Northwestern Kellogg

Northwestern ENGINEERING

MBAi 448 I Winter 2026

Transformers

Alex Castrounis

www.whyofai.com | linkedin.com/in/alexcastrounis

Attendance

Enter the “Magic Word” in Canvas

Do Not Share with Absent Students

Image via ChatGPT

K CO

Positive Impact Investing to Benefit Humanity and Earth

What is natural language?

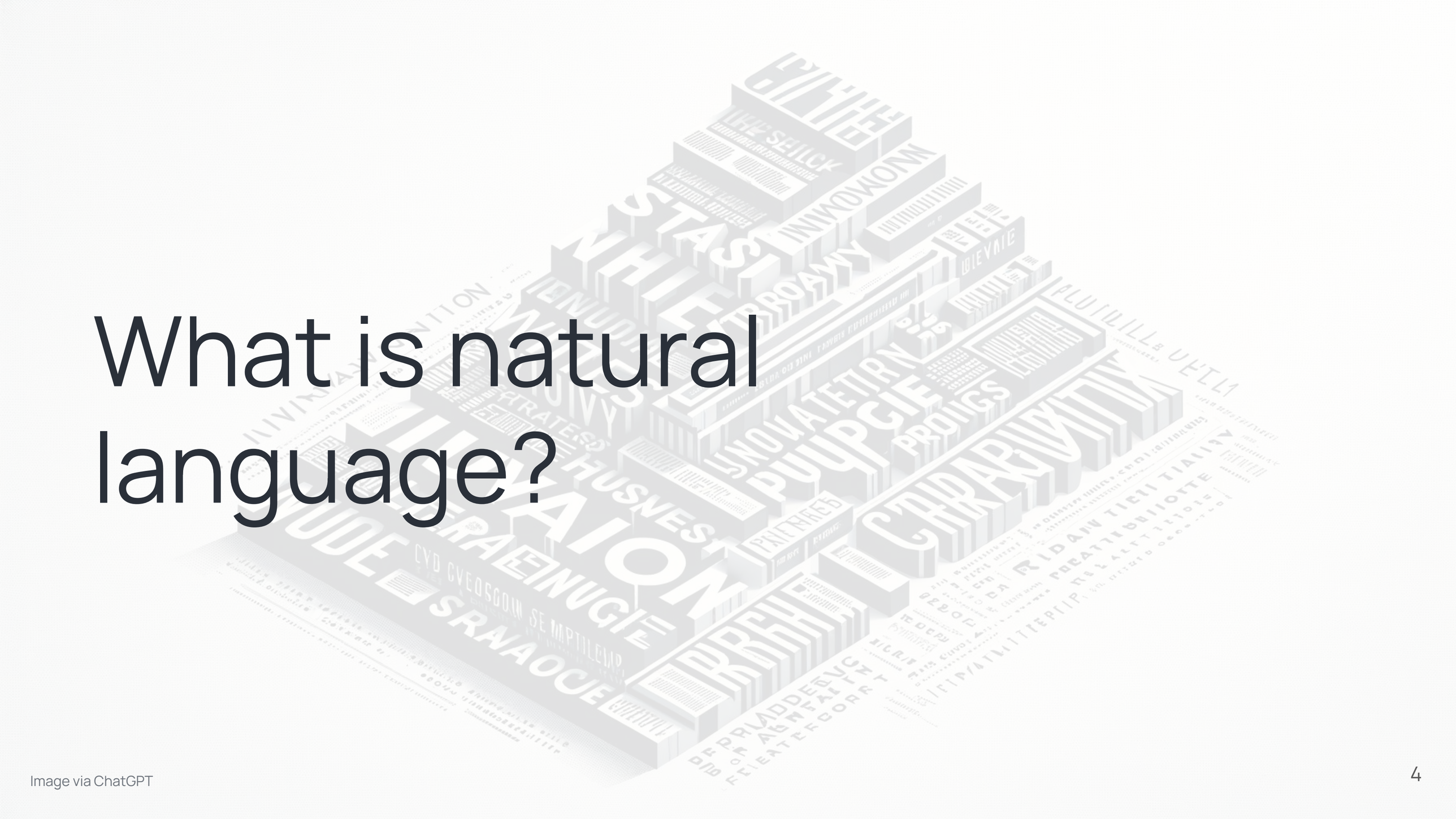
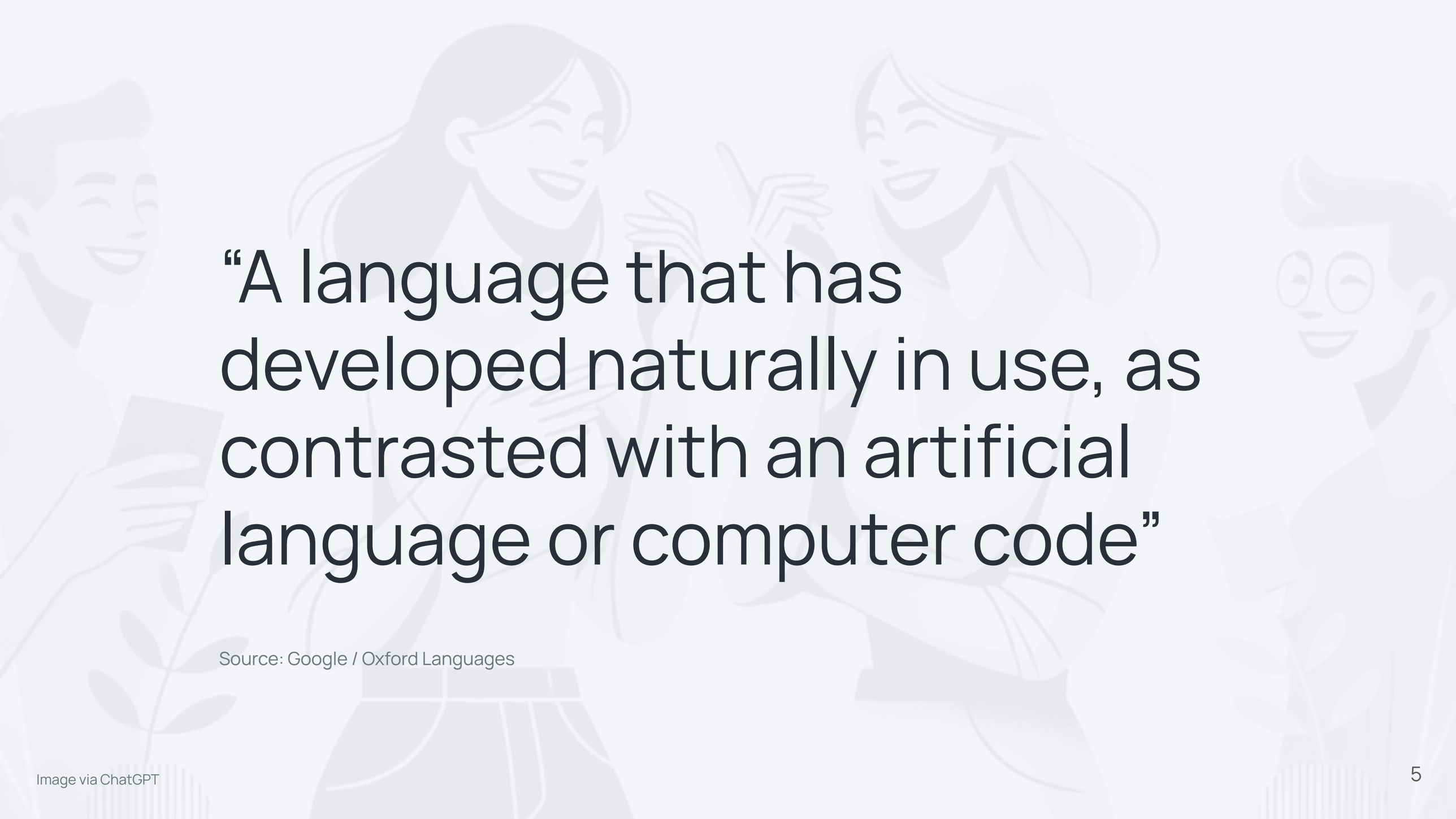


Image via ChatGPT



“A language that has developed naturally in use, as contrasted with an artificial language or computer code”

