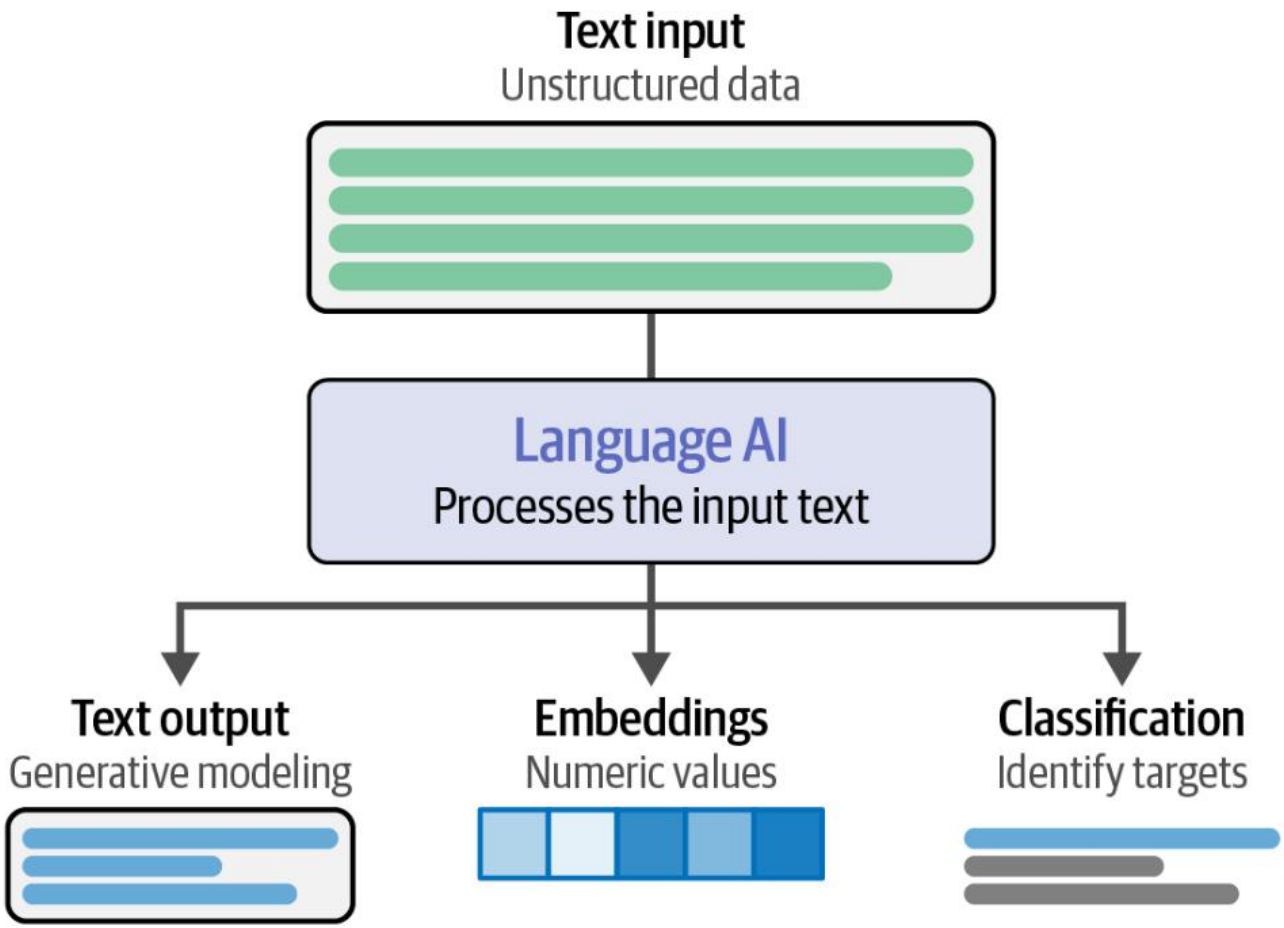
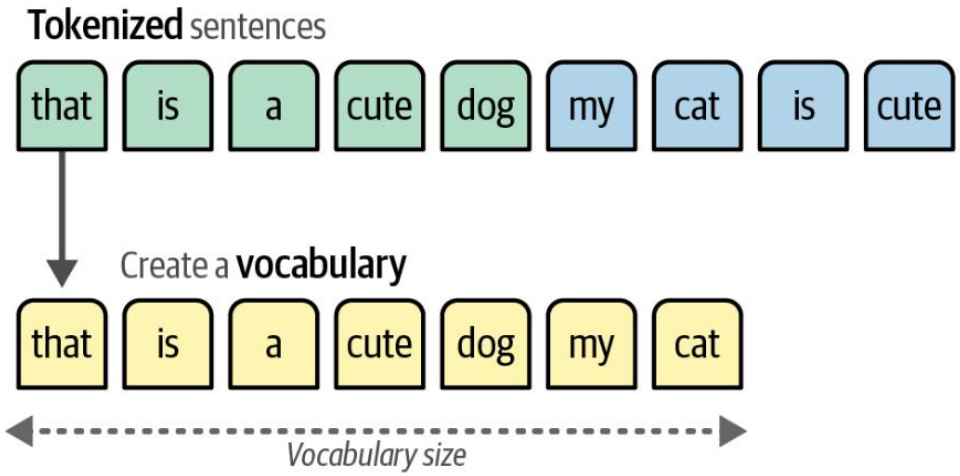
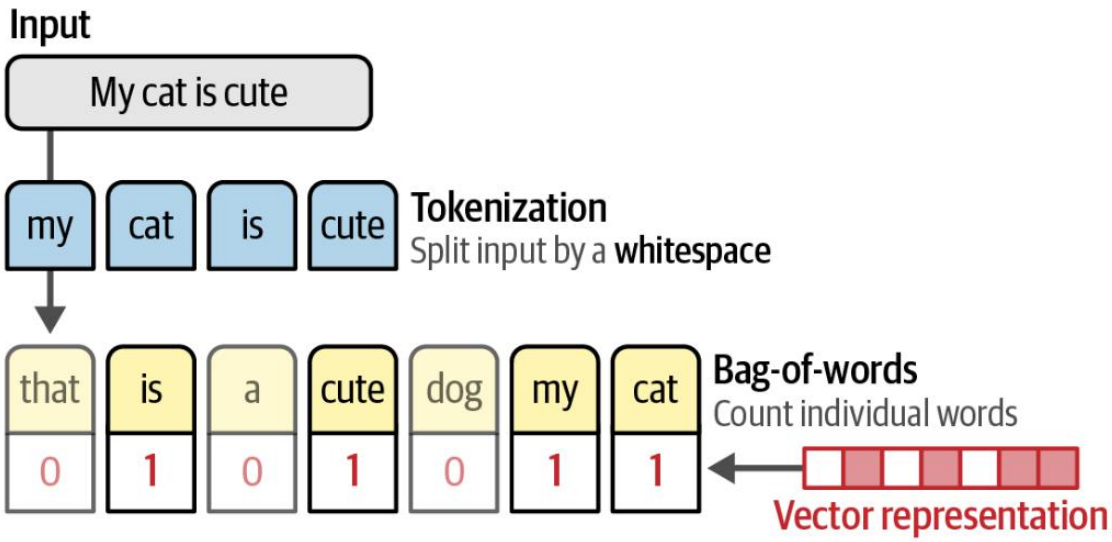
Introduction to Natural Language Processing (NLP)

Tasks


Text input
Unstructured data
Language AI
Processes the input text
Text output
Generative modeling
Embeddings
Numeric values
Classification
Identify targets

