What is a token in artificial intelligence?



www.yzokul.com.tr

What is a token in artificial intelligence?



You have probably heard that artificial intelligence language models work with tokens. What is a token? The short answer: a token is a small piece of text that a language model processes. It is the basic unit the model handles when it reads or writes text. This book explains tokens step by step, starting from zero, so you can understand exactly what they are and why they matter.



A language model does not read a whole sentence at once. Instead, it breaks the text into small pieces. Each piece is called a token. The model then processes these tokens one by one. This breaking is necessary because the model works with numbers, and tokens are the pieces that get turned into numbers. So tokens are the first step in how a model handles text.



For simple short words, one whole word can become one token. For example, the English word cat is short and common, so it is often a single token. The model sees cat as one piece. This shows that a token can be a whole word. But this is not always true, as you will see next.



A token is not always a whole word. Long or uncommon words can be split into smaller parts. Each part becomes a token. For example, the word unbelievable might be split into un, believe, and able. Each of these parts is a token. So a token can be a word part, not just a whole word.



Punctuation marks are also tokens. A comma, a period, or a question mark is a piece of text that the model processes. For example, in the sentence Hello, world! the comma and the exclamation mark are each a token. So tokens include punctuation, not just letters and words. This is important to remember.



Consider a long Turkish word: muzikolojik. This word is not common, so it may be split into several pieces. The model might break it into mu, ziko, and lojik. Each piece is a token. You can see that the original word is divided into smaller parts, and each part is processed separately by the model.



Look at the pieces from the previous example: mu, ziko, lojik. None of these pieces is a whole word by itself. They are just fragments. This shows that a token is simply a piece of text, whatever its size. It does not need to be a meaningful word. The model treats each piece as a token.



The model does not work with letters directly. It works with numbers. So each token is turned into a number. This number is called a token ID. The model processes these numbers. For example, the token cat might become the number 1234. The text becomes a list of numbers that the model can calculate with.



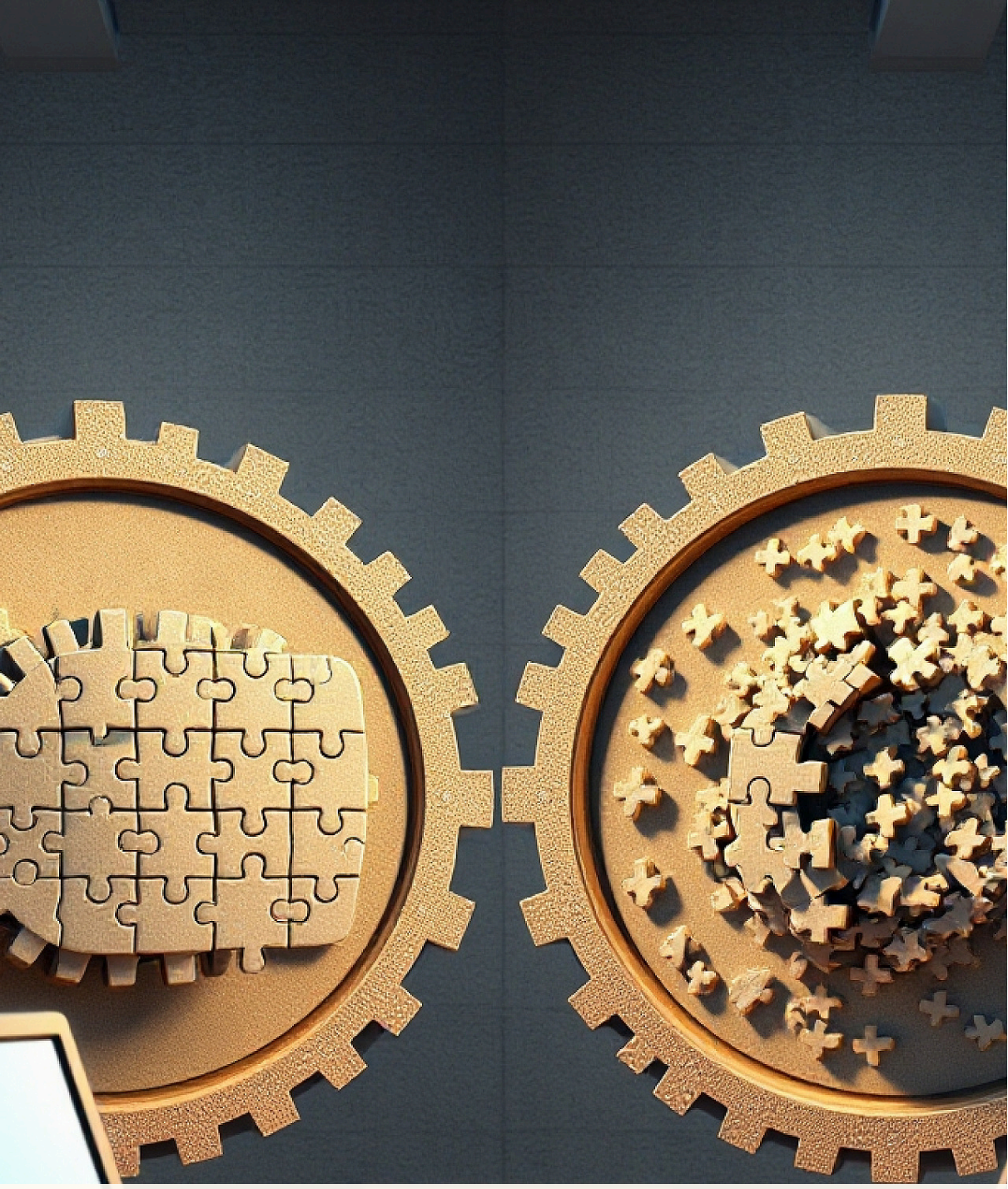
Why numbers? Because the model does math. It uses mathematical operations to predict the next token. Letters and words cannot be used in math directly. So every token gets a number, and the text becomes a sequence of numbers. This is how the model understands and generates text. Numbers are the language of the model.



The number of words and the number of tokens are not the same. One word can be several tokens. For example, the word unbelievable might be three tokens. So if you have a sentence with five words, it might have more than five tokens. Counting words does not give you the token count. You must count tokens separately.



The language used affects how text splits into tokens. Some languages have long words or complex word forms. For example, Turkish words can be very long and may split into many tokens. English words are often shorter and may split into fewer tokens. So the same idea in different languages can give different token counts.



The tokenization method also affects the result. Tokenization is the process of splitting text into tokens. Different methods split text in different ways. One method might split a word into two tokens, while another method might keep it as one token. So the token count depends on the method used.



A common mistake is thinking that a token is always a whole word. But as you have seen, a token can be a word, a word part, or a punctuation mark. For example, the word unbelievable might be three tokens, and a comma is one token. So do not assume token equals word. Tokens are pieces of text.



Another common mistake is counting words and assuming that is the token count. Words and tokens are counted differently. A sentence with ten words might have fifteen tokens. So if you need the token count, you must count tokens, not words. This matters when you use language models, because they have limits on tokens.



Here is a short text: I love AI. If we split it by words, we get three tokens: I, love, AI. But if we split it by characters, we get more tokens: I, space, l, o, v, e, space, A, I. The text is the same, but the pieces are different. This shows the method changes the tokens.



To sum up, a token is a piece of text that a language model processes. It can be a whole word, a part of a word, or a punctuation mark. Each token is turned into a number. Tokens are not the same as words, and the token count depends on the language and the tokenization method. Remember these key points.