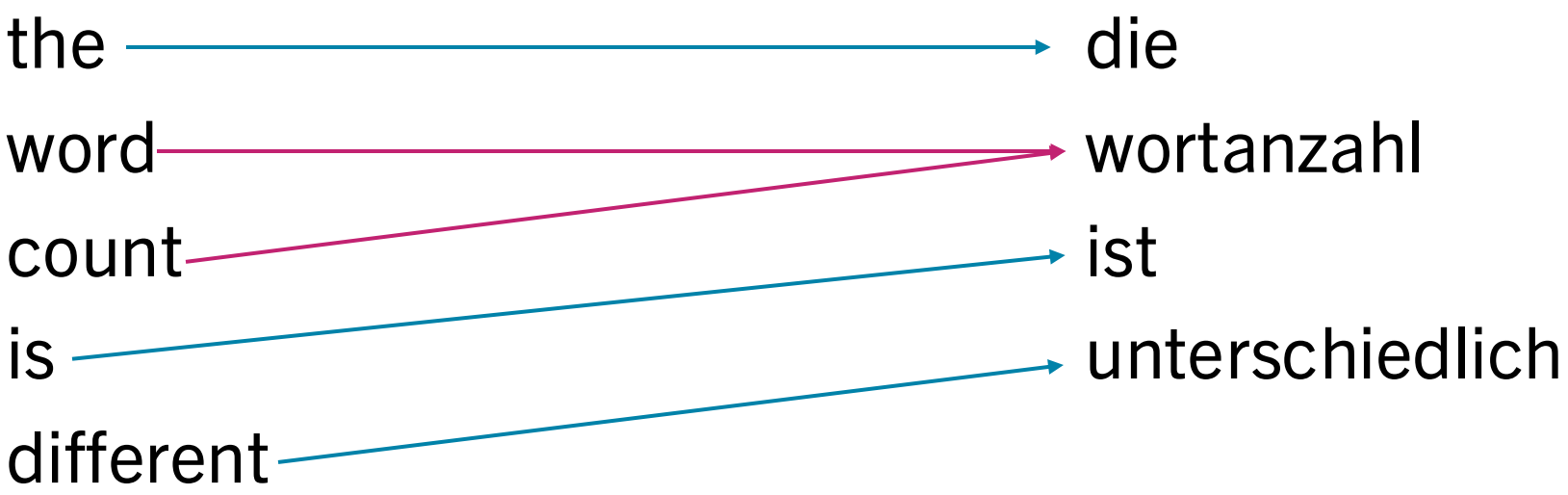
# Encoder-Decoder Networks



# Attention



# Attention


the
word
count
is
different
die
wortanzahl
ist
unterschiedlich

# Attention

- Attention mechanism predicts how much each word *depends* on the words in the input
- By understanding this relationship, prediction power for text is greatly improved

| | le | chien | blue |
|---|---|---|---|
| the | 1 | | |
| blue | | 0.2 | 1 |
| dog | | 1 | 0.2 |

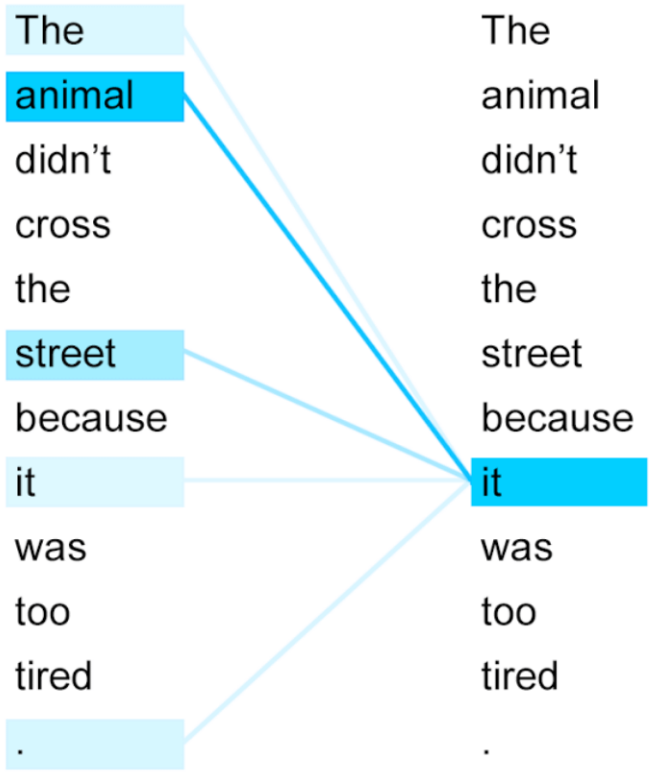
# Attention

- This dependency matrix can then be multiplied against the word embeddings to create new, contextual word embeddings

| | wortanzahl | variiert |
|---|---|---|
| word | 0.5 | |
| count | 0.5 | |
| differs | | 1 |