Source: Google / Oxford Languages

Image via ChatGPT

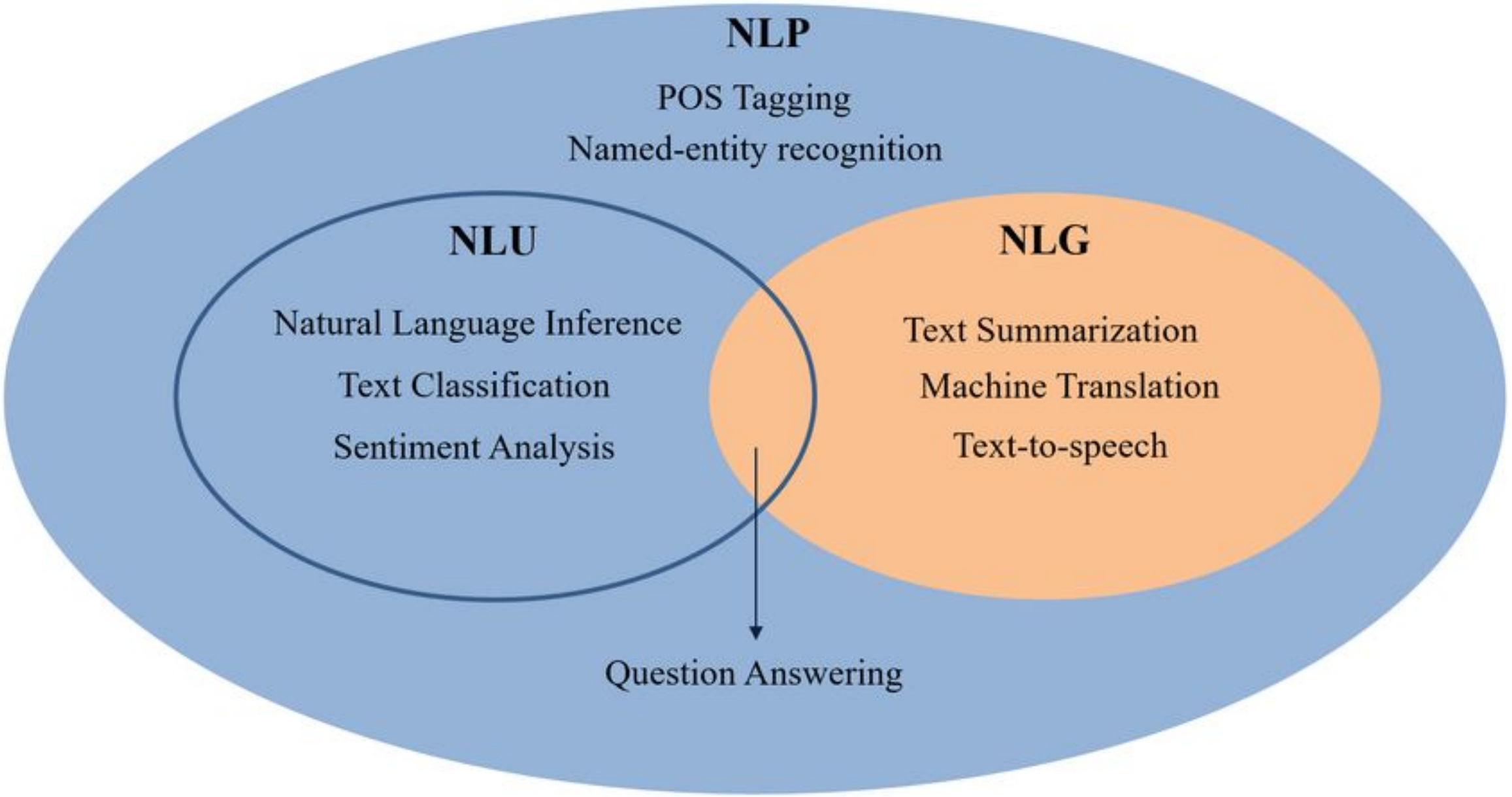


Image via Archana Tikayat Ray

“NLP is a field of linguistics and machine learning focused on understanding everything related to human language. **The aim of NLP tasks is not only to understand single words individually, but to be able to understand the context of those words.**”

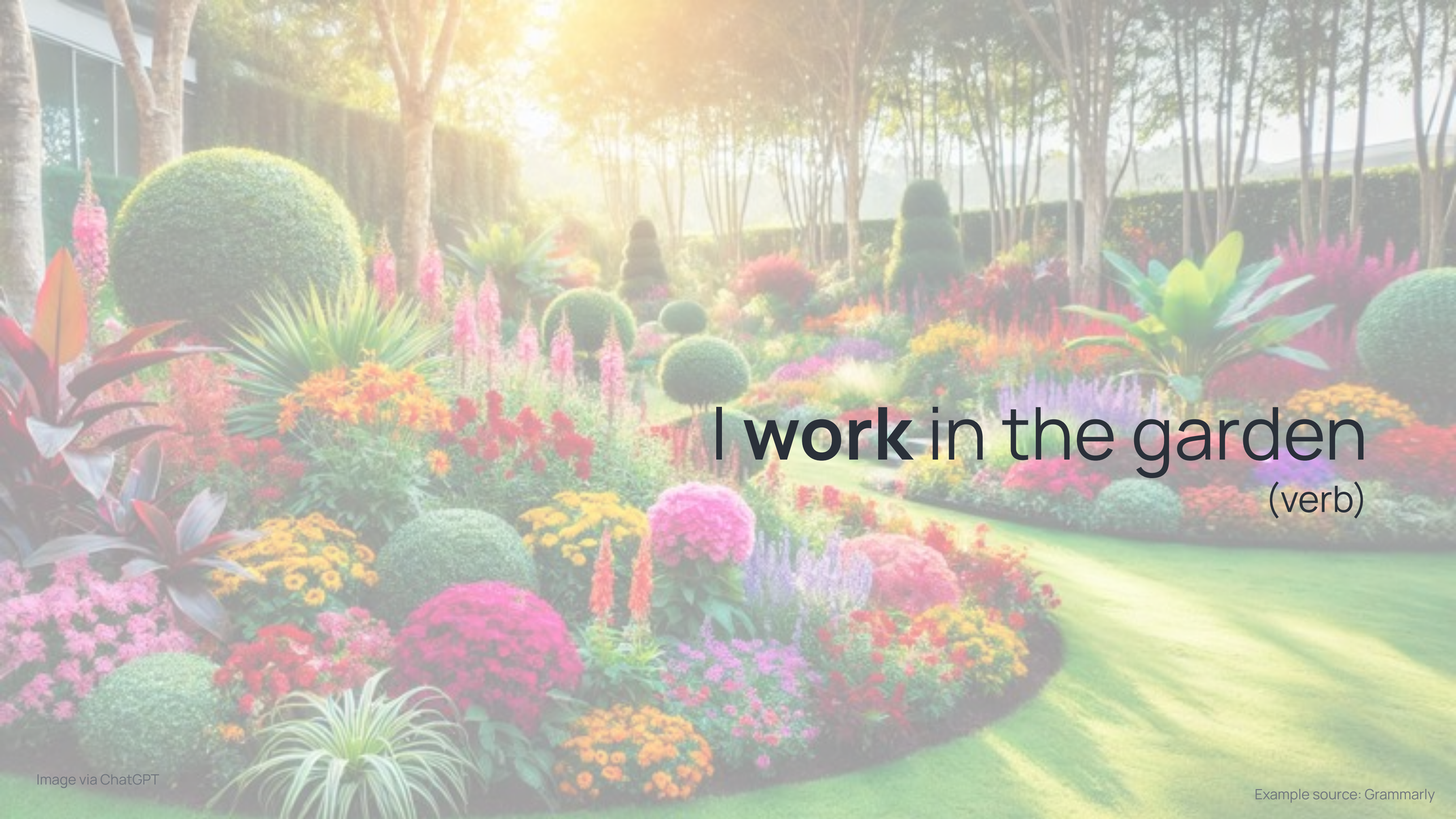
Source: Hugging Face

HUGGING FACE

Natural Language is Hard

Challenges Faced by Language Models

- Word interdependencies and long-range dependencies
- Semantic and contextual relationships and dependencies
- Coreference resolution (aka entity resolution)
- Word ambiguity and general ambiguity (e.g., context, intent, entities)
- Understanding parts of speech (POS)
- Understanding semantic roles (aka thematic relations)
- Parsing language (e.g., tokens, words)
- Synonyms
- Respecting gender and plurality (especially in different languages)
- Language-specific word “genders” and orders
- Word, phrase, and sentence variations (e.g., slang)
  - CV suffers from similar problems with image variations
- Understanding grammar rules (e.g., subject-verb agreement, possessive noun)



I **work** in the garden

(verb)

Image via ChatGPT

Example source: Grammarly

I went to **work**

(noun)

Image via ChatGPT

Example source: Grammarly

I arrived at the **bank** after crossing the river

Example source: Google

Image via ChatGPT

I arrived at the **bank** after crossing the river

Example source: Google

Image via ChatGPT

Server, can I have the check?