Text representation. BoW

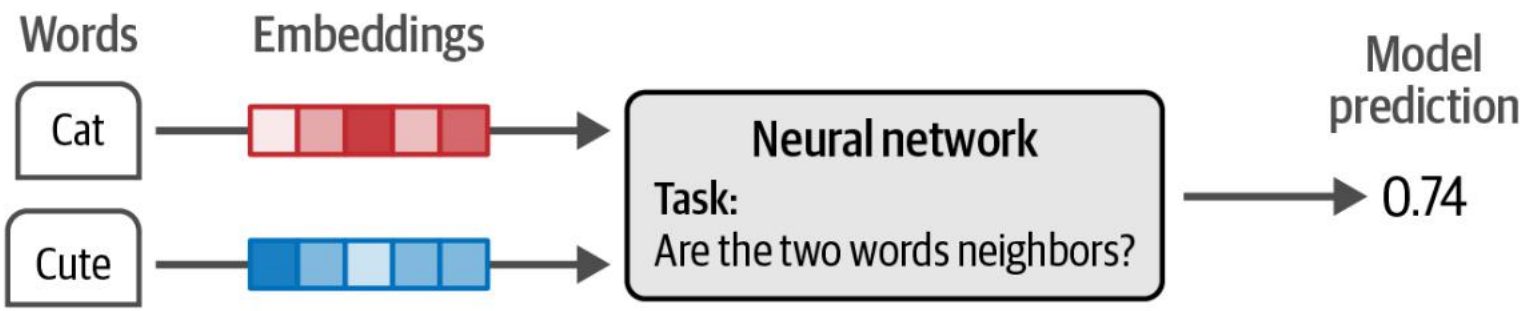




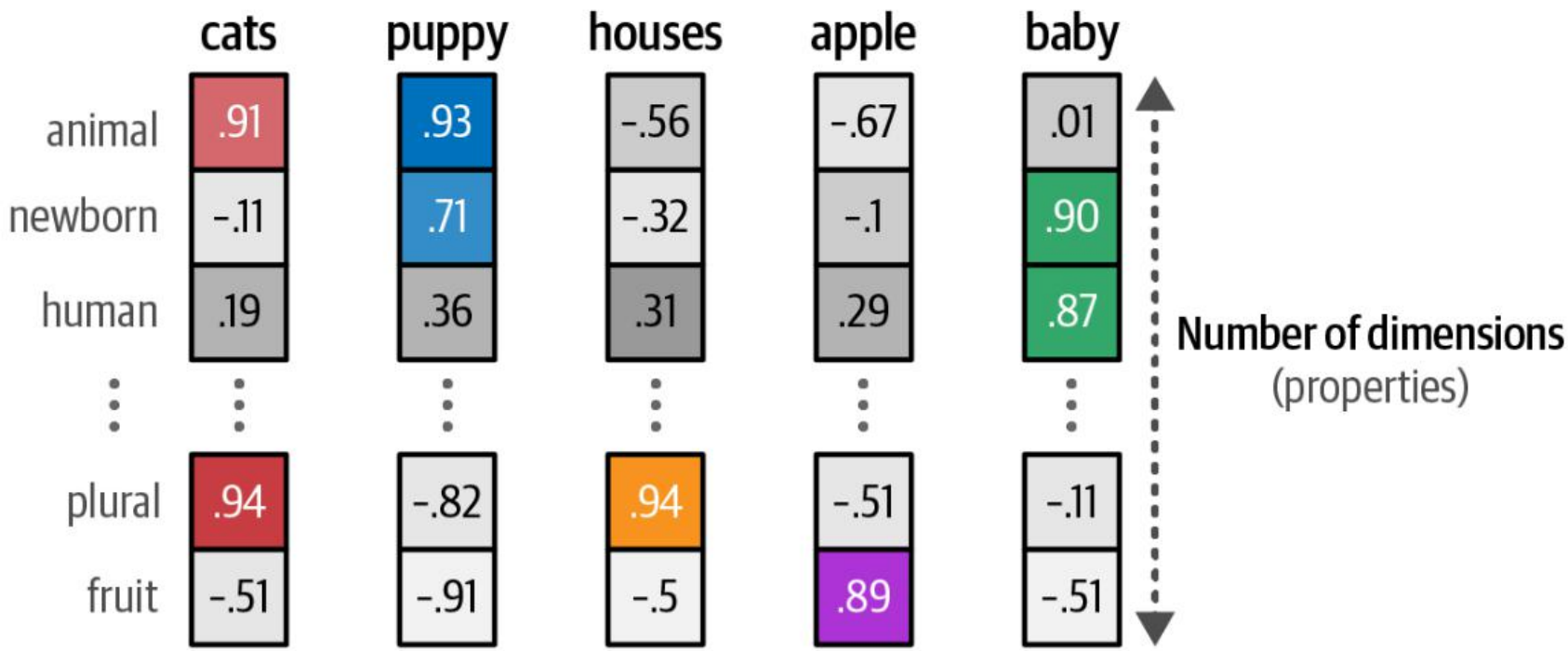
+ PCA

Better representation: Dense Vectors

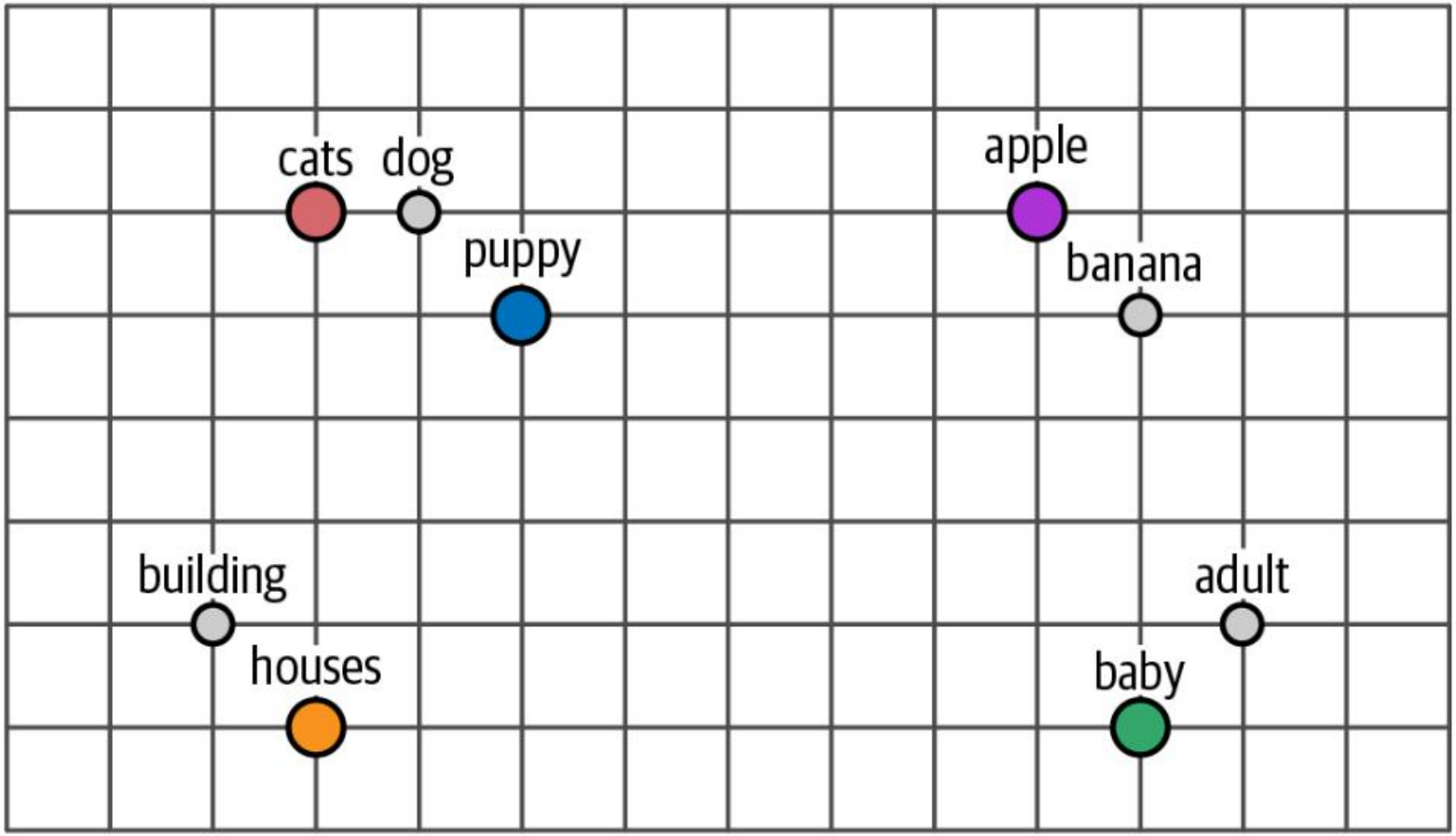
- Word2Vect: 1-hot encoding to vector



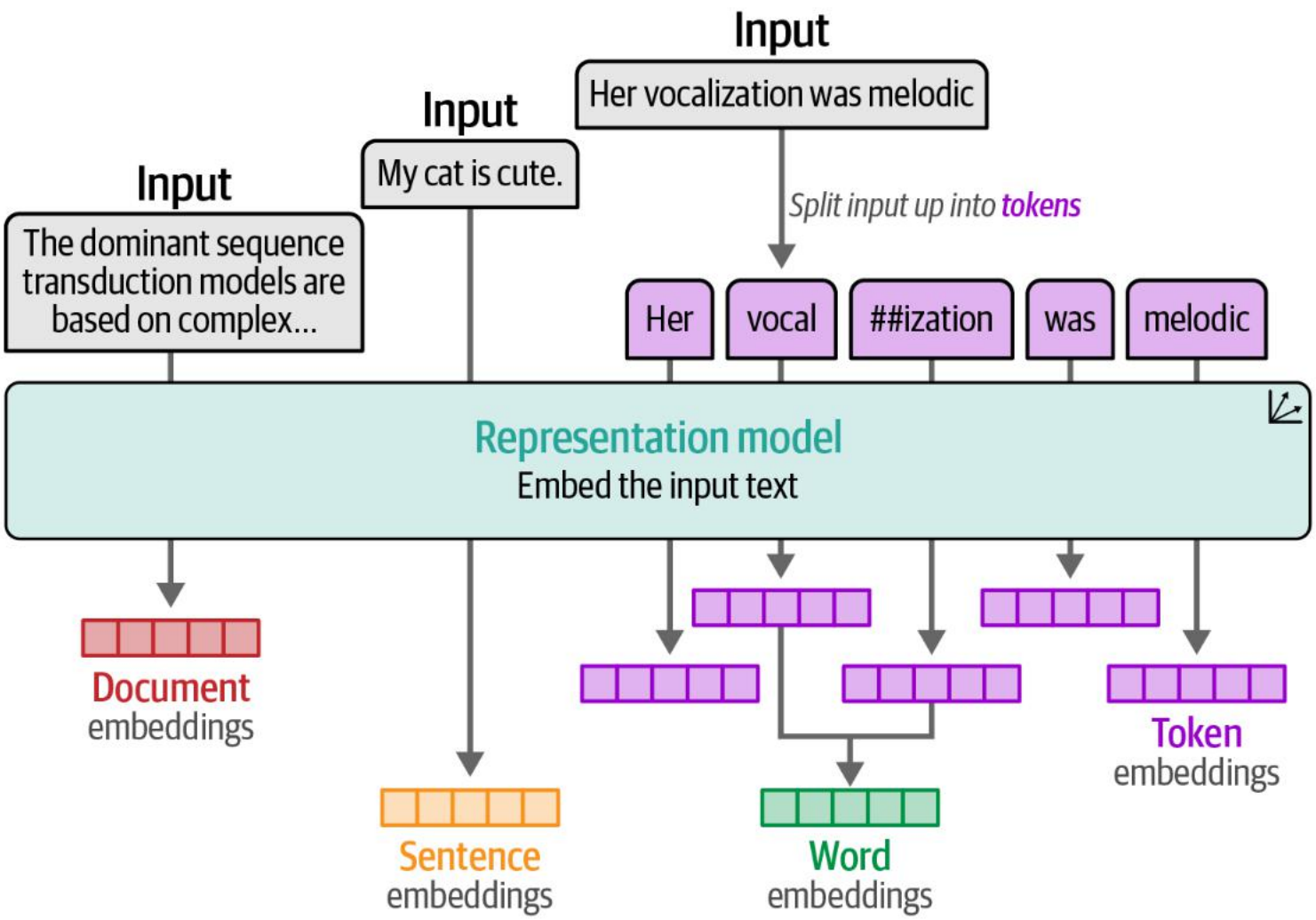