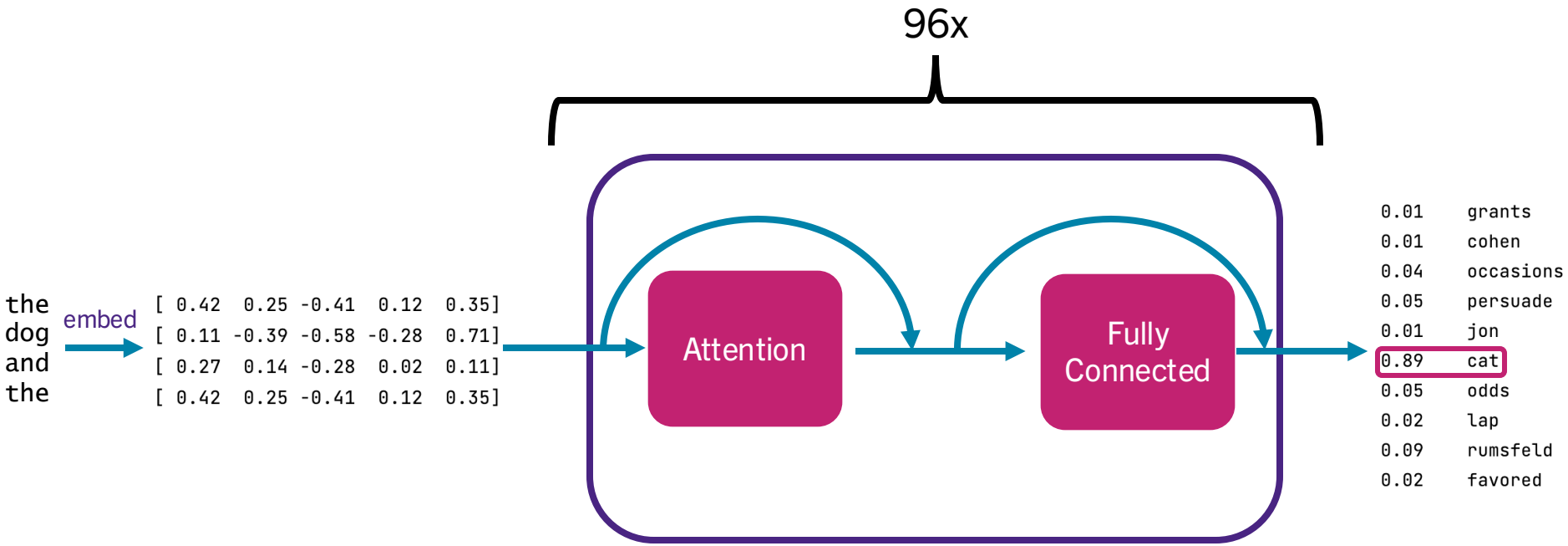
# Self-Attention

- In models like ChatGPT, we use *self-attention* — simply put, the relationship is now between the phrase and itself