Example source: Dale Markowitz

Image via ChatGPT

Looks like I just crashed the **server**

Example source: Dale Markowitz

Image via ChatGPT

Image via ChatGPT

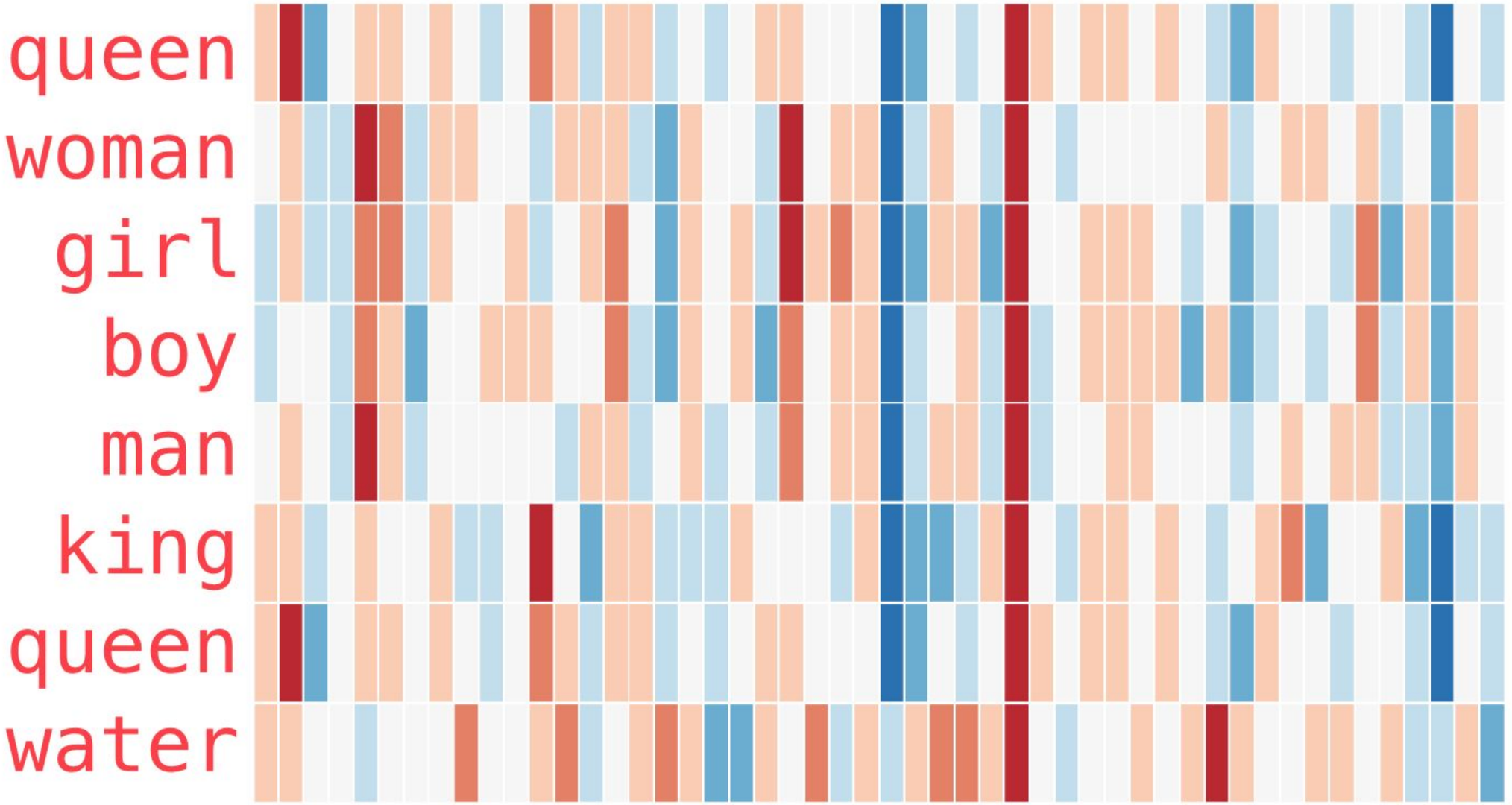


Image: The Illustrated Word2vec

What about context?

