

SECTION 6

Embeddings

How an LLM turns tokens into learned vectors

Slides based on Section 6 of the paper

1

From tokens to vectors

- **Look-up table:** tokens are the keys
- Vectors enter the model at **layer 0**
- Tokens replaced by **learned vectors**
- Layers then process these vectors

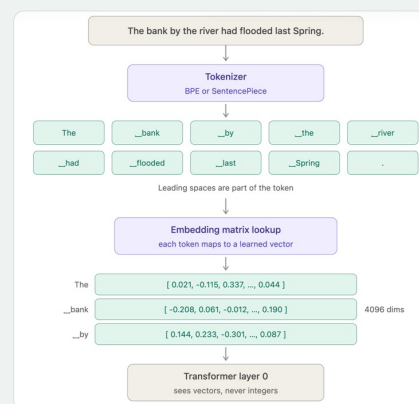


Figure 3: Words become learned vectors

2

| Tokenization comes first

- **Separate tokenizer**, trained on a corpus before the LLM itself
- **First pass**: regular expression splits on whitespace and punctuation
- **Byte-pair encoding**: repeatedly count adjacent symbol pairs, merge the most frequent into a new symbol
- **Stop** when a target vocabulary size is reached
- Vocabulary is driven by **frequency, not meaning**

| Tokenizer example

SAMPLE SENTENCE · OPENAI'S O200K_BASE TOKENIZER

Despite the unanticipated 37% downturn, ACME Corp.'s shareholders — surprisingly — remained optimistic about Q3 2024 earnings.

- Near one-to-one mapping of tokens to words, with occasional exceptions
- “unanticipated” splits into “un” + “anticipated”
- **Not morphological awareness**: no notion of a prefix
- “un” and “anticipated” simply frequent enough to merge; the full form was not

Neighbors of “parent”

- Visualizes words that frequently **co-occur** with “parent”
- Embedding holds **10,000 words**
- Vector space has **200 dimensions**
- Produced by Project Tensorflow (projector.tensorflow.org)

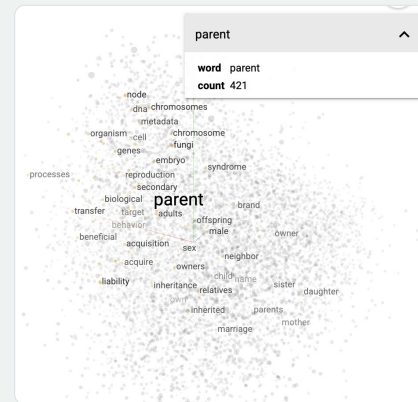


Figure 4: Words that co-occur with “parent”

Section 6 · Embeddings

6

Most similar words to “parent”

- Top eleven words similar to “parent”
- Includes **offspring, organism, chromosome, sister, child**
- “mother” is in **spot eight**
- “father” is a **no-show** among the top 50 nearest terms

Nearest points in the original space:	
offspring	0.620
organism	0.638
chromosome	0.649
chromosomes	0.657
parents	0.664
sister	0.668
child	0.693
mother	0.700
reproduction	0.701
acquisition	0.701
biological	0.717

Figure 5: Words similar to “parent”

Section 6 · Embeddings

7

| Vector arithmetic in static embeddings

- GloVe and Word2Vec support **semantic vector arithmetic**

king – man + woman = queen

- Adding two vectors yields a **combined** one
- “red” + “apple” lands near other red fruits, such as strawberries

Section 6 · Embeddings

| How embeddings are learned in LLMs

- LLMs are (pre)trained by ingesting **a lot of text**
- Initially **random vectors** change to capture associations with other vectors
- Vectors evolve into a **structured space**

Section 6 · Embeddings

| Neighbors do not define a concept

- Words around “parent” **do not define** the concept
- A definition would show the **role** of a parent, e.g., offering emotional support.
- These are words that **co-occur**

Section 6 · Embeddings

11

| More than next-word prediction

- Pre-training objective: **token prediction**
- Computation runs over **vectors**, not tokens
- Vectors carry **far more information** than the tokens do
- “Just predicting the next word” **understates** what happens internally

Section 6 · Embeddings

12

| Tokens versus vectors

Tokens

- The **input**
- The **output**
- The **training target**
- The **unit of the loss**
- Appear in the **interface** to the LLM

Vectors

- The **medium of processing**
- Carry far more information than the tokens do
- Evolve from random values into a structured space