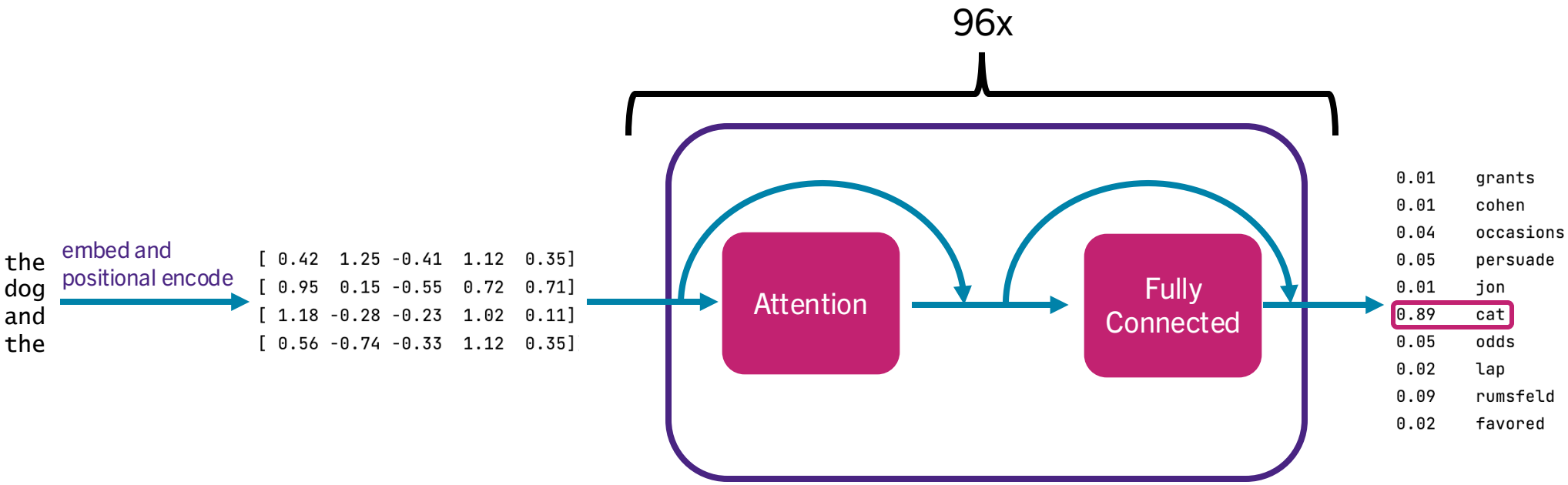


# Building GPT

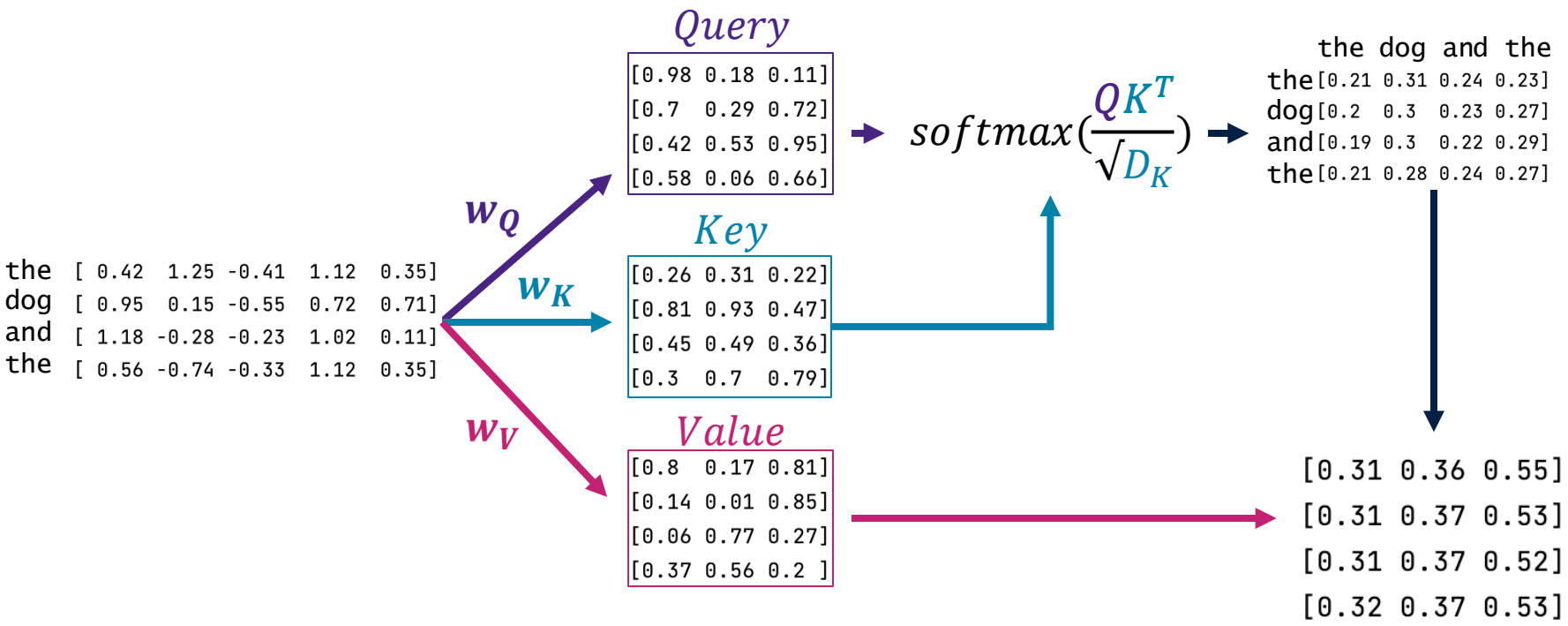
# Building GPT: The