What is supposed to learn



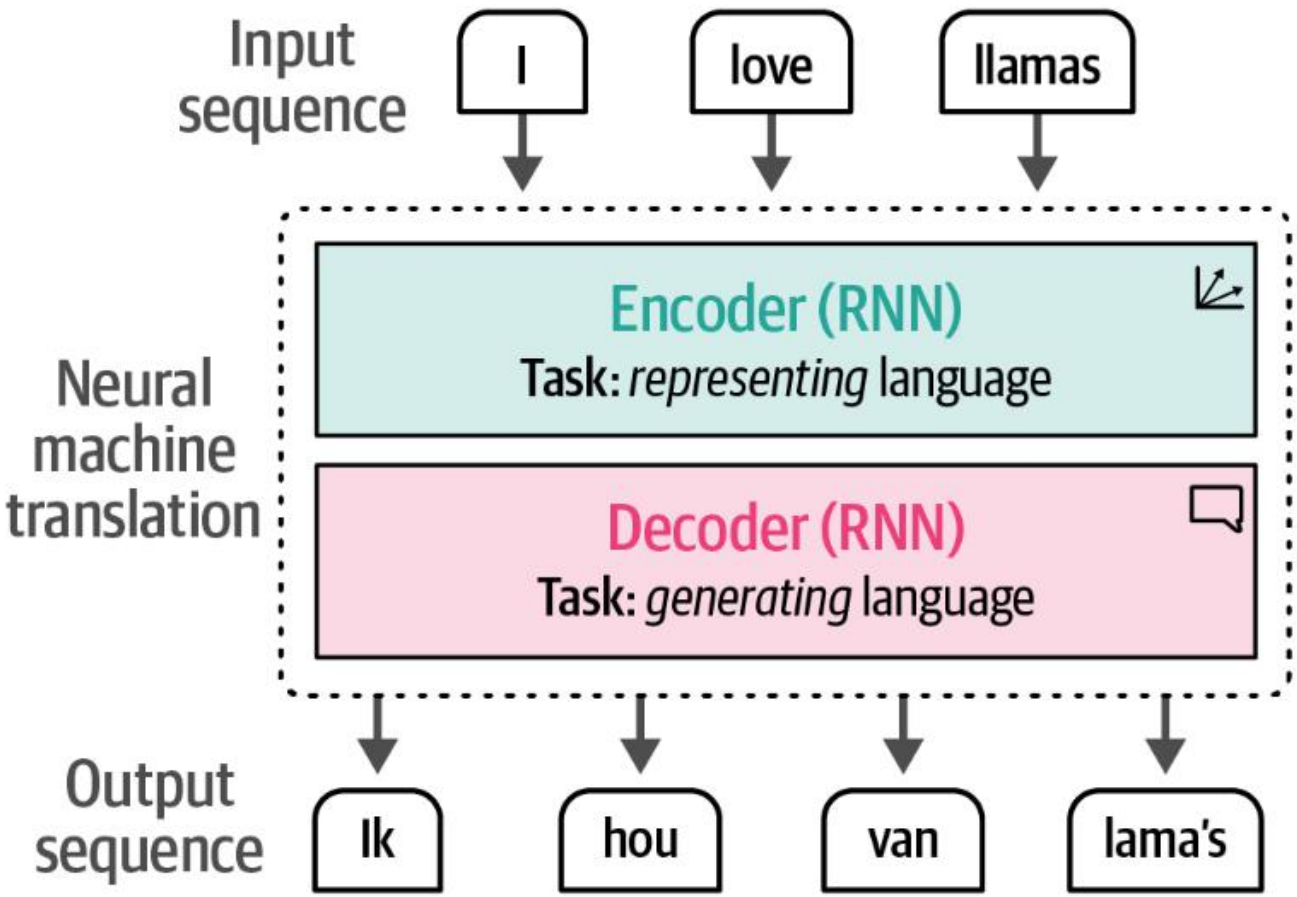
Distances are preserved



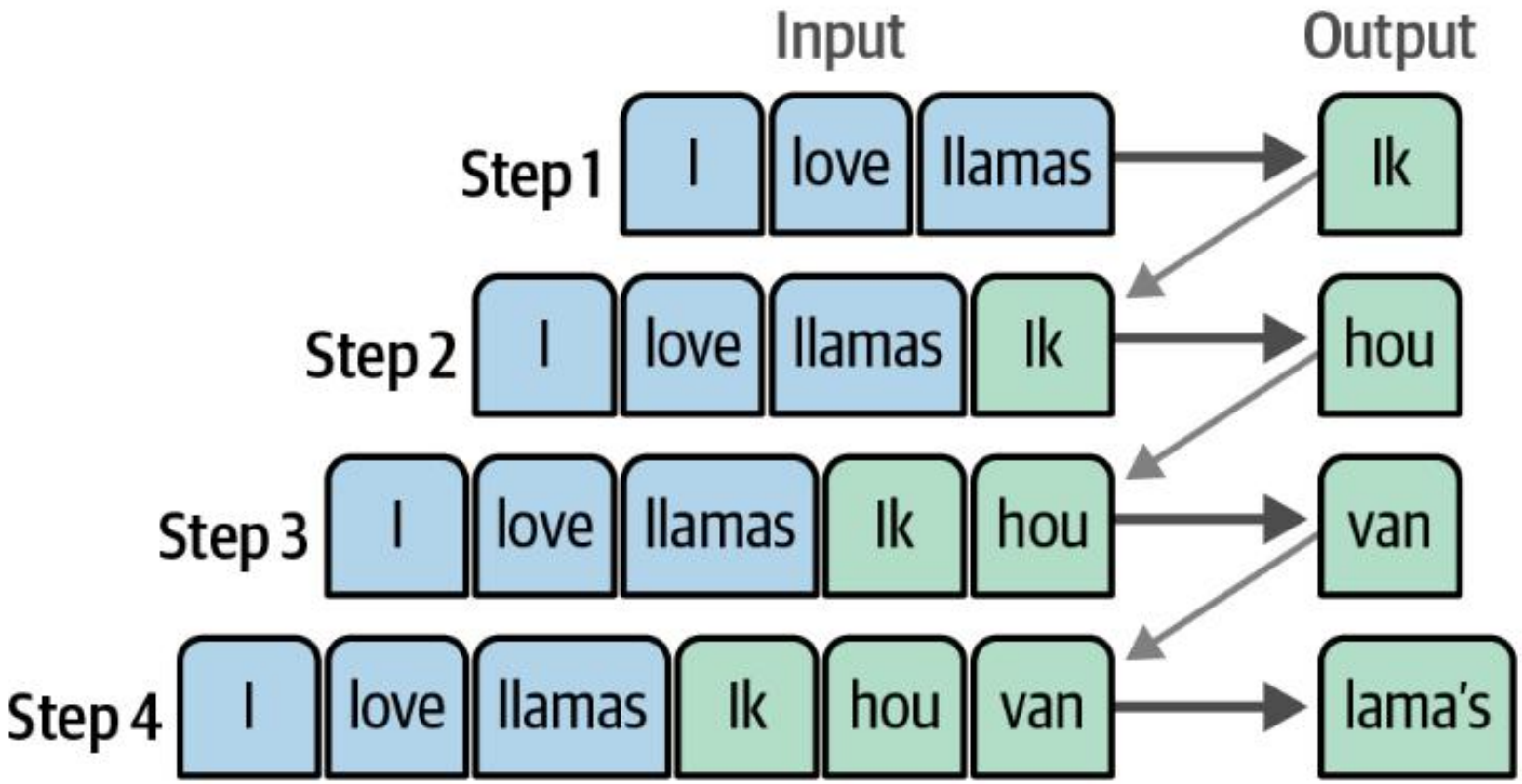
Types of embeddings



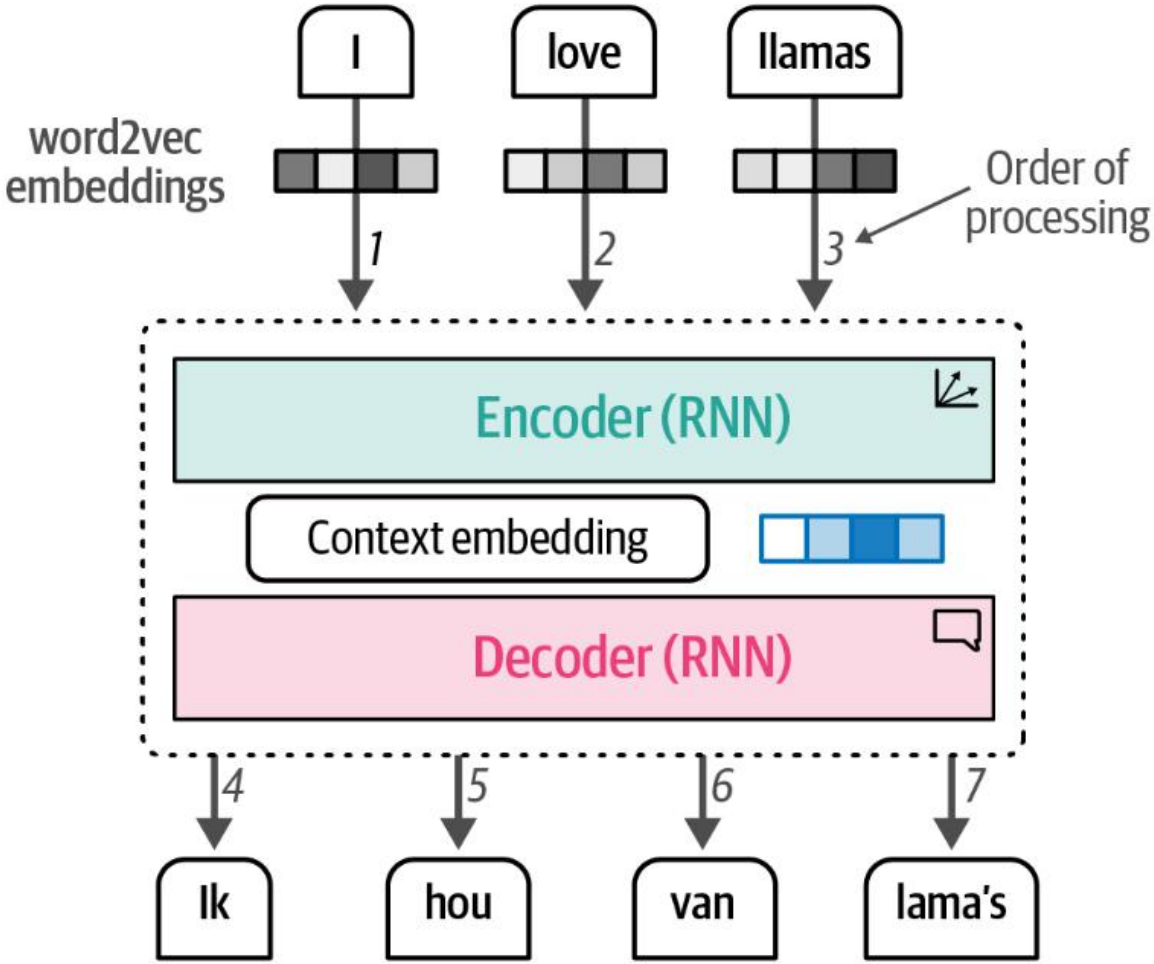
Encoding and Decoding Context with Attention



