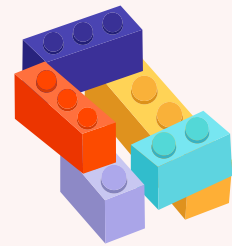


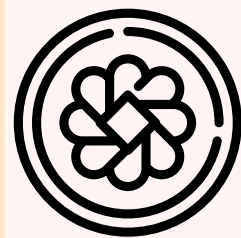
# Next Token Prediction in Decoder-Only Models

## Key Terms

- **Tokens** are the basic building blocks of text for an AI model (can be a full word, fragment of a word, or punctuation). *They help the AI model understand the text.*
- **Decoder-Only Model** is a type of model that takes a prompt and processes it token by token, predicting the next one until it finishes the answer.
- **Decoder-Layers** are multiple layers inside a decoder-only model that process tokens and refine them to predict better.
- The **Causal Mask** is a rule inside the decoder-only model to stop it from looking at future words and only look at the past ones.



They are pieces, but when put together, form a sentence.



Chat-GPT is a decoder-only model



Like multiple editors



No cheating

## Next-Token Prediction

Next-token Prediction is the main task of many decoder-only models like *GPT*. The goal of it is to predict the next word or symbol in a sequence given the context of the previous one.



Like finishing sentences.

The dog sat on the

Token 1 Token 2 Token 3 Token 4 Token 5

Each word is turned into a token

Causal Mask

Decoder Layer 1

Decoder Layer 2

Probability of next token

## Next-Token Prediction

mat

72%

floor

15%

chair

5%

Decoder layers improve context by refining the tokens' meaning.

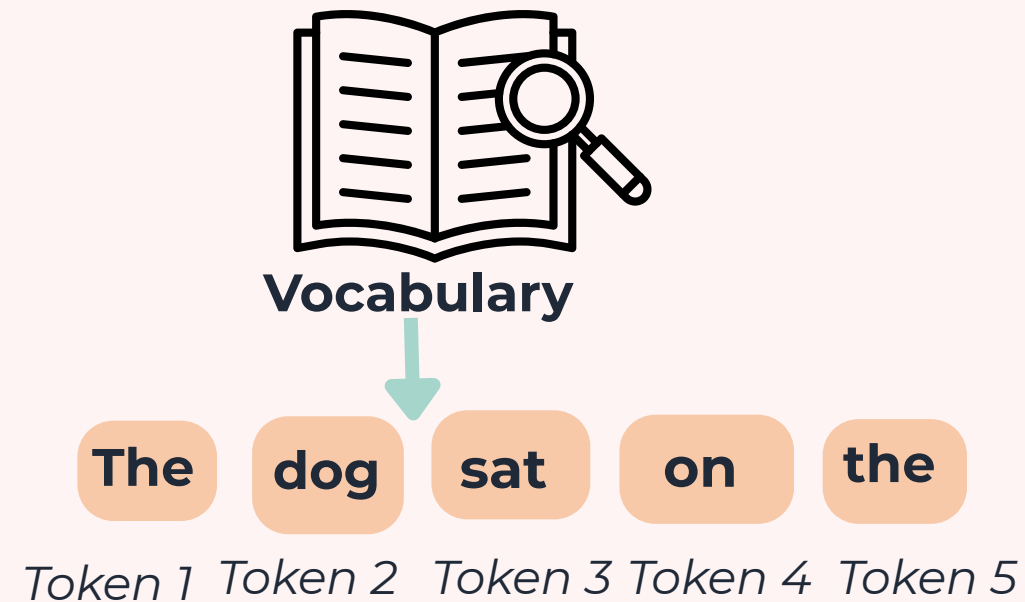
Word is chosen: mat

This process is a simplified version of next-token prediction, in reality the process has dozens of layers of decoders and is trained on many tokens at once. However, the process remains the same.