Transfomer



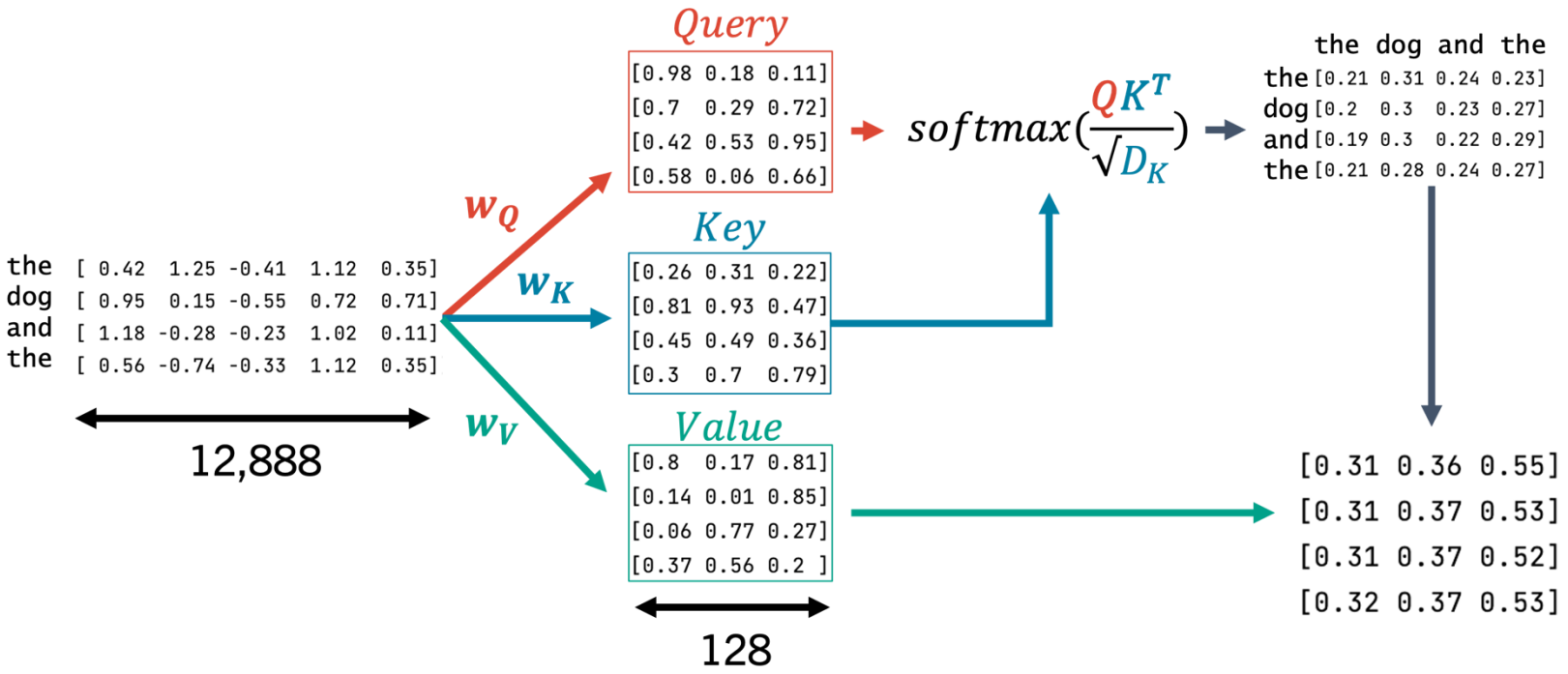
# Building GPT



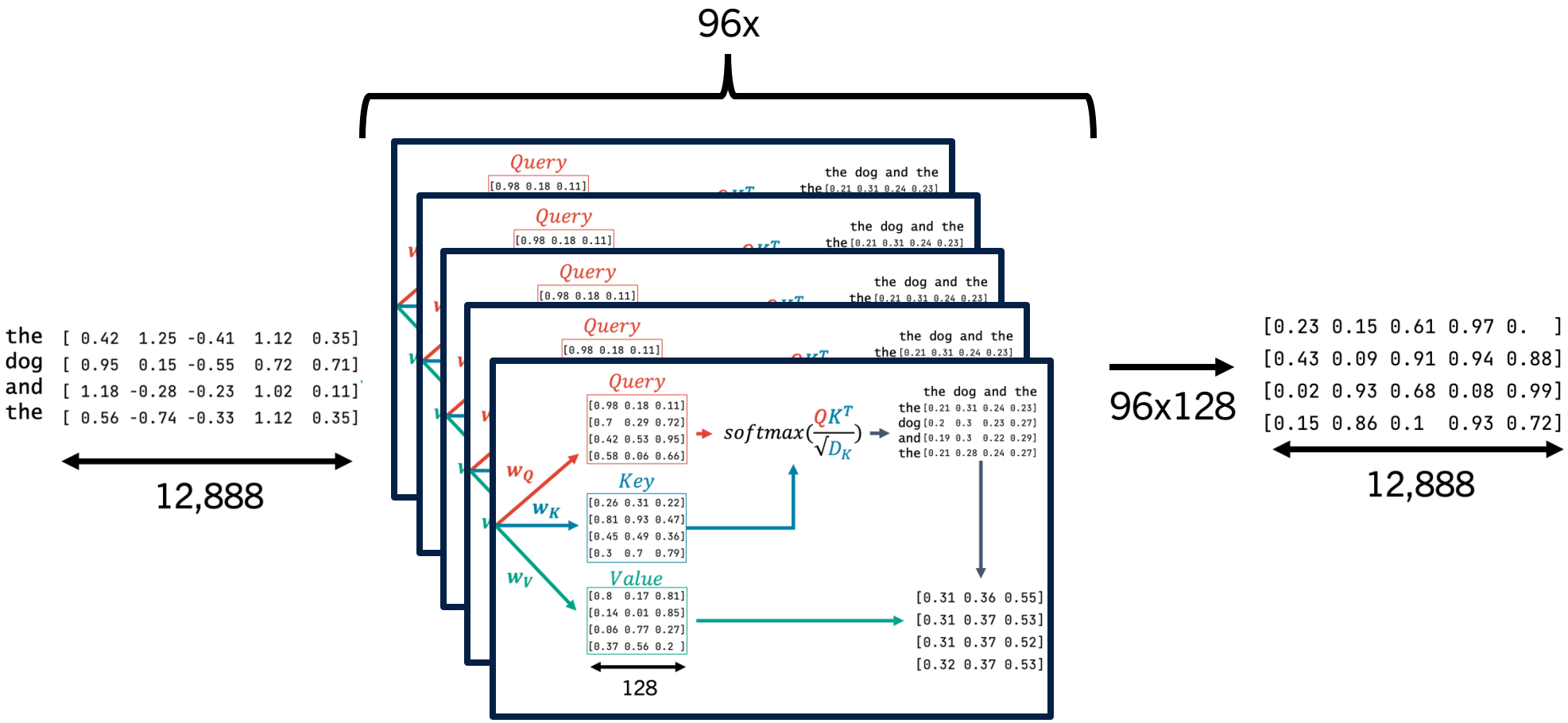
