# What Are Tokens? From Characters to GPT's Tokenizer

## Discussion on Tokens in Neural Networks

- Early neural networks processed text character by character
- Limited character set (about 100), efficient memory use
- Hard for networks to learn both word formation and meaning from characters
- Switched to word-level input; vocab exploded due to many words, proper nouns
- Rare words often left out, limiting understanding
- Breakthrough: use chunks/fragments of words as tokens
- Tokens can be words, word parts, or common word pairs
- This middle ground keeps vocab manageable, covers rare words
- Efficient, fast learning, aligns with language structure (e.g., word stems)
- Tokens are the first input to language models (token IDs)
- Tokens are not the same as vectors; vectors come later in the network

## Tokenizer Tool

- GPT uses a specific tokenizer to convert text into tokens
- Tool available at platform.OpenAI.com/tokenizer for hands-on exploration
- Useful for understanding how text is split into tokens