Image via ChatGPT

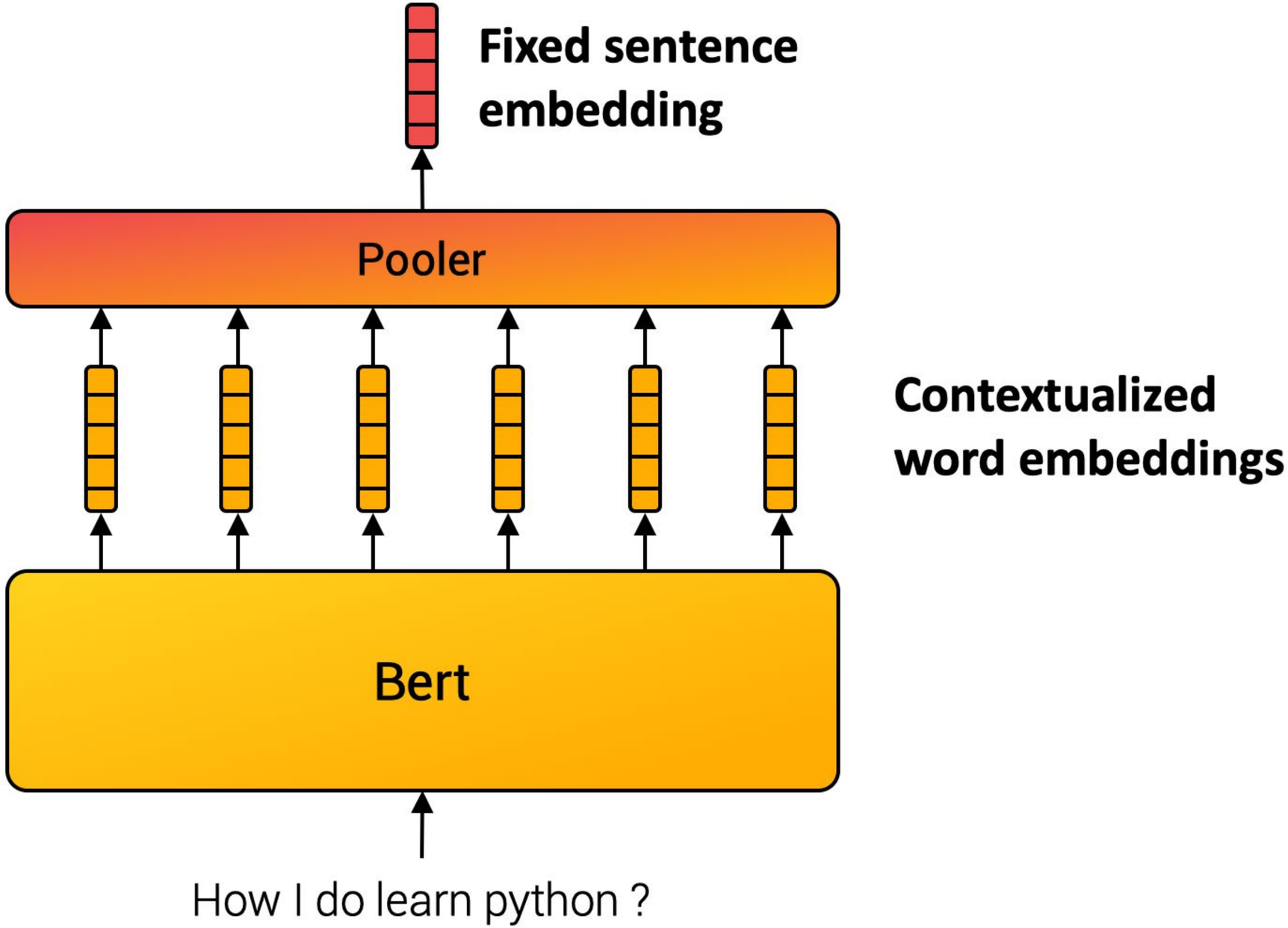


Image credit: Hugging Face

We’ll Circle Back to This Soon

Image via ChatGPT

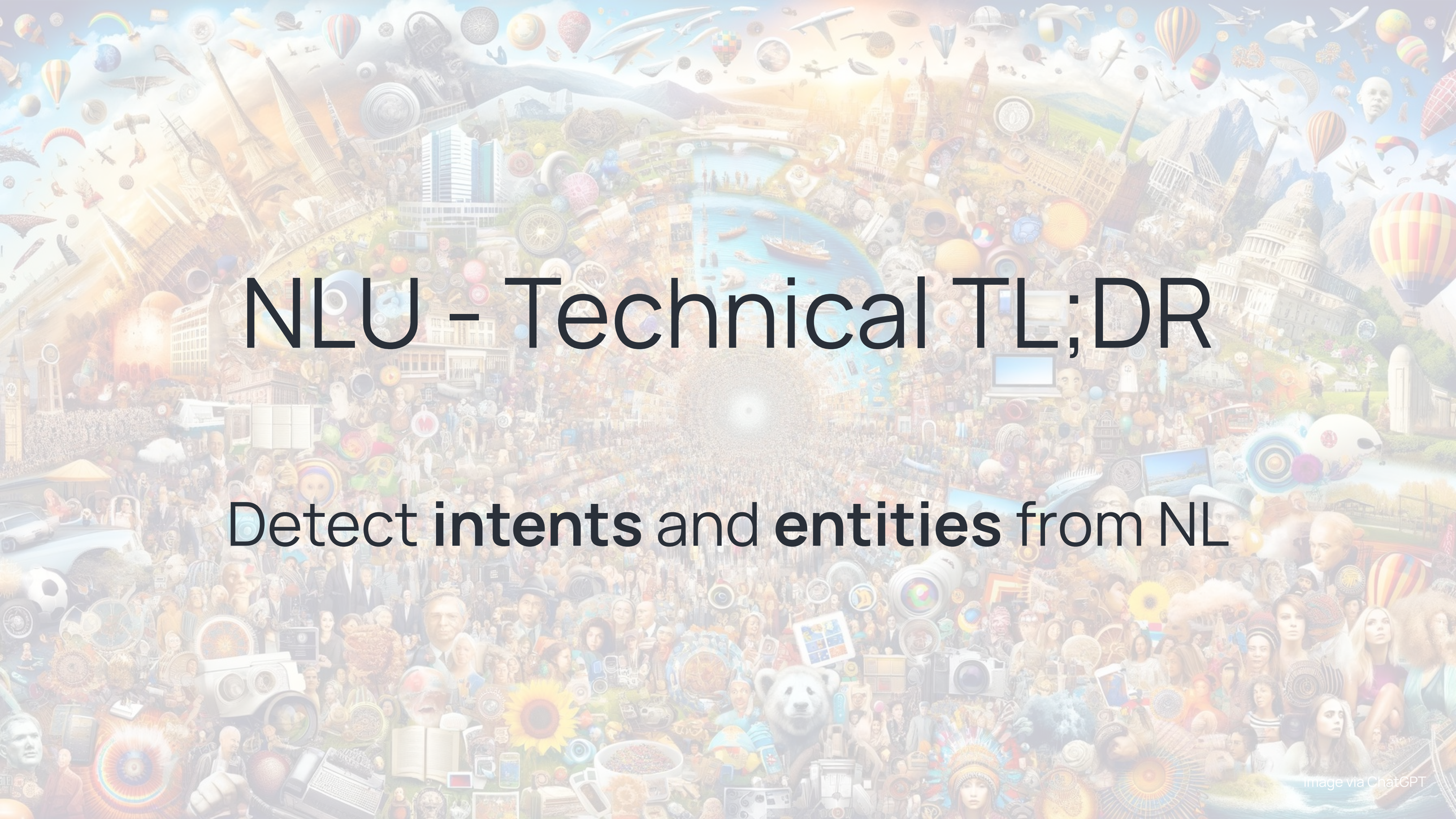
What is natural language understanding?

“Natural-language understanding (NLU) is the comprehension by computers of the structure and meaning of human language (e.g., English, Spanish, Japanese), allowing users to interact with the computer using natural sentences.”

Source: Gartner

NLU - Technical TL;DR

Detect **intents** and **entities** from NL



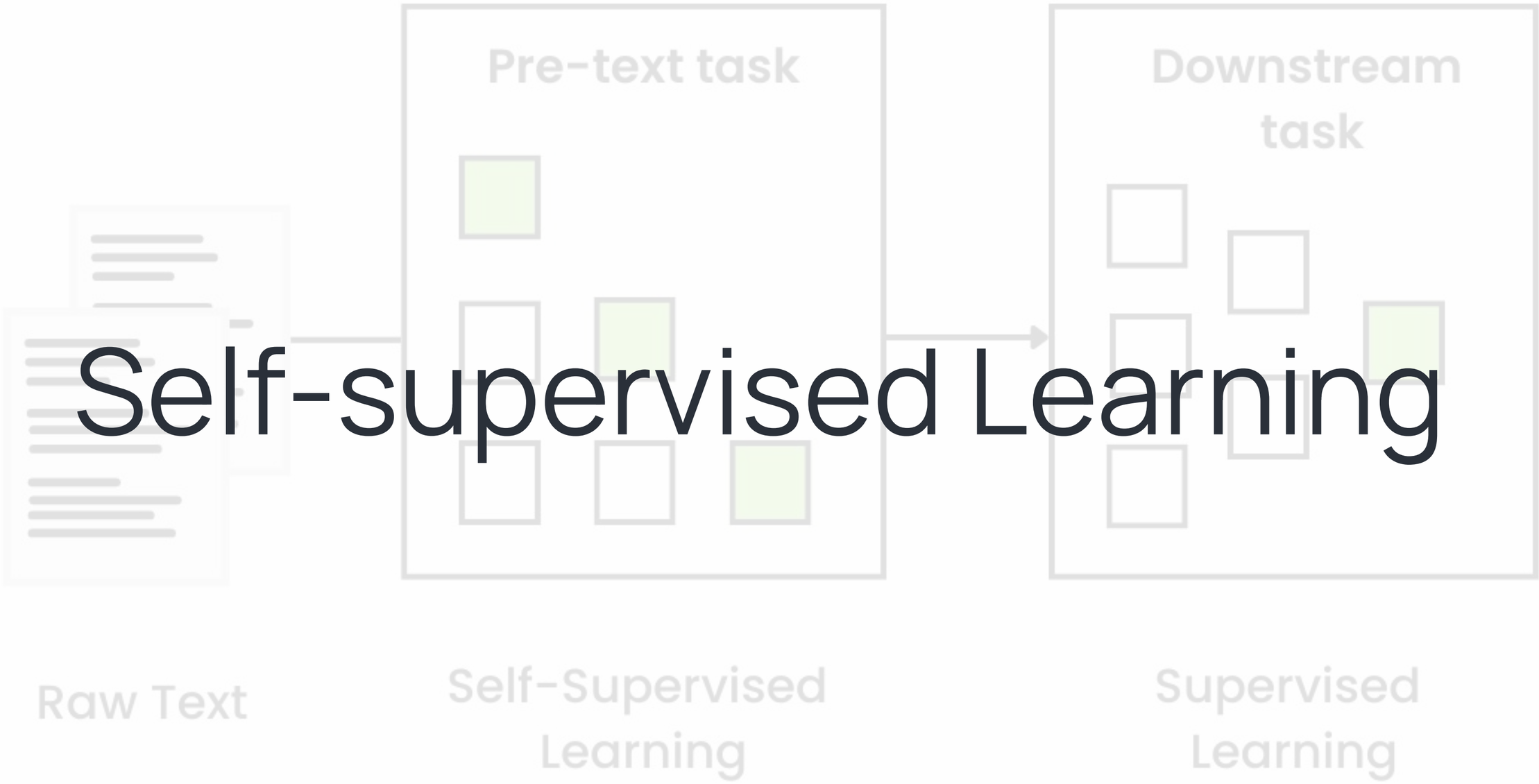
Example Intents and Entities

Intents

- Get a question answered
- Get information
- Get help or support
- Place an order
- Book or schedule something
- Get directions
- Check something (e.g., balance)
- Calculate something
- Send something (e.g., an email)

Entities

- Dates
- Numbers (quantity, price, acct nums)
- Names (places, people, products, orgs)
- URLs
- Addresses (emails, physical)
- Phone numbers
- “Things”