# Building GPT: Attention



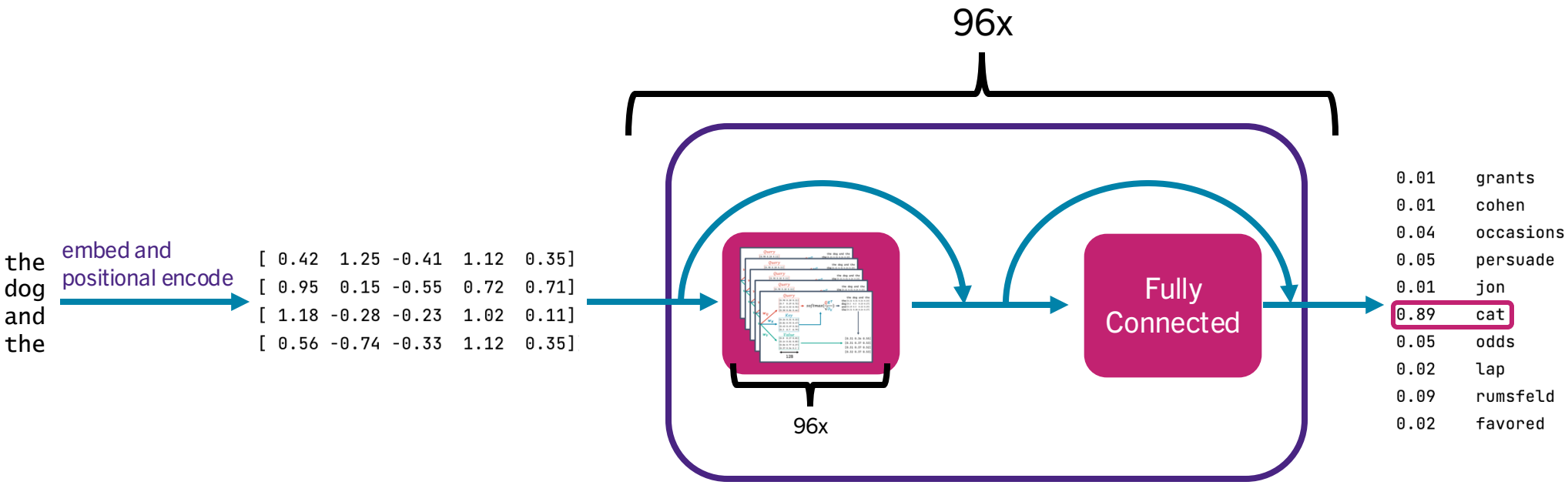
# Building GPT: Attention



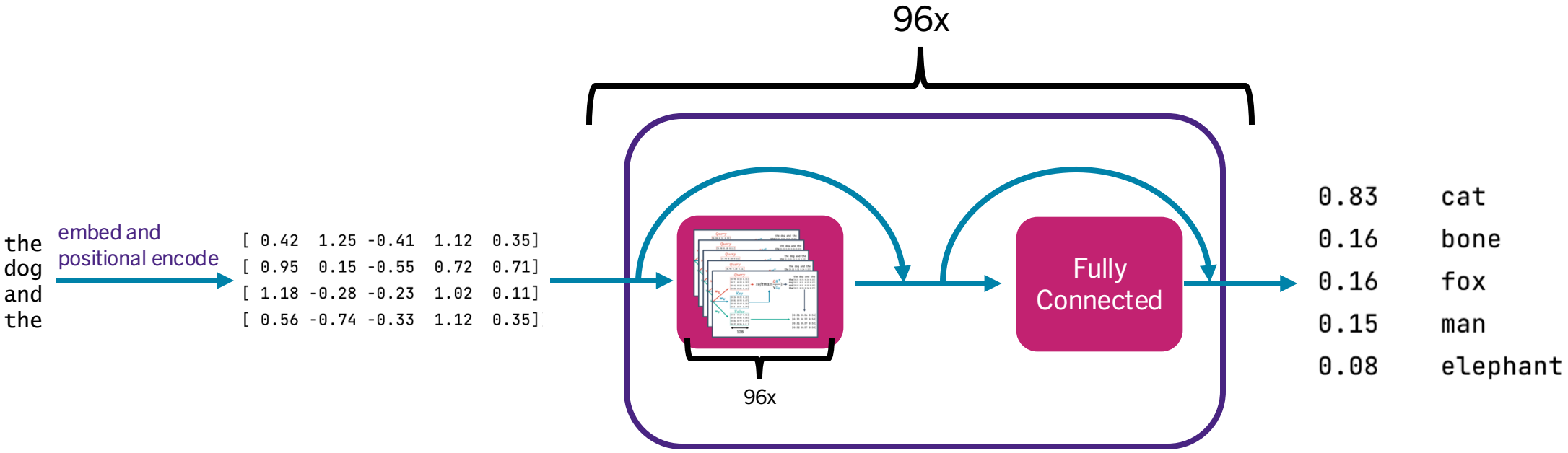
