NO PREP

Tokens, Parameters, and Why Bigger Isn't Always Better

MORE LESSONS LESS PLANNING · MLLP · FREE TODAY · FREE TOMORROW · FREE FOREVER

Includes: Student reading, case study, differentiated questions, teacher key

Name: Period: Date:

Tokens, Parameters, and Why Bigger Isn't Always Better

Name: Period: Date:

Tokens, Parameters, and Why Bigger Isn't Always Better

The Vocabulary Problem: Breaking Language Into Chunks

When an AI model reads text, it does not process letters or complete words the way a human reader does. It processes units called tokens. A token is a chunk of text that the model's vocabulary system has learned to recognize as a meaningful unit. Tokens are not always whole words. Common short words like 'the,' 'is,' and 'and' are typically single tokens. Longer or rarer words might be split into multiple tokens. The word 'unbelievable,' for example, might be tokenized as 'un,' 'believ,' and 'able.' Different AI models use different tokenization systems. GPT-4 uses a system called byte-pair encoding (BPE), which starts with individual characters and merges the most common pairs into single tokens through a training process. Most large language models have vocabularies of roughly 30,000 to 100,000 unique tokens. Everything the model can understand and generate must be expressible in terms of these tokens.

Why Tokens Matter for Users

Tokens are the unit that determines how much text an LLM can process at once. Every AI model has a context window, which is the maximum number of tokens it can hold in 'working memory' during a conversation. Early GPT models had context windows of about 4,000 tokens (roughly 3,000 words). Recent models have expanded dramatically. Claude's context window can handle over 200,000 tokens. Some specialized models handle millions. Context windows matter practically. If you are asking an AI to summarize a very long document, the document has to fit within the context window, or the model has to process it in chunks and may lose track of earlier information. For conversational AI, the context window determines how far back in a conversation the model can 'remember.' When a conversation exceeds the context window, the model can no longer access the earlier parts, which is why very

long AI conversations sometimes seem to forget earlier context. API pricing for AI tools is also typically based on token counts, both for the input (the text you send) and the output (the response you receive). If you are building a product using an AI API, managing token usage is a significant cost consideration.

Parameters: The Scale of What AI Knows

If tokens are the vocabulary, parameters are the memory. The number of parameters in a neural network model determines how much information it can encode and how complex its internal representations can be. A parameter is simply a number, one of billions of weights in the neural network, that gets adjusted during training. After training, the parameters are fixed, and their values encode everything the model learned. GPT-3 had 175 billion parameters. GPT-4 is estimated to have over a trillion, though OpenAI has not officially confirmed its size. Llama 3, Meta's open-source model, comes in versions ranging from 8 billion to 405 billion parameters. These numbers are staggering but also a bit abstract. What they represent practically is the model's capacity: a model with more parameters can, in theory, learn more nuanced patterns and represent more complex knowledge.

Scaling Laws: More Data and More Parameters Help, Up to a Point

For much of the 2010s and early 2020s, AI researchers observed what they called scaling laws: consistent improvements in model performance as they increased training data, compute, and parameter count. Making a model bigger while training it on more data reliably produced better results across almost every benchmark. This is part of why companies raced to build larger and larger models. The relationship was codified in 2022 by researchers at DeepMind who published what are now called the Chinchilla scaling laws. These showed that many large models were actually undertrained: they were too large for the amount of data they were trained on, and a smaller model trained on more data would outperform them. This shifted the field's thinking away from pure size and toward more efficient use of data and compute. The Chinchilla model, with 70 billion parameters, outperformed GPT-3's 175 billion parameters because it was trained on far more data.

Why Bigger Isn't Always Better

The Chinchilla findings reflect a broader truth: bigger is not always better when it comes to AI models. A very large model trained on insufficient data will have poor

performance. A large model may also require far more computing power to run inferences (generate responses), making it slower and more expensive to operate. A smaller, well-trained model can be faster, cheaper, and sometimes equally accurate for specific tasks. This is why you now see AI companies releasing families of models at different sizes. Anthropic offers Claude Haiku (small and fast), Claude Sonnet (balanced), and Claude Opus (large and powerful). OpenAI offers GPT-4o Mini alongside the full GPT-4o. For many practical tasks like answering a quick question or summarizing a short document, a smaller model does the job just as well as a larger one at a fraction of the cost. Understanding this helps users and developers choose the right tool for the right task.