Self-supervised Learning

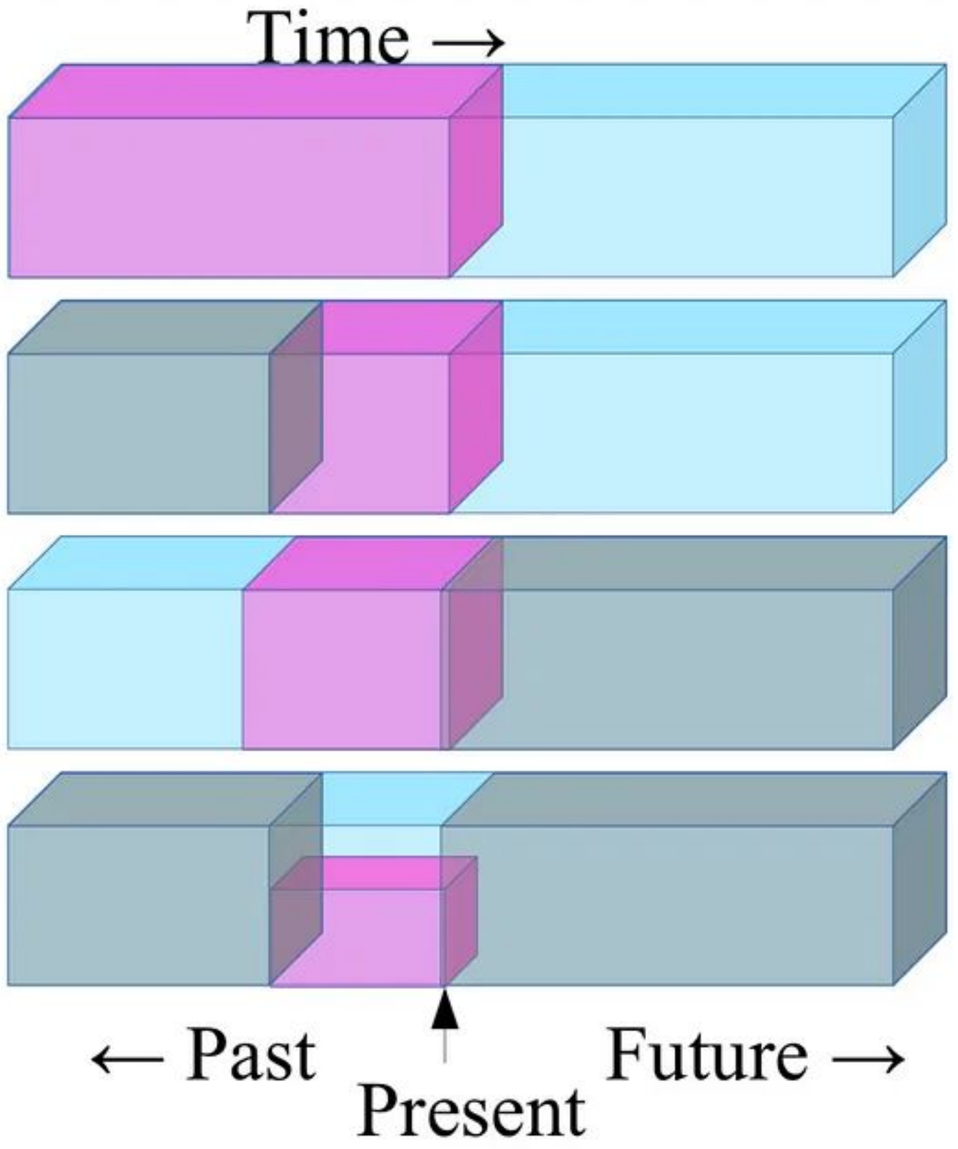
Image via Turing

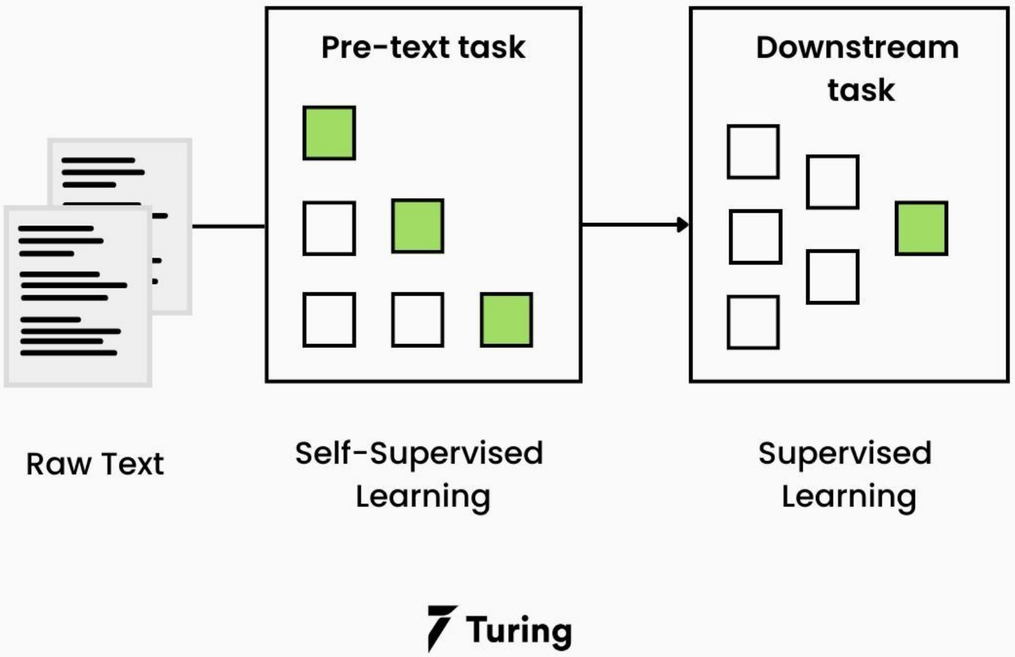
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Image credit: GPT-3 Original Paper | Language Models are Few-Shot Learners

Self-Supervised Learning

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.





For Text

For Images

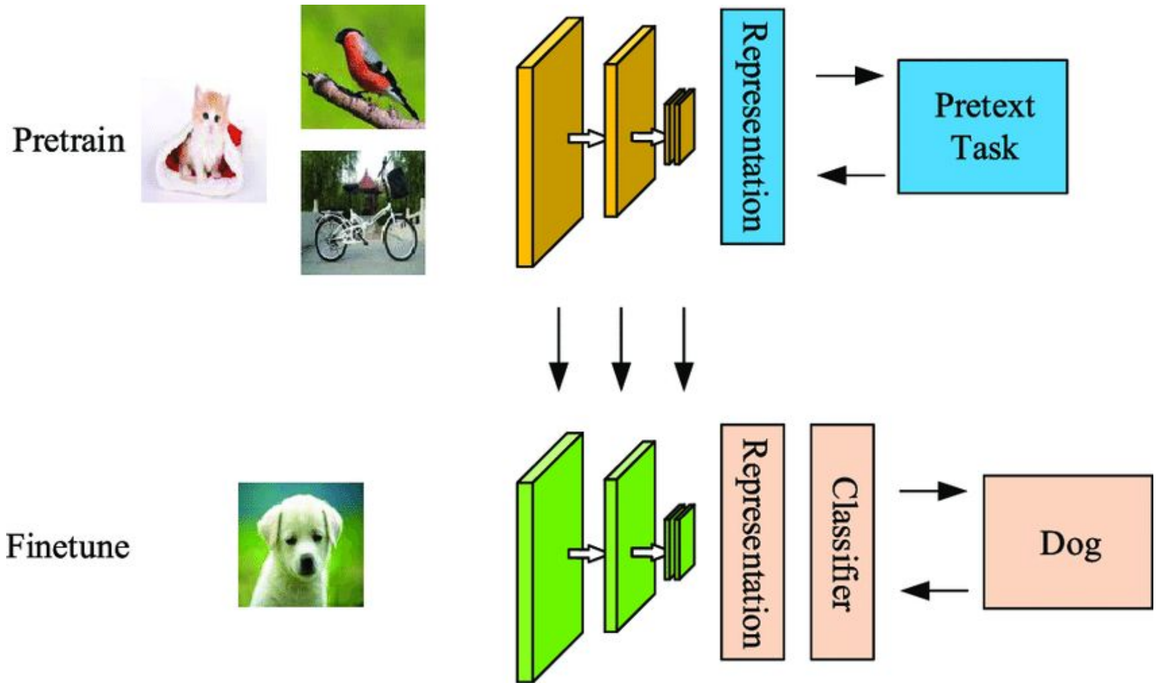


Image credit: V7 Labs

Activity

Design the Pretraining Objective

Use Case: Customer support AI learns from (very few labels):

- Chat logs (text)
- Screenshots (images)
- Voice notes (audio)
- Screen recordings (video)

Goal: Design two **self-supervised pretraining objectives** where the **target comes from the data itself**.

***** Don’t use any devices for assistance**
***** Follow guidance on activity steps**

Activity

Self-supervised pretraining loop:

X (raw) → $\tilde{X}$ (mask/corrupt/shuffle/pair) → predict → compare to X (or true pair)

Design 2 objectives (each):

- Modality: text / image / audio / video (or 2-modal)
- Change → Target: what you alter → what to predict
- Learns: capability gained

Options:

- Mask → Predict
- Corrupt → Reconstruct
- Shuffle → Fix order
- Match → Correct pair

***** Don't use any devices for assistance**
***** Follow guidance on activity steps**

Activity

Discussion:

- What exactly was the input and what was the target?
- Where did the “label” come from in your setup?
- What capability would this pretraining build (meaning, layout, sound patterns, sequence/motion)?
- Which objective seems most business-useful for support triage—and why?
- What could go wrong if the raw data is messy or biased?