# Building GPT: Attention



# Building GPT



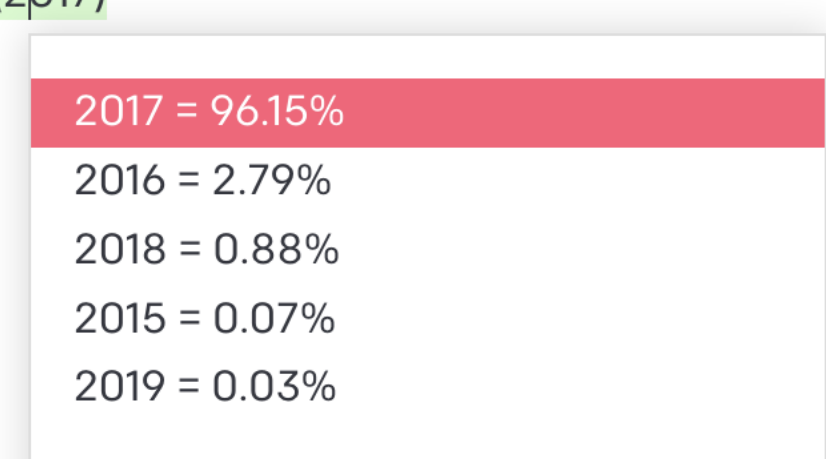
# Building GPT: Top-P