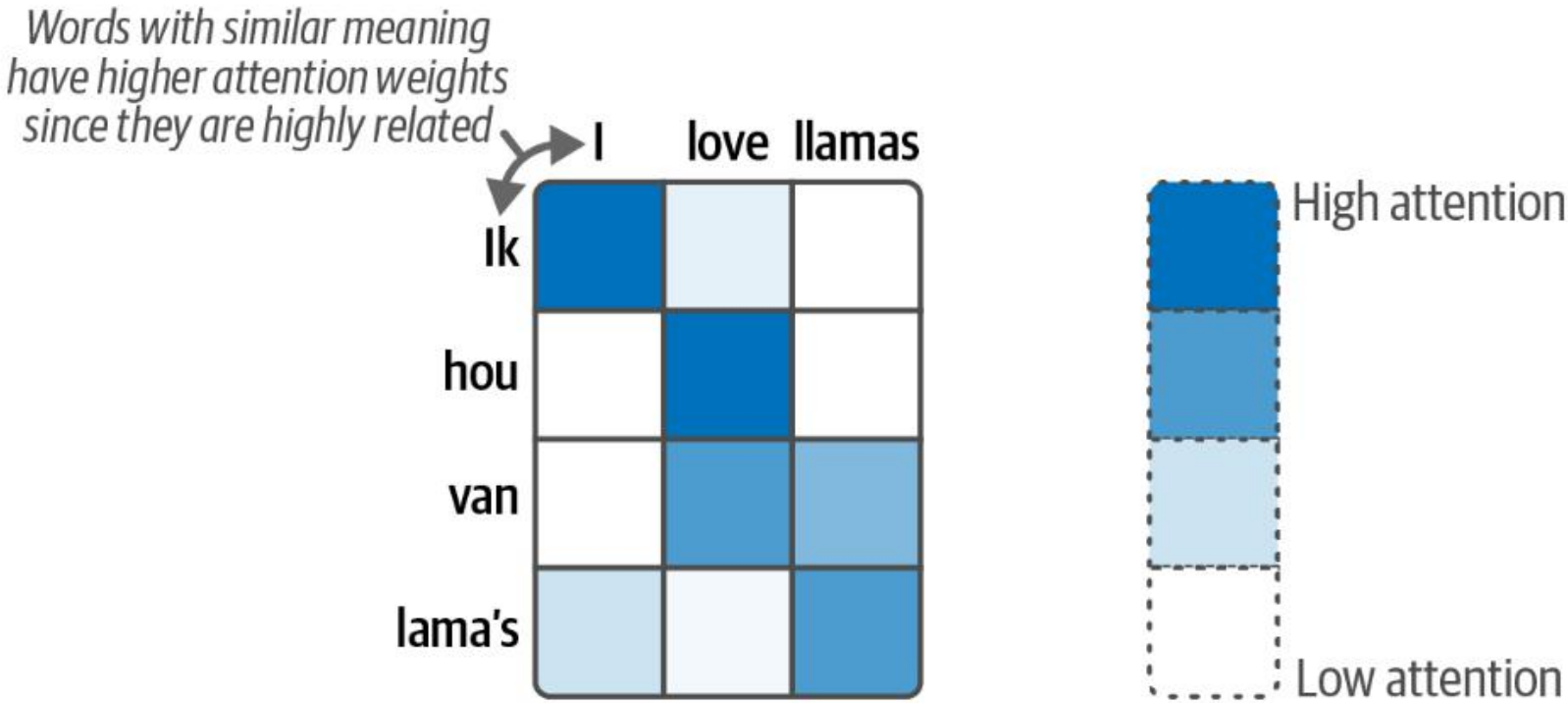
Decoder in action



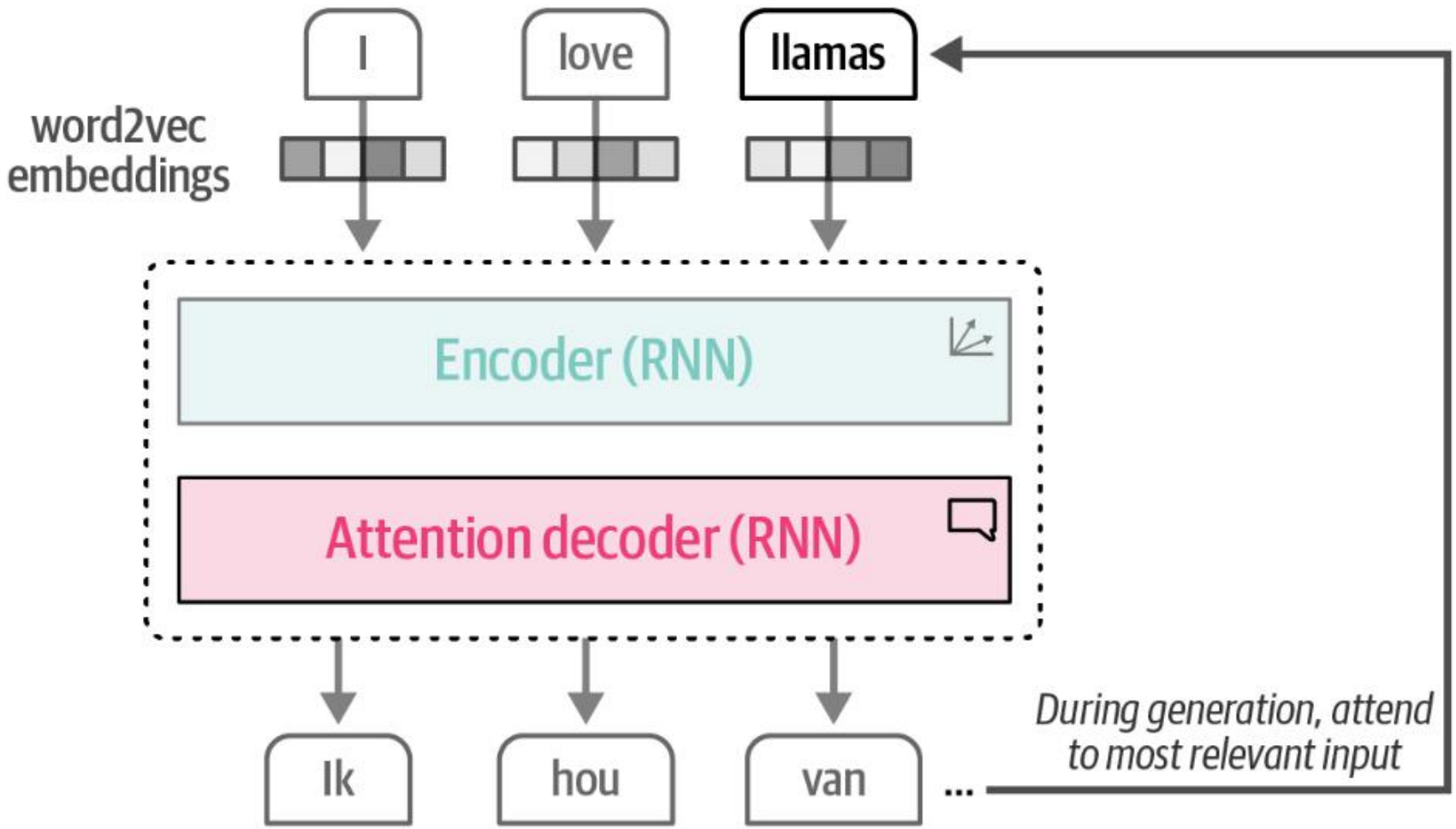
Using word2vect



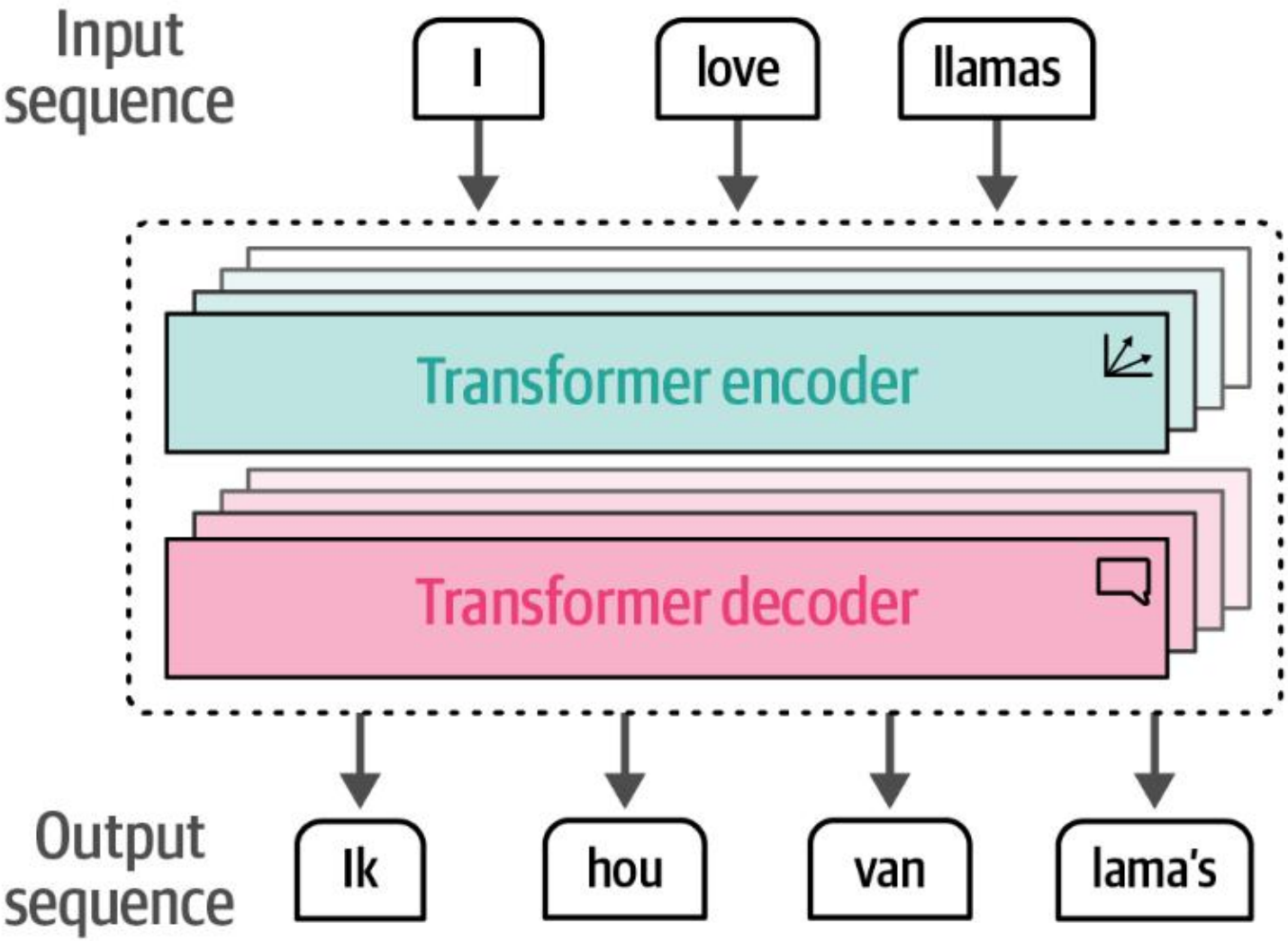
Attention mechanism



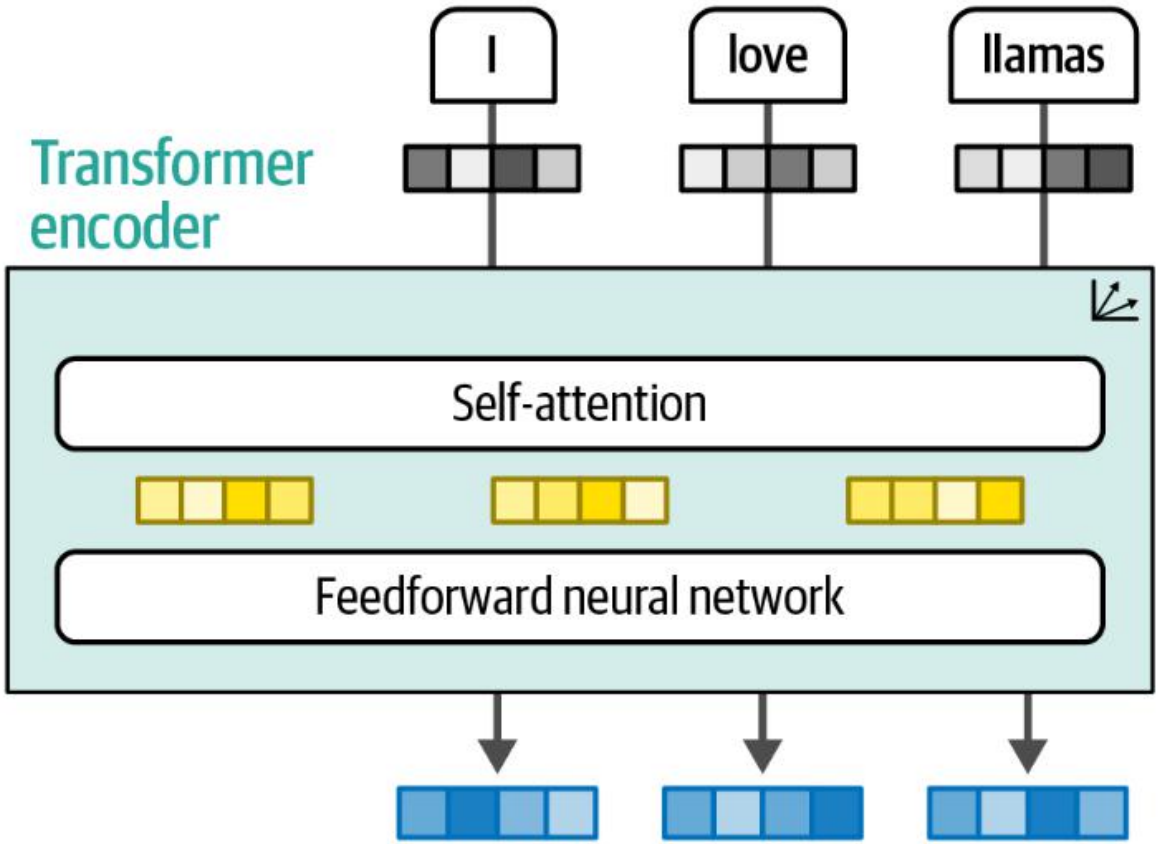
Adding attention

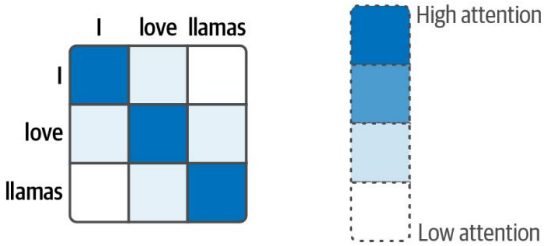