Efficiency Advances Are Changing Everything

Researchers have also discovered that the efficiency of training and running AI models improves substantially over time, independent of scale. Techniques like quantization (representing numbers with less precision to save memory), pruning (removing unnecessary parameters), and distillation (training a small model to mimic a large one) can dramatically reduce model size without proportional losses in performance. In early 2025, a Chinese AI company called DeepSeek released models that performed comparably to much larger Western models while being far cheaper to train and run, reportedly at a fraction of the cost. This shocked the AI industry and demonstrated that raw scale may be less important than previously thought. The field is moving toward more efficient architectures and training methods, which may mean that future frontier models do not need to be as large or expensive as today's. Mixture of Experts (MoE) architecture represents one of the most significant recent innovations in making large models more efficient. In a standard transformer model, every parameter is activated for every input. In an MoE model, only a subset of the network's parameters, called experts, are activated for any given input. A router component decides which experts to activate based on the specific input. This means an MoE model can have a very large total parameter count but only use a fraction of those parameters per token, making inference faster and cheaper. GPT-4 is widely believed to be an MoE model, though OpenAI has not officially confirmed its architecture. Mixtral, released by French AI company Mistral AI in 2023, is a publicly documented MoE model that achieves strong performance with efficient inference. The energy cost of inference, running a model after training, has become

an increasingly important consideration as AI usage scales. Training costs are one-time or periodic expenses, but inference costs are incurred with every query, every document summary, every code completion. For large models serving millions of users simultaneously, inference costs can exceed training costs over time. This is a major reason why model compression and quantization techniques are receiving so much research attention. Reducing a model's memory footprint through quantization, from 32-bit floating point representations down to 4-bit or even 2-bit, can dramatically reduce inference cost with relatively modest performance degradation, making deployment of powerful models economically feasible at scale.

Key Information: Tokens and Parameters

Detail	Information
Token definition	A chunk of text the model processes as a unit (not always a full word)
Tokenization method	Byte-pair encoding (BPE): merges common character pairs into tokens
Context window	Maximum tokens an LLM can process at once (Claude: 200,000+)
Parameter definition	A numerical weight in the neural network encoding learned knowledge
GPT-3 size	175 billion parameters
Chinchilla finding	Smaller models trained on more data can outperform larger undertrained models
Scaling laws	Performance improves predictably with more data, compute, and parameters
Quantization	Reducing number precision to make models smaller and faster
Distillation	Training a small model to mimic a larger one

Detail	Information
DeepSeek significance	Demonstrated frontier-level performance at dramatically lower training cost

Case Study: Sofia Chooses the Wrong AI Tool

Sofia is interning at a small marketing agency this summer. Her supervisor asks her to summarize a 400-page industry report and extract key statistics for a client presentation. Sofia uses the company's subscription to a large AI model to process the document. She copies the entire report text into the chat and hits enter. The model returns an error: the document is too long for the context window. Sofia is confused. She thought bigger AI tools could handle anything. She splits the document into thirds and processes each section separately, but the final summary misses connections between sections because the model couldn't see the whole document at once. Her supervisor suggests she try a different tool specifically designed for long documents, which uses retrieval-augmented generation to process documents in chunks and retrieve relevant sections as needed. The second approach works much better. 'Token limits and context windows are real constraints,' her supervisor explains. 'You need to match the tool to the task.'

1. Using what you learned about tokens and context windows, explain why the AI tool failed when Sofia pasted the entire 400-page report.

2. What does this case study suggest about how users should approach choosing AI tools for specific tasks?

Multiple Choice

1. What is a token in the context of large language models?

    a) A chunk of text the model processes as a single unit, which may be less than one full word

    b) A labeled data point in the training set that models process sequentially to improve accuracy

    c) A security code required to access an AI system

    d) A digital coin used to pay for AI usage

2. What did the Chinchilla scaling laws reveal about AI model training?

    a) More parameters always lead to faster inference times and reduced computational requirements

    b) The relationship between model size and performance is completely random

    c) Many large models were undertrained and a smaller model with more data could outperform them

    d) Larger models always outperform smaller ones regardless of training data

3. What is model distillation?

    a) Removing harmful content from a model's training data

    b) Increasing the precision of numerical values stored in a model

    c) A process for merging two separate AI models into one

    d) Training a small model to mimic the behavior of a large model

Matching

Write the letter of the correct definition on the line next to each term.

Term	Definition
1. ___ Token	A) A technique that reduces number precision to make models smaller and faster
2. ___ Context window	B) A chunk of text that an LLM processes as a single unit, often less than a full word
3. ___ Parameter	C) Research showing that smaller models trained on more data can outperform larger undertrained ones
4. ___ Chinchilla scaling laws	D) A numerical weight in a neural network that encodes learned patterns
5. ___ Quantization	E) The maximum amount of text an LLM can hold in working memory at once

Fill-in-the-Blank

1. The maximum number of tokens an AI model can process in a single conversation is called its _____ window.

2. In 2025, the Chinese AI company _____ released models that performed comparably to much larger Western models at a fraction of the cost.

3. _____ laws describe the consistent improvements in AI performance as training data, compute, and parameter count increase.

Short Answer

Answer each question in 2-4 complete sentences.

1. Why do AI companies release model families at different sizes rather than just one flagship model?

2. Explain why a 70-billion-parameter model could outperform a 175-billion-parameter model, as the Chinchilla research demonstrated.

3. Why does the context window size matter for practical AI use, and what problems arise when users work with documents that exceed it?