***** Don’t use any devices for assistance**
***** Follow guidance on activity steps**

COMMON CRAWL

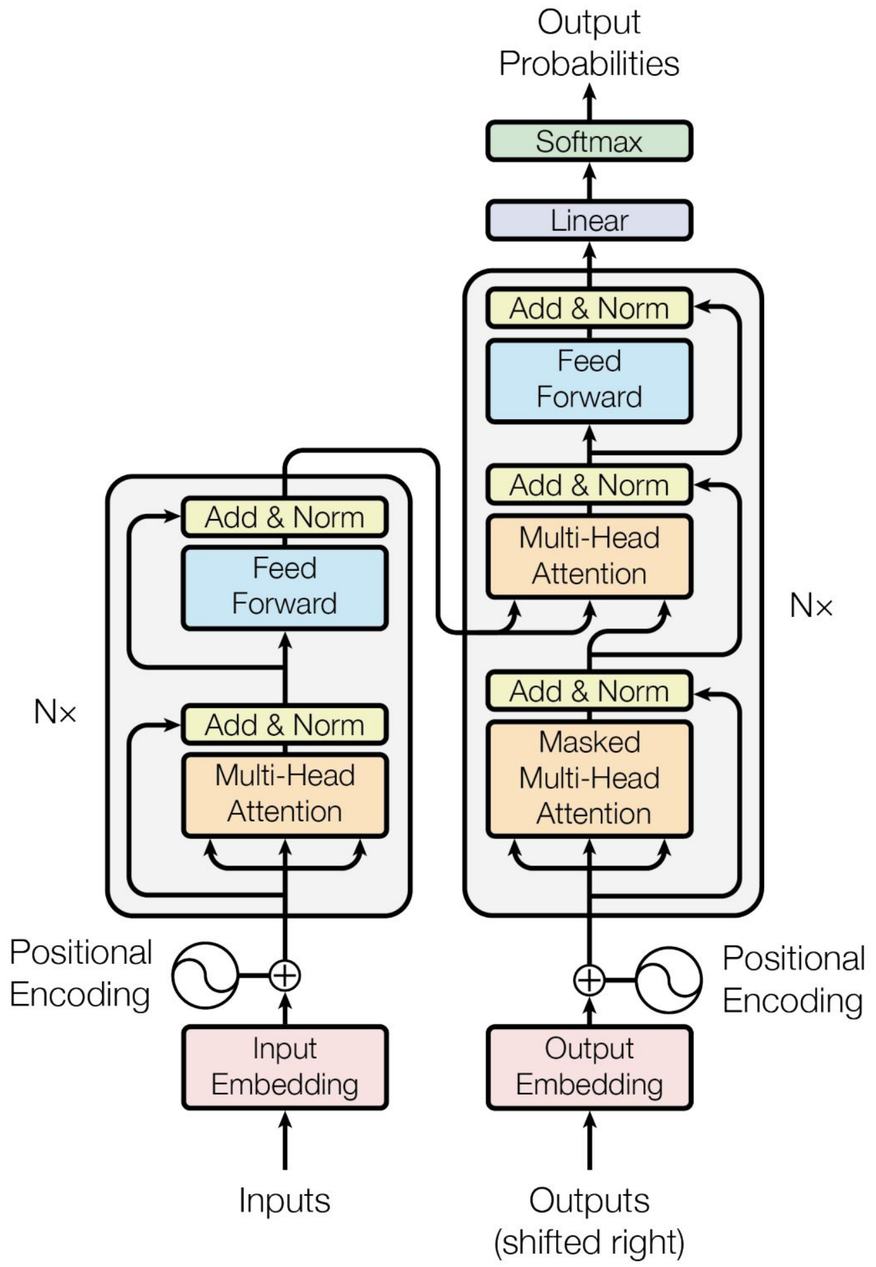
SEMANTIC SCHOLAR
A free, AI-powered research tool for scientific literature

arXiv

Common Corpus
Largest multilingual pretraining data.

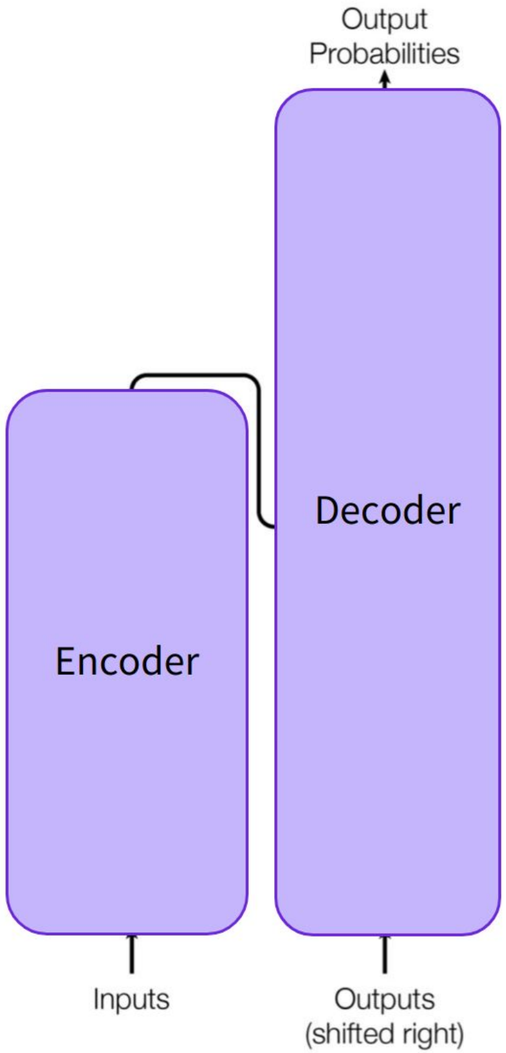
Welcome to Project Gutenberg

Project Gutenberg is a library of over 75,000 free eBooks



Simplified Architecture

Both sides combined = Encoder-Decoder aka Sequence-to-Sequence Transformer



Bi-directional

Self-attention

Uni-directional

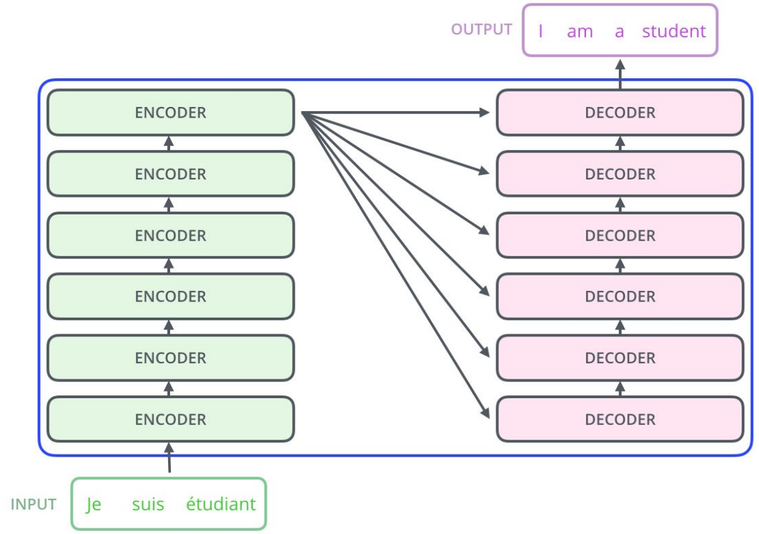
Auto-regressive

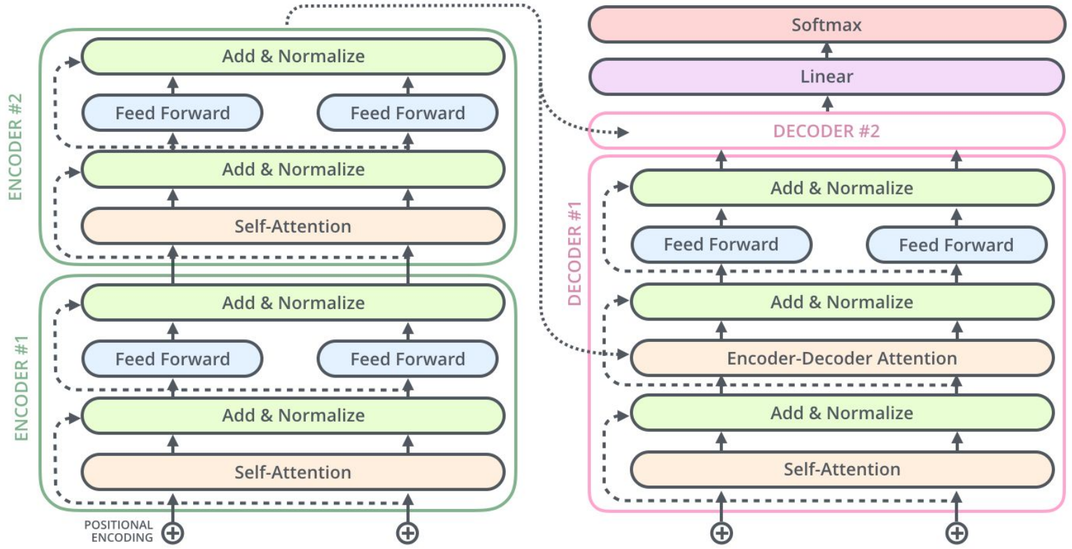
Masked self-attention

Cross-attention

Image credit: Hugging Face

A Closer Look





GPT-3's "magic"

96 decoder layers each with 1.8B parameters!