Attention is all you need

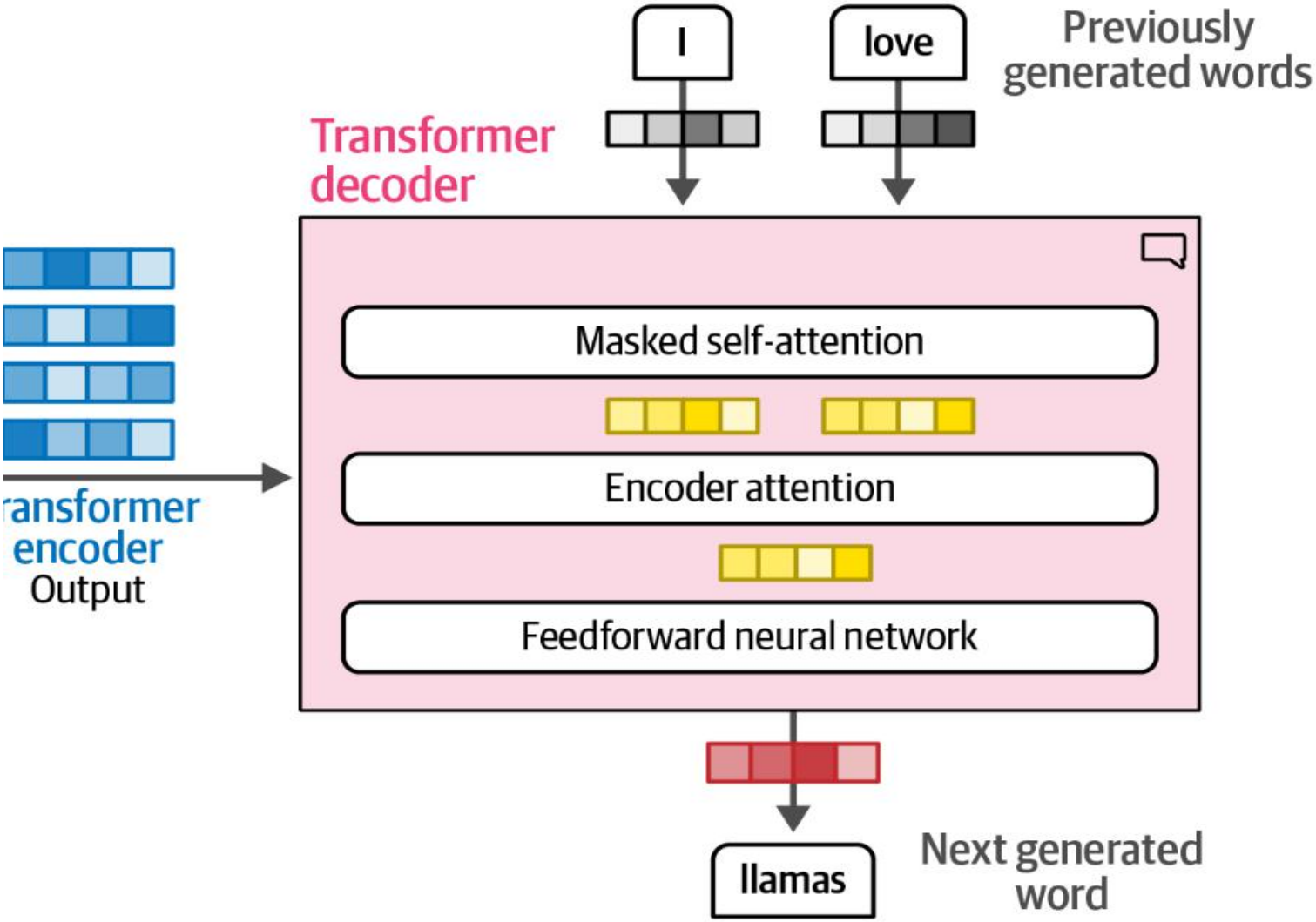


Encoder with self-attention

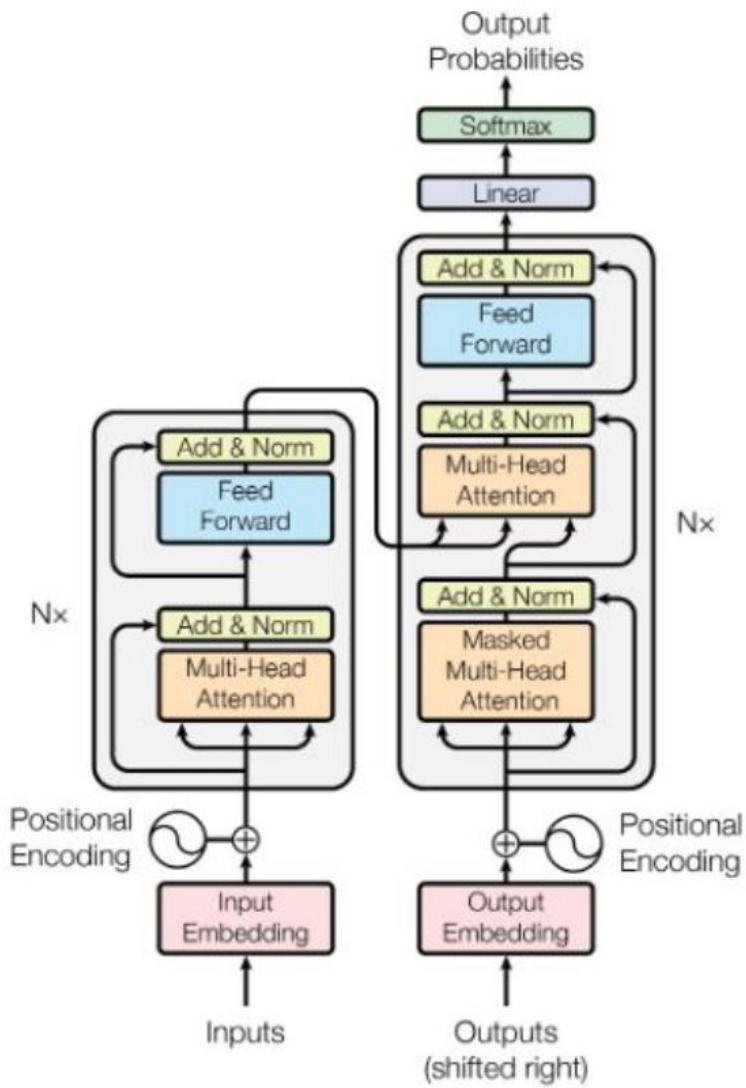




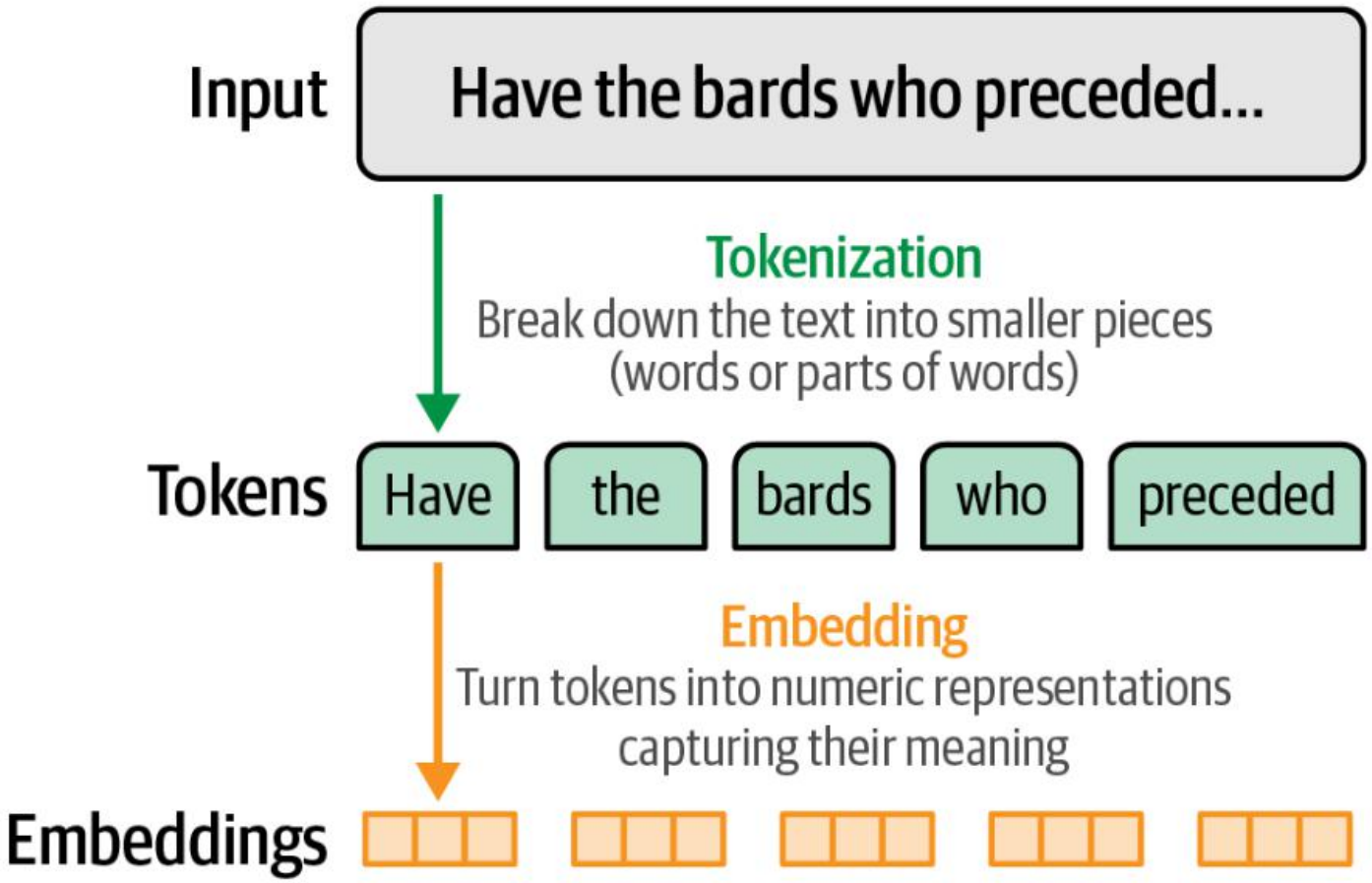
Decoder with attention



The transformer architecture



Tokenizers


Input
Have the bards who preceded...
Tokenization
Break down the text into smaller pieces
(words or parts of words)
Tokens
Have
the
bards
who
preceded
Embedding
Turn tokens into numeric representations
capturing their meaning
Embeddings

A real tokenizer

GPT-3.5 & GPT-4 GPT-3 (Legacy)

Have the bards who preceded me left any theme unsung?