# Training A Decoder-Only Model

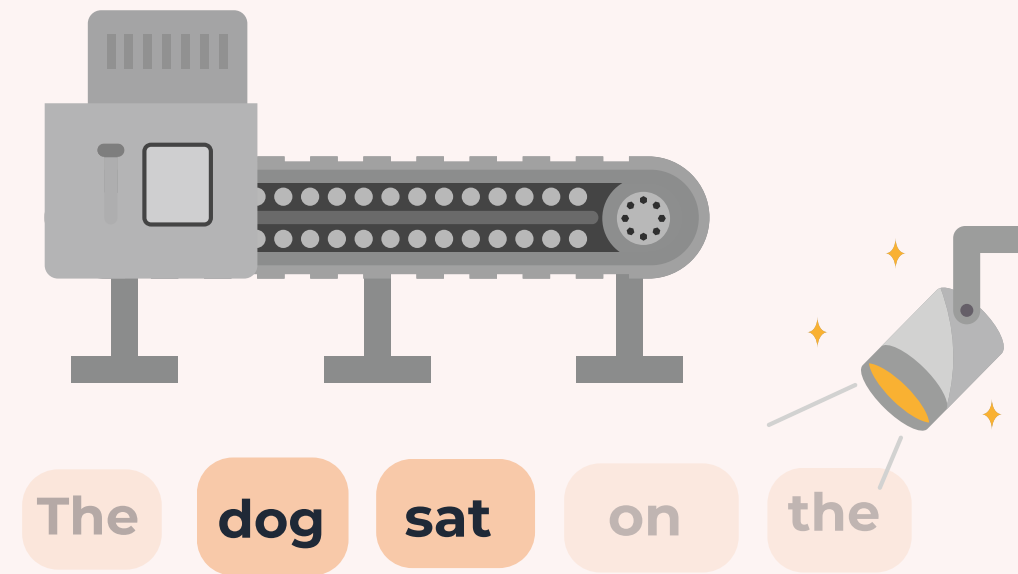
Training – The process of teaching an AI model to learn from data so it can perform a specific task, like predicting the next word in a sentence.

## Tokenization



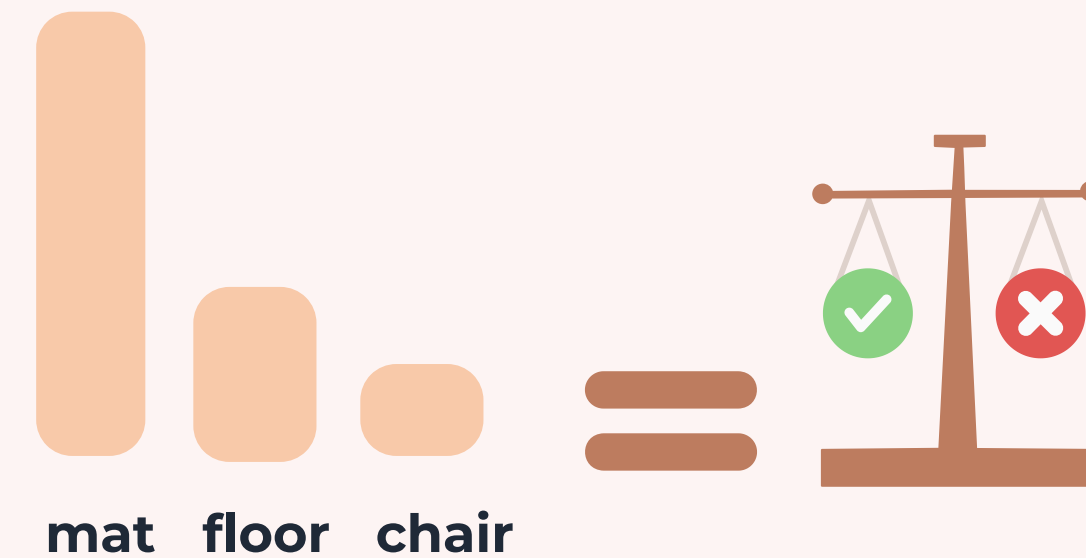
The model learns the patterns of tokens by reading lots of text, building its ability to recognise those tokens later.

## Processing



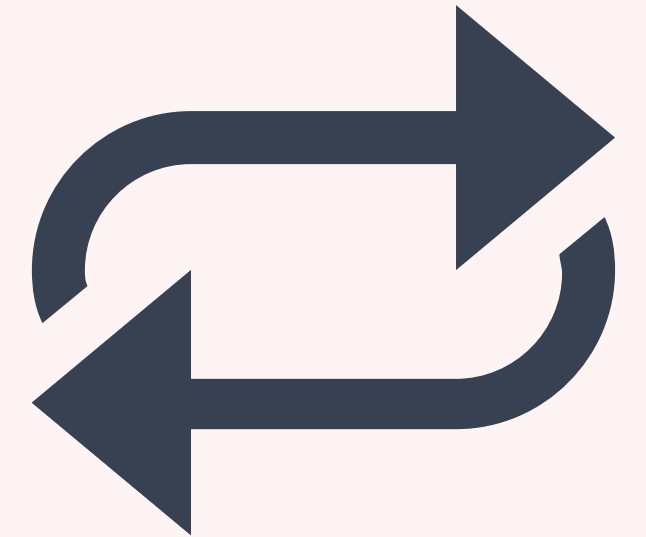
Tokens go into layers → model focuses on important words (self-attention) → rule blocks it from looking ahead (causal mask).

## Prediction & Selection



The model guesses the next token and compares its guess to the correct answer. It learns to make the correct token more likely next time.

## Repeat



This process is repeated billions of times until the model becomes skilled at predicting tokens in many contexts.