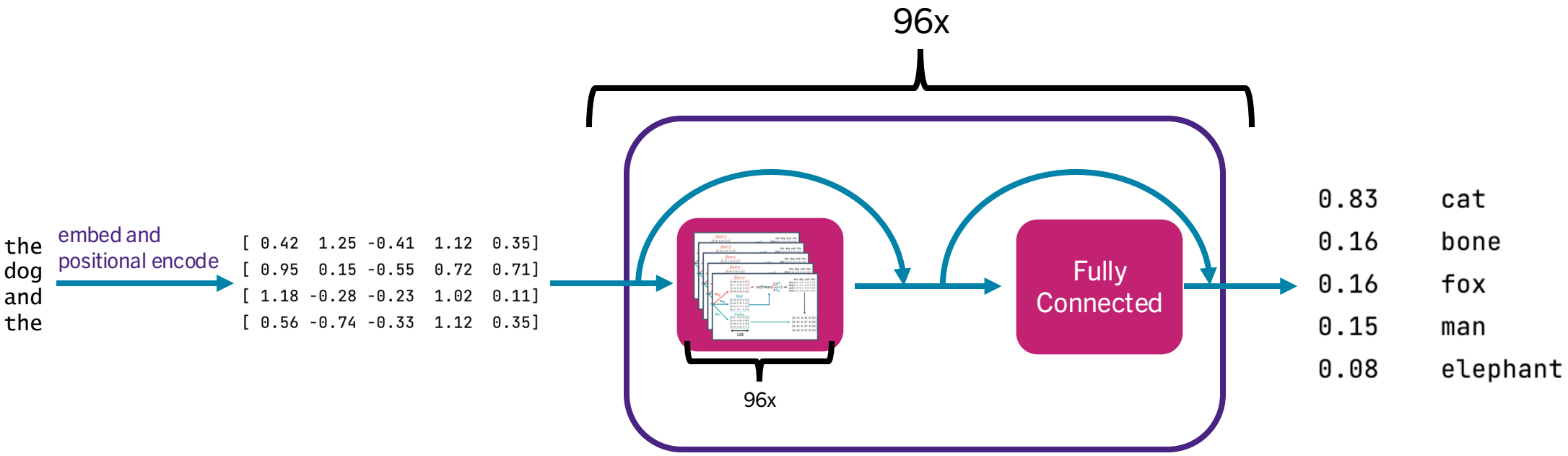
# Building GPT: Top-P

Top 10 documentaries about artificial intelligence:

1. AlphaGo (2017)



# Building GPT



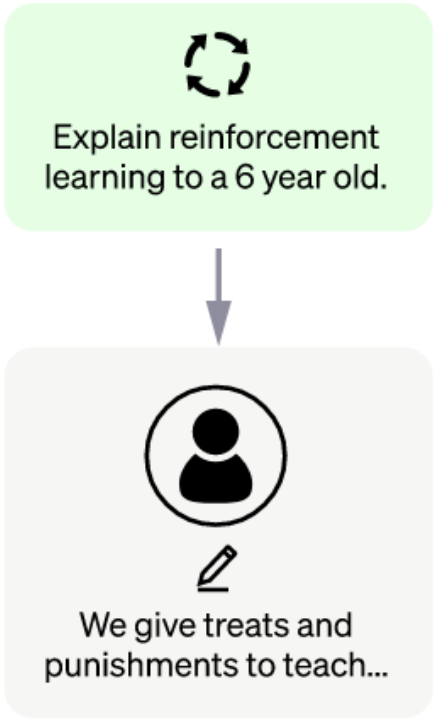
# GPT's Training Data

- 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 |
|---|---|---|

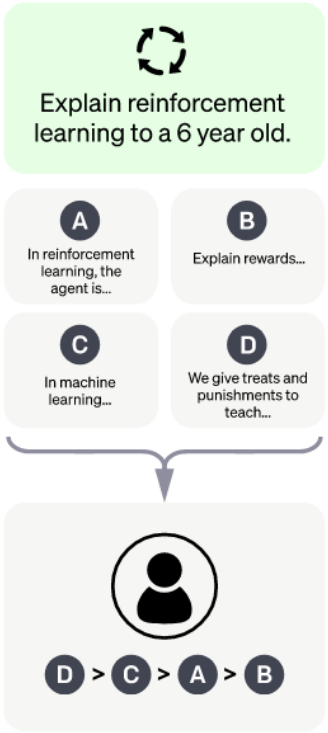
# The training innovation of ChatGPT

Human annotators write answers to questions



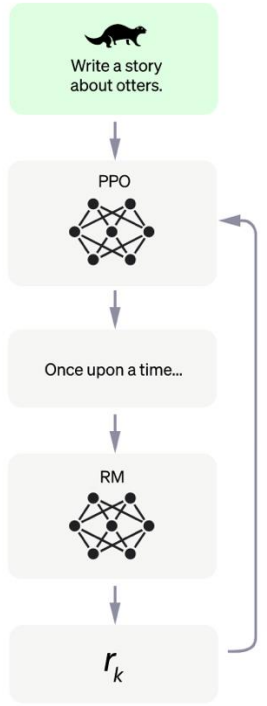

The generalist GPT model is taught from these Q&A pairs

Human annotators write <u>more</u> answers, and someone else ranks them




A <u>separate</u> model learns to rate the quality of an answer

GPT writes answers to sampled questions



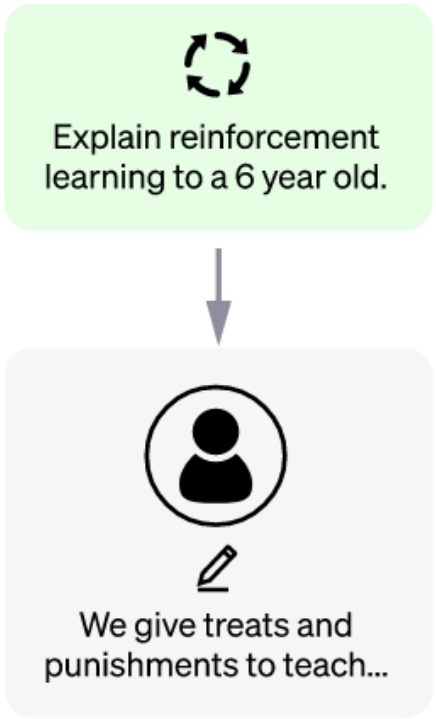

The reward model rates each answer, allowing GPT to keep learning

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

# The training innovation of ChatGPT

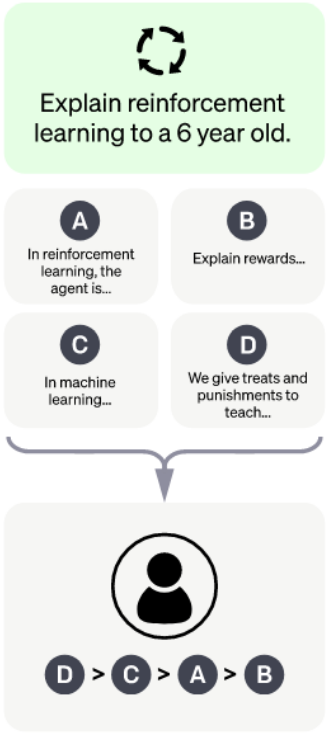
No more humans involved!

Human annotators write answers to questions



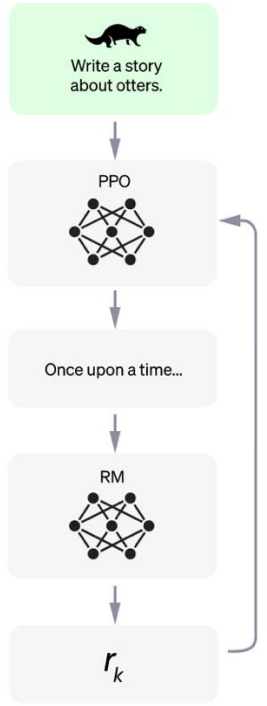

The generalist GPT model is taught from these Q&A pairs

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/