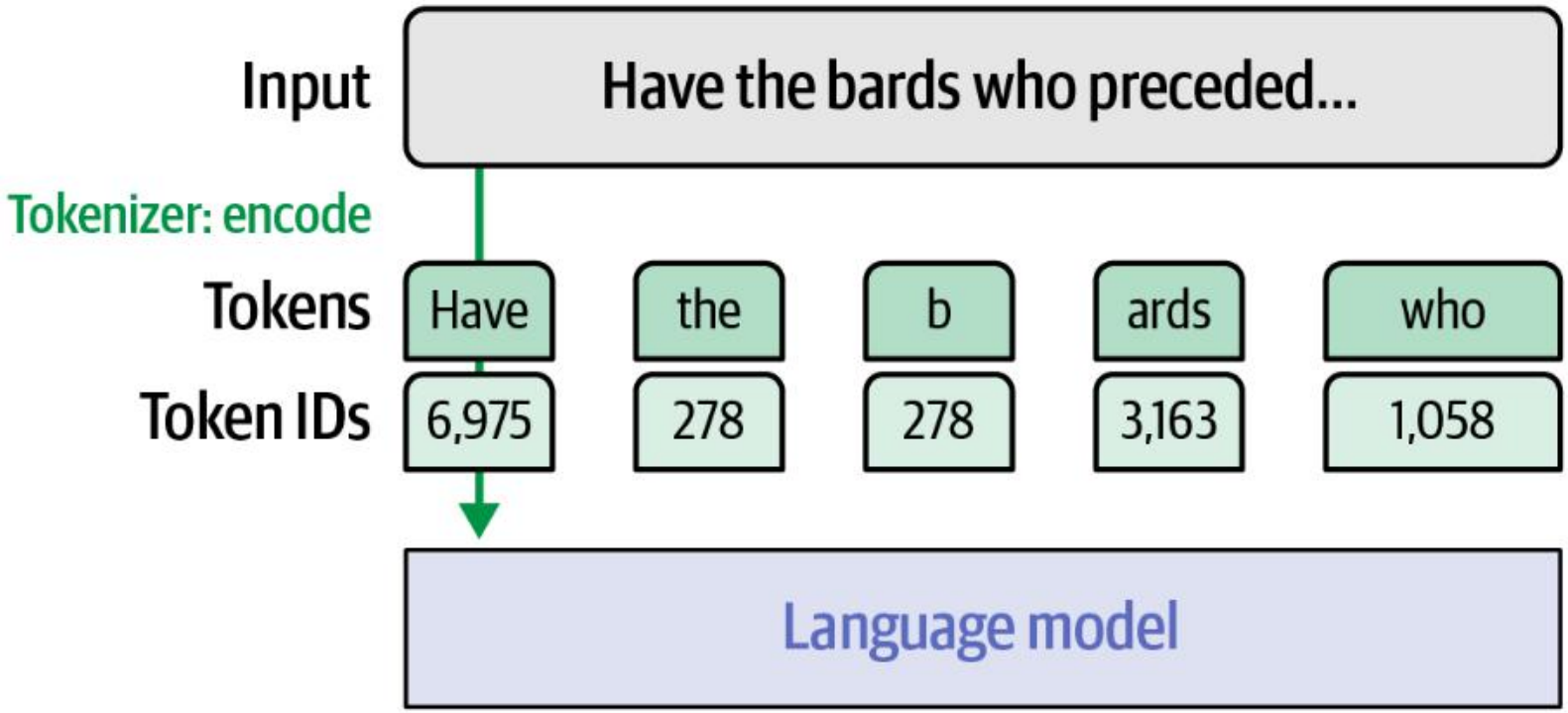
Clear Show example

Tokens	Characters
13	53

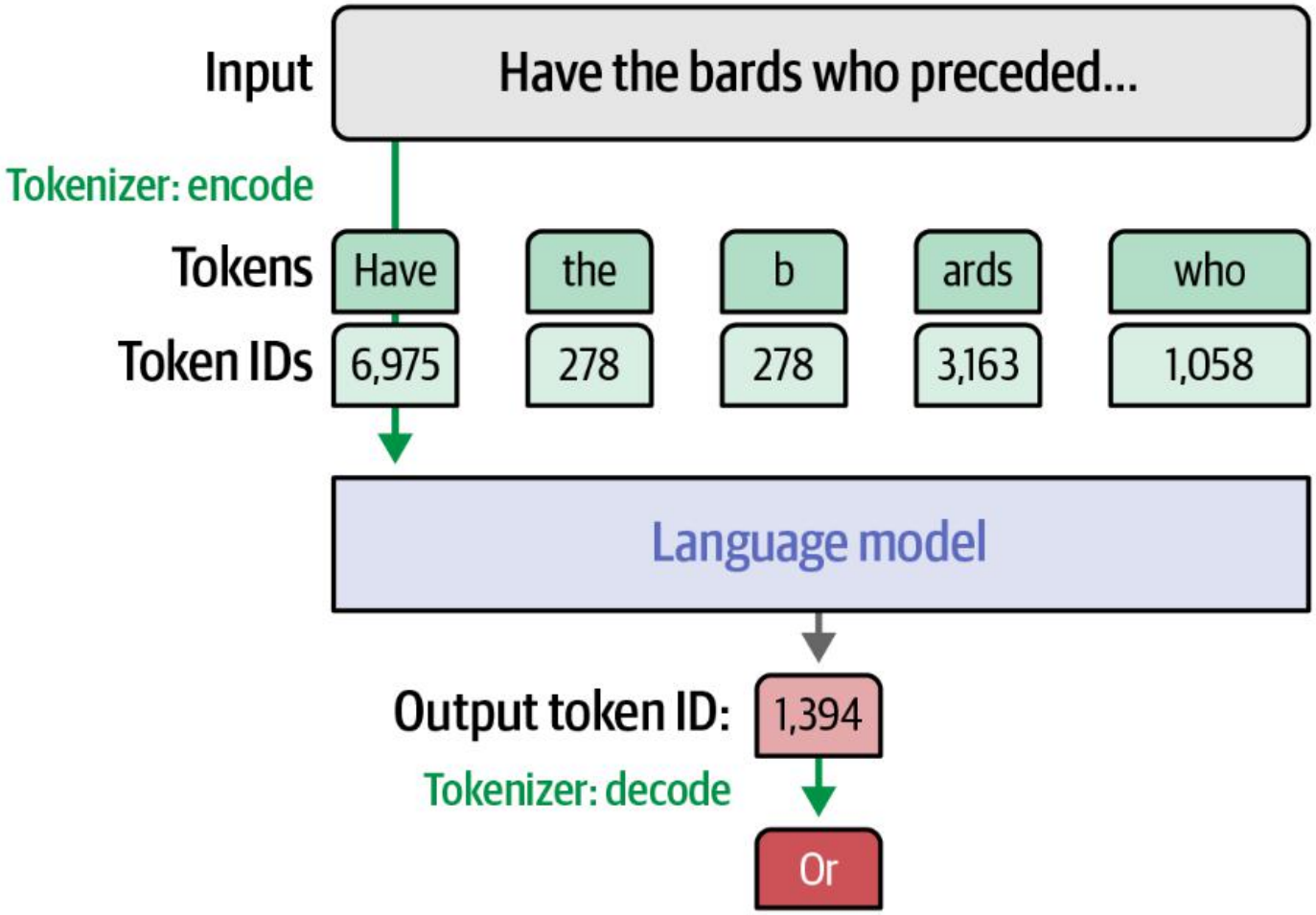
Have the bards who preceded me left any theme unsung?