Images credit: The Illustrated Transformer

Decoders *(ChatGPT, Claude, Gemini, Llama, …)*

Generative Pre-trained Transformer

(good for generative AI, conversational AI, multimodal AI, etc.)

Encoders *(BERT, …)*

Bidirectional Encoder Representations from Transformers

(good for classification, named entity recognition, etc.)

Encoder-Decoder *(BART, …)*

Bidirectional Auto-Regressive Transformers

(good for machine translation, summarization, etc.)

Encoder-Decoder *(T5, …)*

Text-To-Text Transfer Transformer

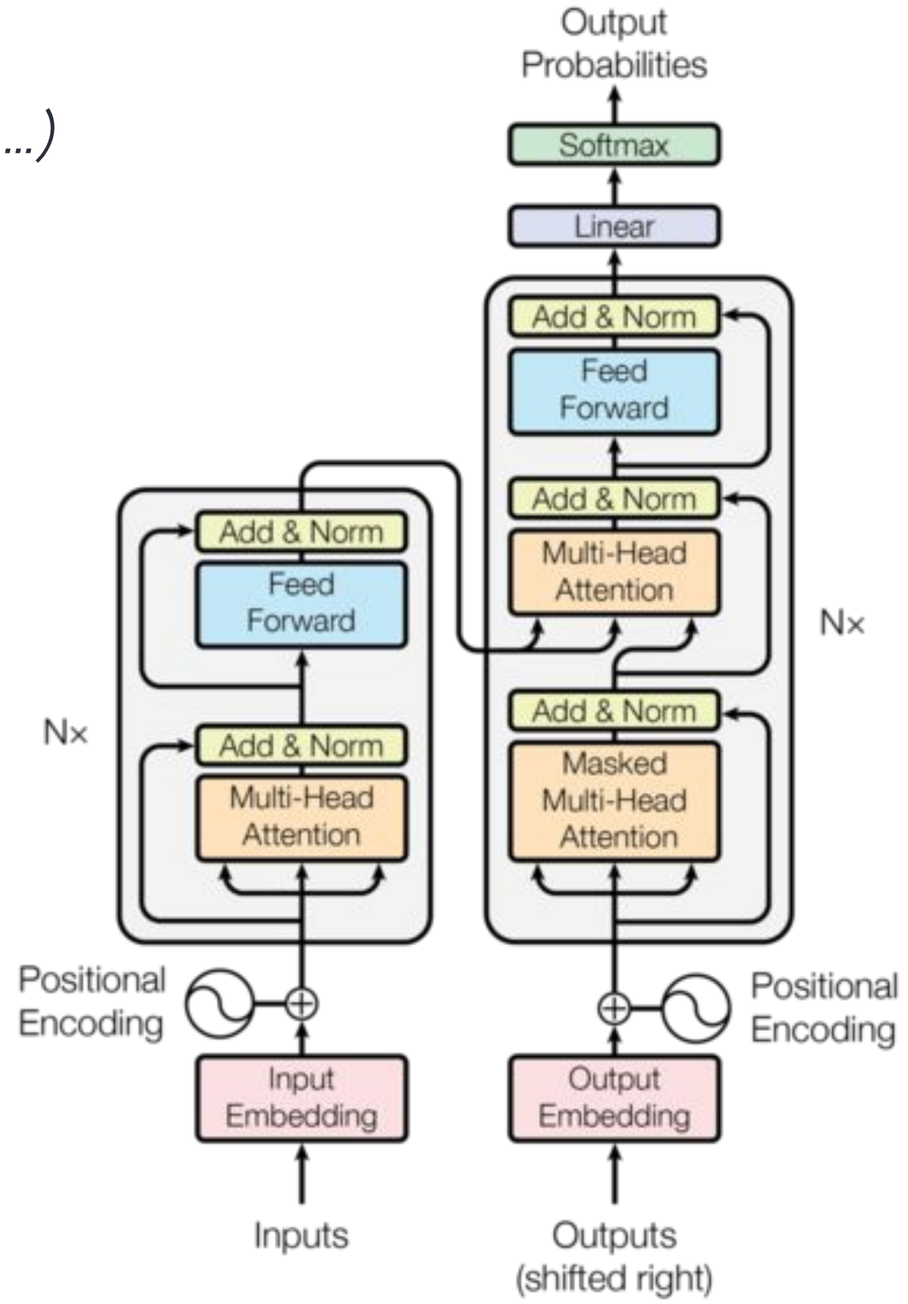


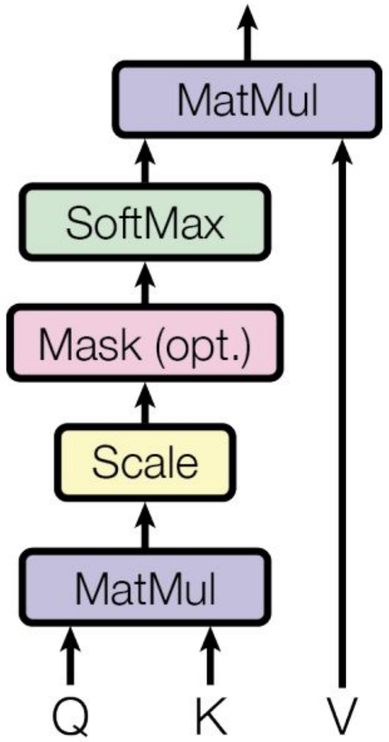
Image credit: Hugging Face

Attention & Attention Scores

Image via ChatGPT

- Attention scores model relationships between all input tokens/words
- Attention scores are calculated for each token/word relative to every other token/word
- Think of attention scores as parameters / weights that help generate new, better representations

