Understanding (AI) Parameters, Tokens, and What You Can Control

medium.com/@varada/understanding-parameters-tokens-and-what-you-can-control-794ccce52a28

Varada July 25, 2025

A guide to the building blocks of AI that every user should understand

Have you ever wondered what makes ChatGPT “smarter” than other AI models? Or why do some AI responses cost more than others? The answer lies in understanding two fundamental concepts that power every AI interaction: **parameters** and **tokens**.

Most people confuse these terms or assume they’re the same thing. They’re not. Understanding the difference is crucial for anyone wanting to use AI effectively, whether you’re a curious beginner or building the next AI-powered startup.

> Let me start with a simple truth: **You cannot send parameters to an AI model.**

I know this might sound confusing, especially when you see API documentation talking about “parameters.” But here’s the thing — there are two completely different types of “parameters” in the AI world, and mixing them up is like confusing the engine of a car with the radio settings.

What Are AI Model Parameters? (The Engine)

> Think of **model parameters** as the AI’s brain — the accumulated knowledge and intelligence built into the system. These are the numerical weights and connections that determine the model’s intelligence.

The Numbers Game

Here’s how the major AI models stack up in terms of parameters:

- : ~1.76 trillion parameters (rumored)
- : ~200 billion parameters
- : ~175 billion parameters
- : ~8 billion parameters

But here’s the kicker: They’re baked into the model during training, like the knowledge in a professor’s brain after decades of study.

Why Parameter Count Matters (And Why It Doesn’t)

More parameters generally mean more capability, but it’s not that simple. GPT-4o mini, with just 8 billion parameters, often outperforms much larger predecessors through better training techniques. It’s like comparing a well-educated specialist to someone who just memorized an encyclopedia.

The industry is also becoming secretive about these numbers. Google won’t tell you how many parameters Gemini has. Anthropic keeps Claude’s specs under wraps. Why? Because the AI arms race has shifted from “bigger is better” to “smarter is better.”

What Are Tokens? (The Fuel)

If parameters are the brain, **tokens** are the language — the way text gets broken down for AI to understand.

Tokenization in Action

When you type “Hello, how are you?” the AI doesn’t see whole words. It sees tokens:

```
 = 1 token = 1 token = 1 token = 1 token = 1 token = 1 token
```

Context Windows: The Memory Limit

Every AI model has a **context window** — the maximum number of tokens it can process at once:

- : 200,000 tokens (~150,000 words)
- : 8,192 to 128,000 tokens
- : Up to 2 million tokens

This is like short-term memory. A model might have trillions of parameters (vast knowledge), but only handle 200,000 tokens at once (limited working memory).

What You Can Control: API Parameters

Here’s where the confusion happens. When developers discuss “sending parameters” to an AI, they refer to **API parameters** — configuration settings that control how the model behaves.

The Settings You Can Adjust

These are like adjusting the bass and treble on your stereo — you're not changing the fundamental music (model parameters), just how it's presented.

Temperature: The Creativity Dial

Temperature is the most important setting you'll encounter:

- : Deterministic, always gives the same answer
- : Balanced Creativity (recommended for most uses)
- : Very creative, sometimes unpredictable
- : Maximum randomness (often incoherent)

Max Tokens: The Length Limiter

Max tokens control response length:

- : ~37 words (tweet-length)
- : ~375 words (short paragraph)
- : ~1500 words (full article)

Real-World Applications: How to Use This Knowledge

For Developers

When building AI applications, understand your token economics:

```
 openairesponse = openai.ChatCompletion.create(model=, messages=[{: , : },{: , :
}],temperature=, max_tokens=, top_p= )
```

For Business Users

Choose models based on your needs, not just parameter count:

- : GPT-4o mini (8B parameters, very fast)
- : Claude 4 Opus (reasoning-focused)
- : Claude (200K context window)
- : Perplexity (multi-model access)

For Content Creators

Optimize your prompts for token efficiency:

- : “I would appreciate it if you could kindly help me write a blog post about artificial intelligence and machine learning technologies and their applications in modern business environments.”
- : “Write a blog post about AI/ML applications in business.”

The Hidden Economics of AI

Understanding tokens isn’t just technical — it’s financial. API pricing is typically per token:

- : $30–60 per million tokens
- : $3–15 per million tokens
- : Often, the most cost-effective

A 1,000-word article might cost $0.05–0.15 to generate, but a 10,000-word document could cost $0.50–1.50. Those tokens add up.

The Future of AI: Beyond Parameter Counts

The industry is evolving beyond the “bigger is better” mentality:

- : Models like GPT-4o mini prove that smaller can be better
- : Claude 4 and Gemini 2.5 emphasize thinking over size
- : Using multiple specialized models instead of one giant one
- : Processing text, images, audio, and video together

Practical Takeaways

For AI Users:

- = Model capability (you can’t change)
- = Input/output text (you control the amount)
- = Behavior settings (you fully control)

For Developers:

- Monitor token usage for cost optimization
- Use temperature strategically for different tasks
- Choose models based on the context window needs
- Consider vector databases for AI applications

For Business Leaders:

- Model size doesn’t always correlate with performance

- Token costs can impact scaling decisions
- Context windows determine document processing capabilities
- Multi-model strategies often outperform single-model approaches

The Bottom Line

Understanding parameters and tokens isn’t just technical trivia — it’s practical knowledge that can help you:

- on AI API costs
- for your specific needs
- through proper configuration
- with efficient token usage

The AI revolution is just beginning, and the winners will be those who understand not just what these tools can do but how they work under the hood.

Remember: You’re not sending parameters to the AI — you’re configuring how a pre-trained model with billions or trillions of parameters processes your specific input tokens. Master this distinction, and you’ll be ahead of 90% of AI users.

What’s your experience with AI parameters and tokens? Have you optimized your AI usage for cost or performance? Share your insights in the comments below.