Text Token IDs

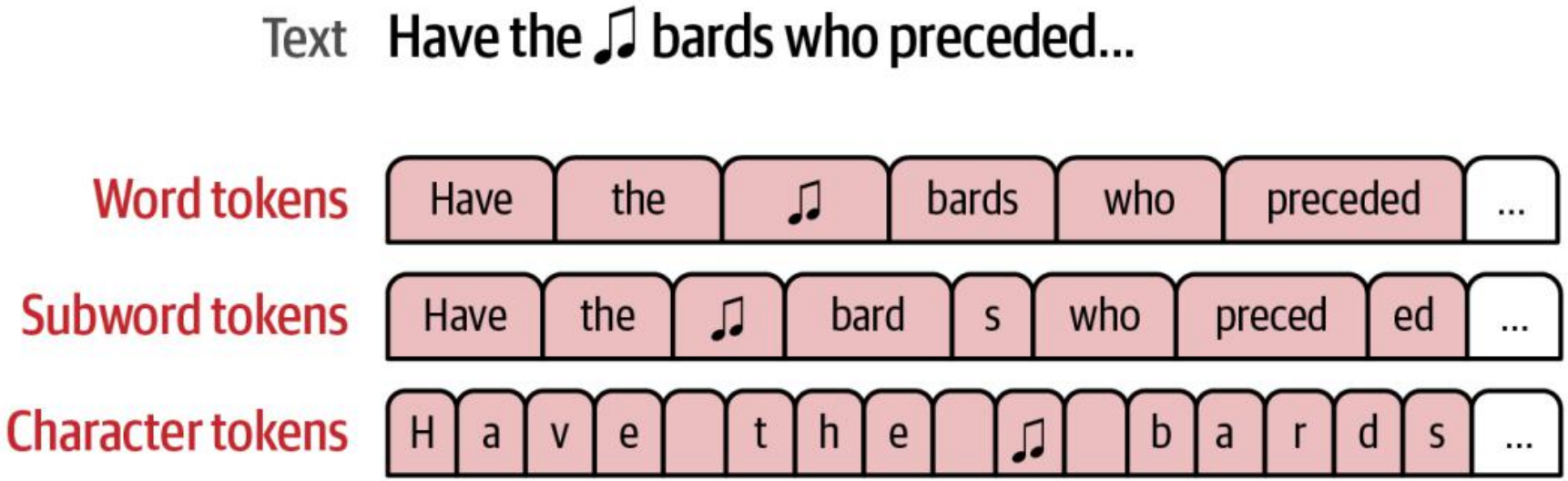
Each token has unique ID



Reverse process



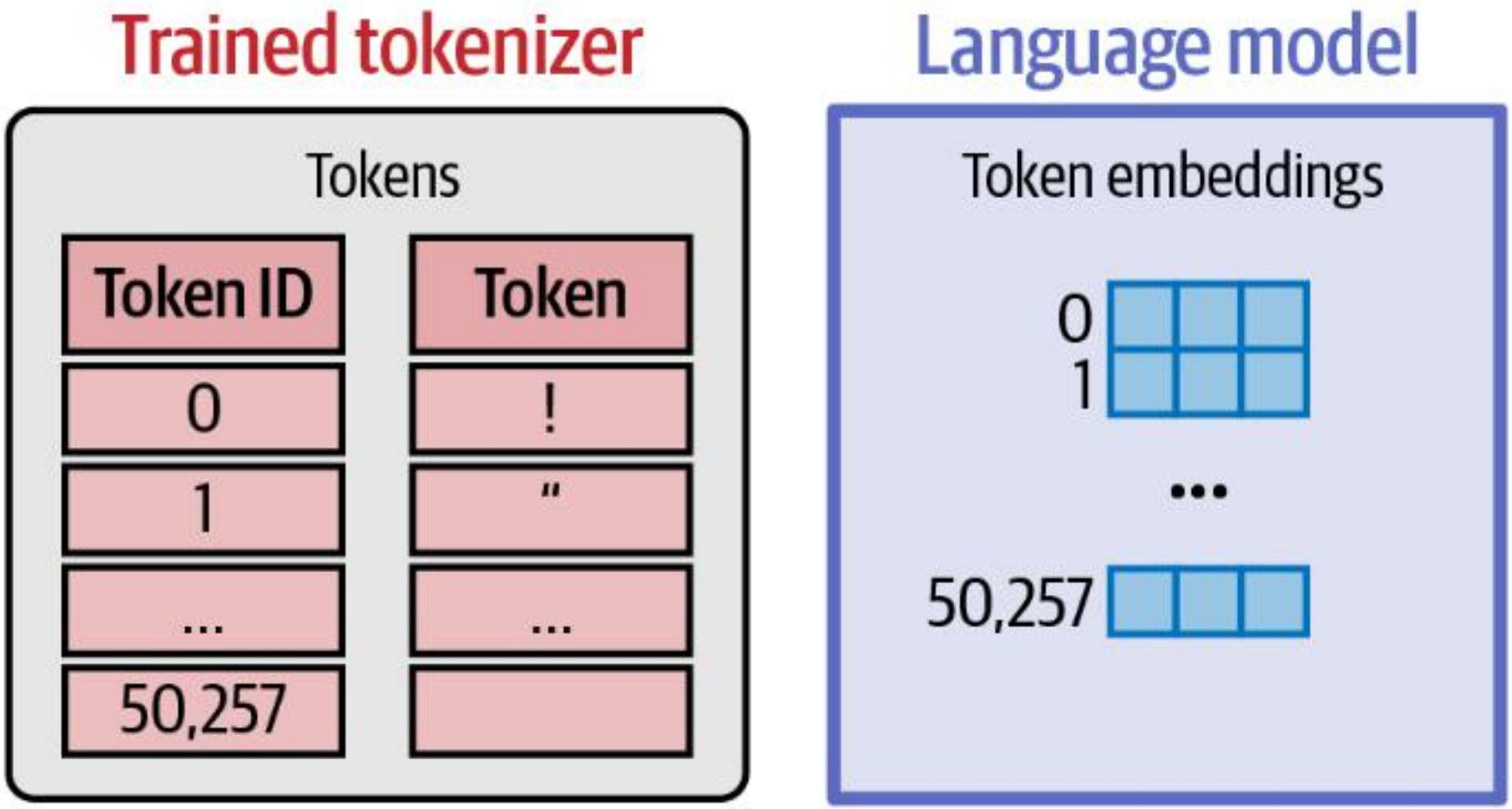
Types of tokenizers



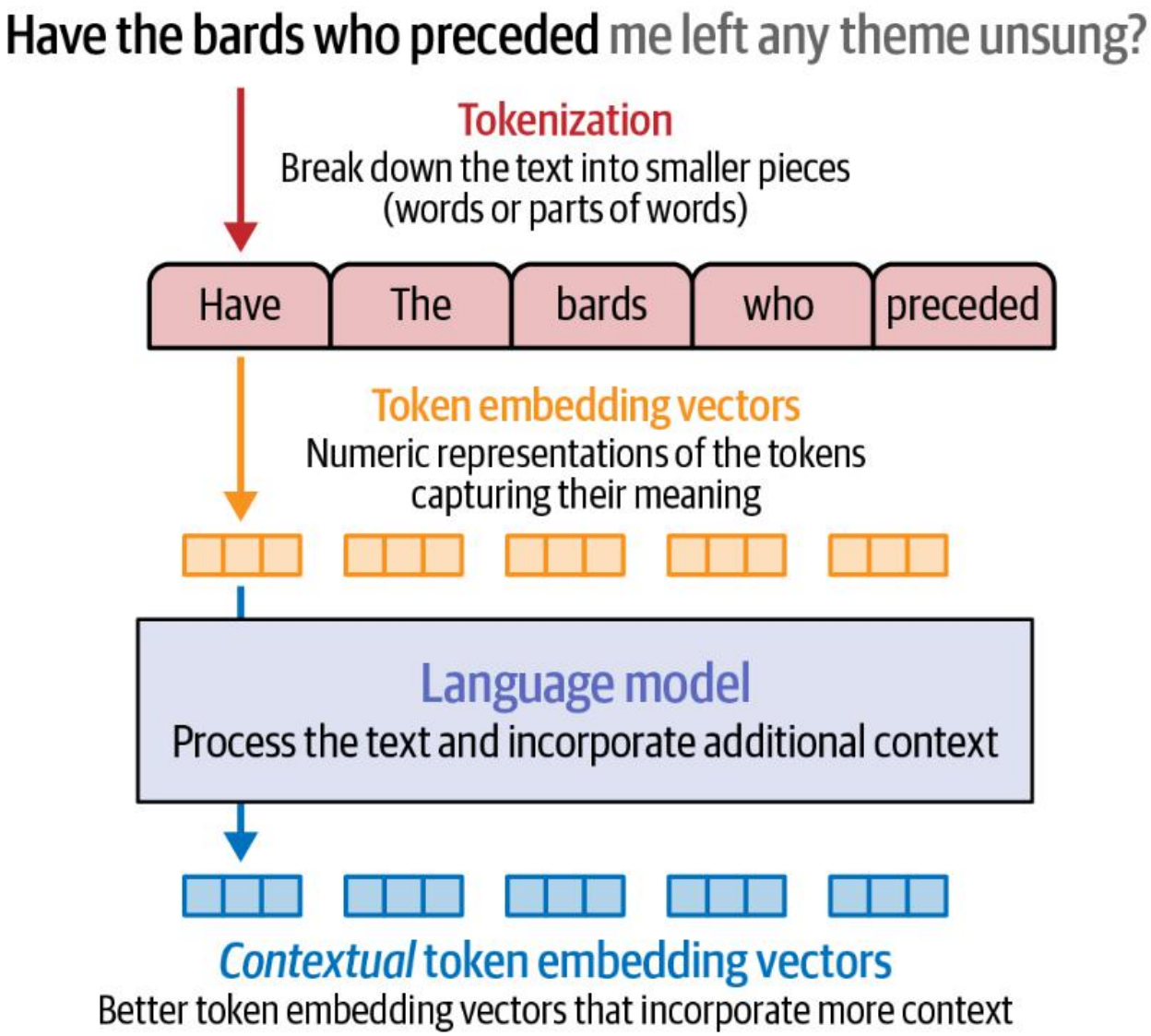
Some subword tokenizers

BERT base model (uncased)	[CLS] english and capital ##ization [UNK] [UNK] show _ token ##s false none eli ##f = = > = else : two tab ##s : " " three tab ##s : " " 12 . 0 * 50 = 600 [SEP]
BERT base model (cased)	[CLS] English and CA ##PI ##TA ##L ##I ##Z ##AT ##ION [UNK] [UNK] show _ token ##s F ##als ##e None el ##if = = > = else : two ta ##bs : " " Three ta ##bs : " " 12 . 0 * 50 = 600 [SEP]
GPT-2	English and CAP ITAL IZ ATION Ġ Ġ Ġ Ġ Ġ Ġ show _ t ok ens False None el if == >= else : two tabs :" " Three tabs : " " 12 . 0 * 50 = 600
FLAN-T5	English and CA PI TAL IZ ATION <unk> <unk> show _ to ken s Fal s e None e l if = = > = else : two tab s : " " Three tab s : " " 12. 0 * 50 = 600 </s>
GPT-4	English and CAPITAL IZATION Ġ Ġ Ġ Ġ Ġ Ġ show _tokens False None elif == >= else : two tabs :" " Three tabs : " " " 12 . 0 * 50 = 600

Token Embeddings



Adding context



Text embedding