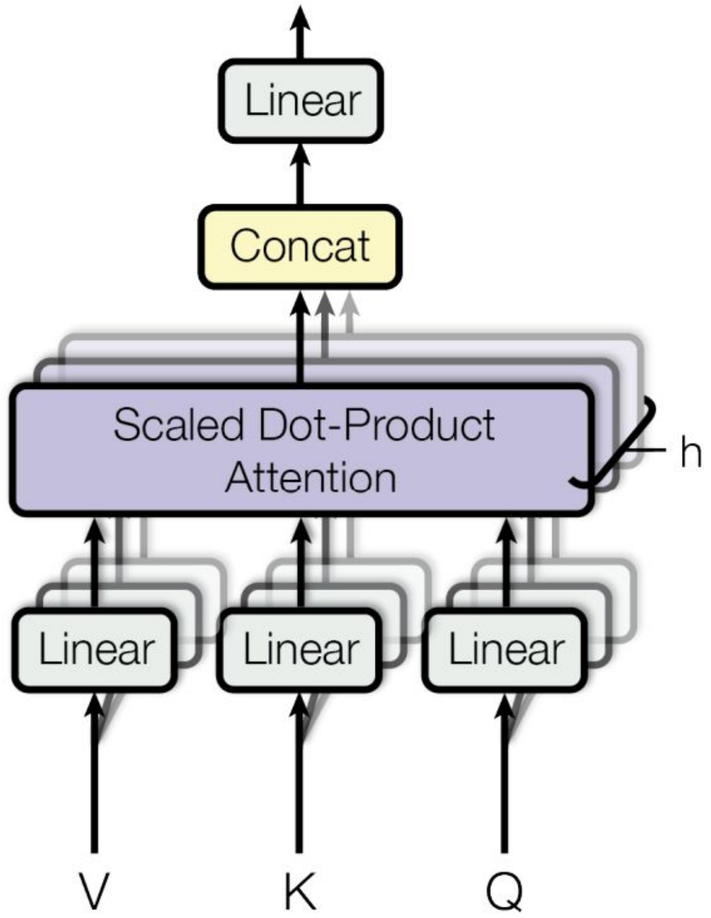
Scaled Dot-Product Attention



Input: Vector of word embeddings

Output: Vector representation of contextualized embeddings

Multi-Head Attention

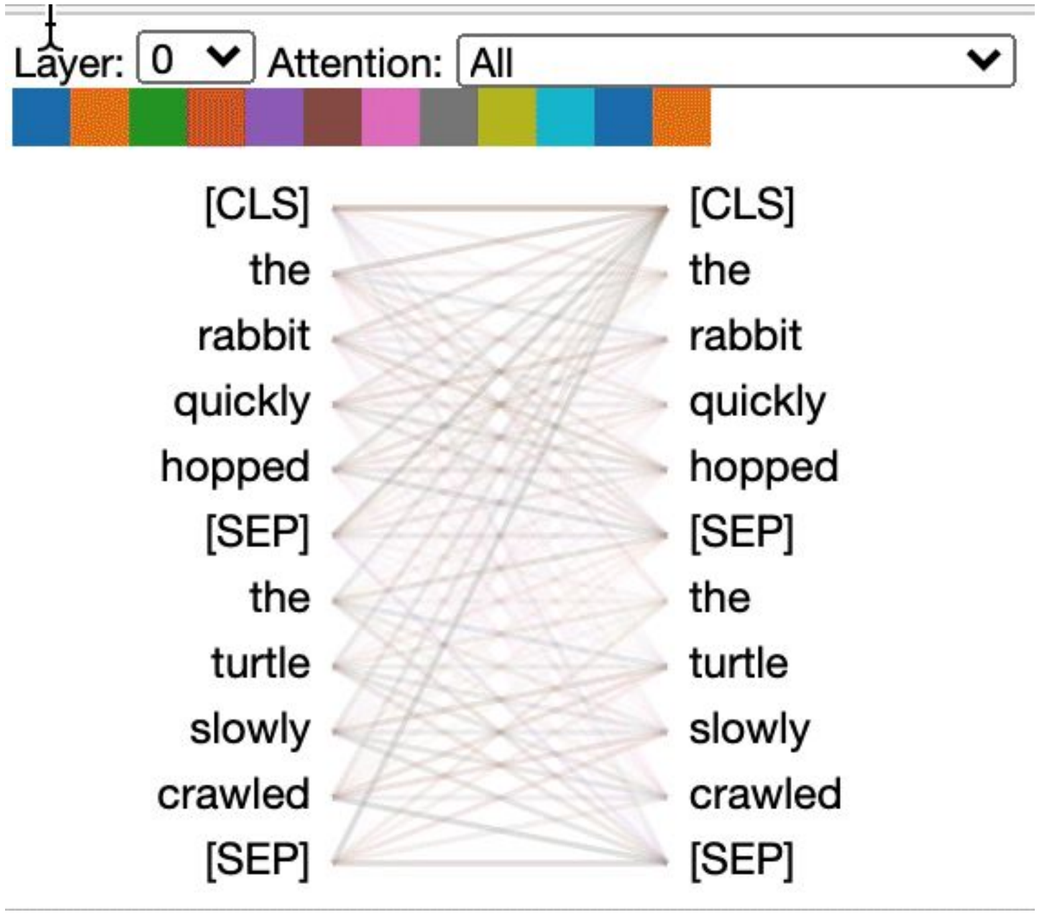


Input: Vector of word embeddings

Output: Vector representation of contextualized embeddings with deeper semantic / contextual understanding

Source (paper): Attention Is All You Need

Visualizing Attention



Try it out!

Animations credit: BertViz on GitHub

Activity

Route vs. Write (Transformers + Attention)

Use Case: Helpdesk Copilot for a company's internal support inbox

Goal: Decide what an encoder vs. decoder is best for, and what attention must focus on

***** Don't use any devices for assistance**
***** Follow guidance on activity steps**

Activity

Incoming Ticket (same input for both tasks)

“Hi. my flight tomorrow got cancelled. Can you move my reservation to Friday and refund the seat upgrade? Thanks, Sam.”

Task A: Route (Encoder-style output)

Pick 1 category: Travel Change I Refund I Other
Pick urgency: High I Normal
Select 5–7 words the model should attend to most

Task B: Draft (Decoder-style output)

Write the first 12–18 words of the reply
Underline words in the ticket your draft relies on most

***** Don’t use any devices for assistance**
***** Follow guidance on activity steps**

Activity

Discussion:

- Which words got the most “attention” for routing vs. drafting? Why?
- Encoder vs. decoder: what is each optimized to do with the same input?
- Where could a drafted reply “hallucinate” in this use case?
- What is one practical control to reduce risk without killing usefulness?

***** Don’t use any devices for assistance**
***** Follow guidance on activity steps**

Appendix

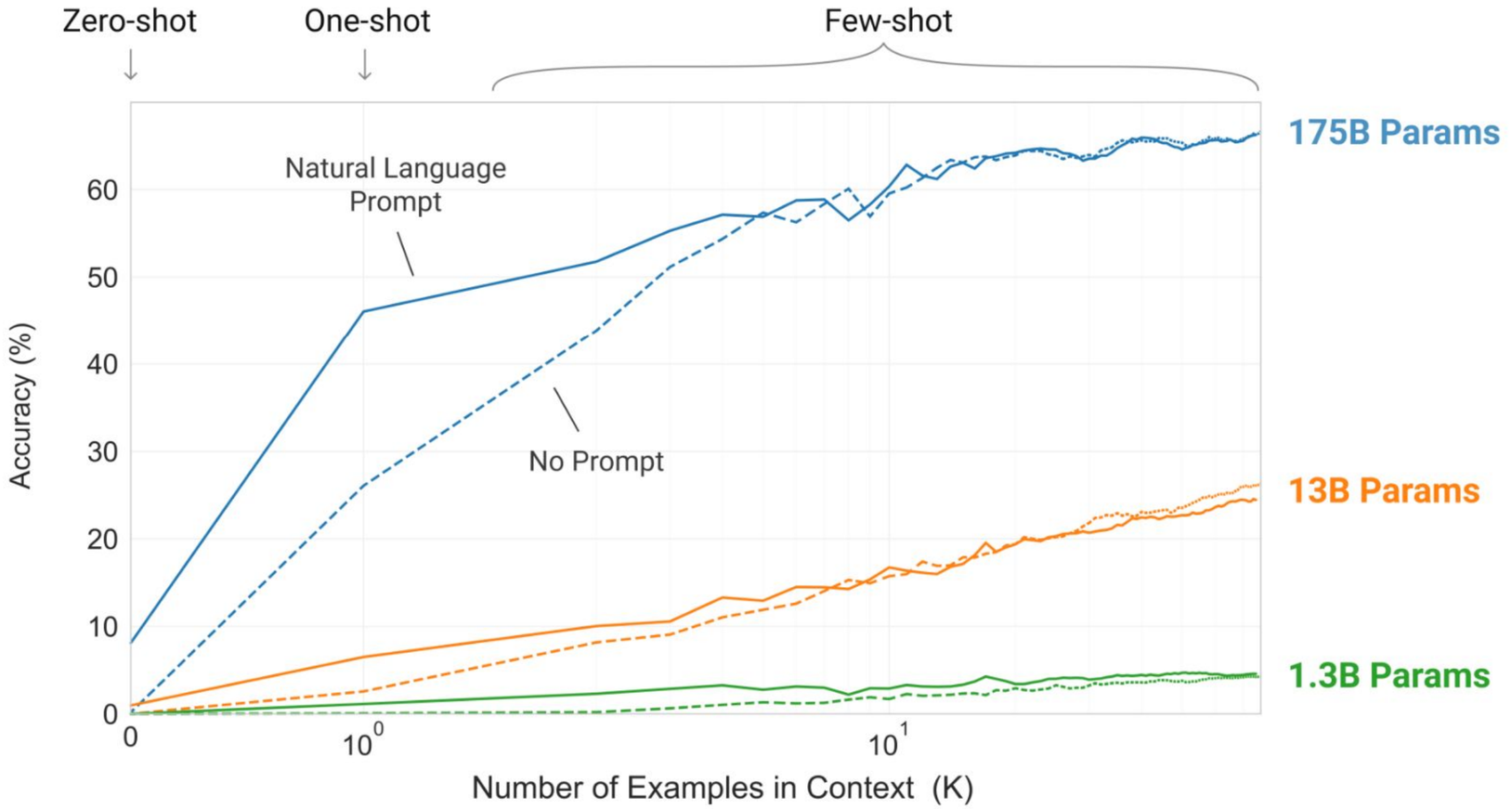


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning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

Source: Brown et al