AI Tokens: Speed & Cost, Explained

The tiny unit behind every AI reply — how it's counted, why it costs what it costs, and why some answers feel instant while others crawl.

tk $ ⚡

What Exactly Is a Token?

AI models don't read in full words — they break text into small chunks called **tokens**. A token might be a whole word, a piece of a word, or a punctuation mark.

1,000 tokens ≈ **750** words

◆ SEE IT IN ACTION

"Tokenization is fun!" → split into 5 tokens. Longer or unusual words get chopped into more pieces than short, common ones.

Your Question Costs Tokens Too

Before the AI answers anything, it turns your message — the prompt — into **input tokens**. The longer or more detailed your prompt, the more tokens it takes to process it.

💬 **A short question**

"What's the capital of France?" → about 8 tokens

📄 **A pasted document**

One page of text → roughly 400–600 tokens

🖼️ **Images & files**

Also converted into tokens — often more than you'd expect

The Answer Is Tokens, Too

Every word the AI writes back to you is an **output token**. Generating them takes real computing work — which is why output usually costs more than input.

📥 Input

What you send in. The AI just has to "read" it — relatively cheap and fast.

Reading = lighter work

📤 Output

What the AI writes back. It has to "think" and generate each token — heavier work.

Writing = pricier, per token

Rule of thumb: **output tokens often cost 2–5× more** than input tokens on most AI platforms.

How Tokens Turn Into $

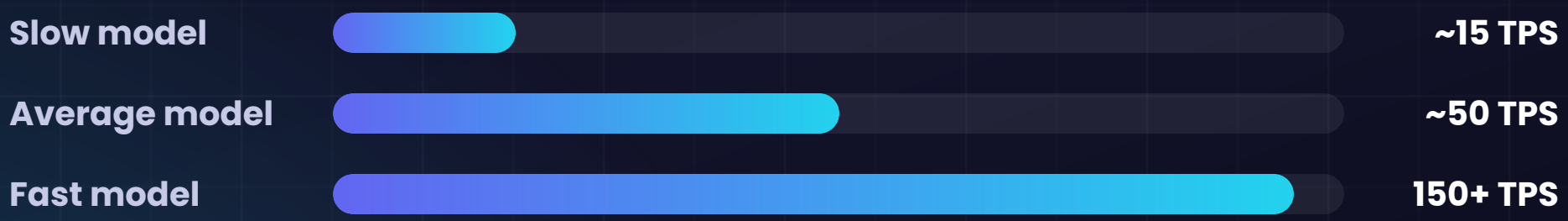
Providers charge per 1,000 or 1,000,000 tokens — with separate rates for input and output. Here's an illustrative example:

TYPE	EXAMPLE RATE	PER
Input tokens	**$0.15**	1M tokens
Output tokens	**$0.60**	1M tokens
A 500-word chat reply	**~$0.0004**	one reply

Total cost = **(input tokens × input rate)** + **(output tokens × output rate)**

Tokens Per Second = Typing Speed

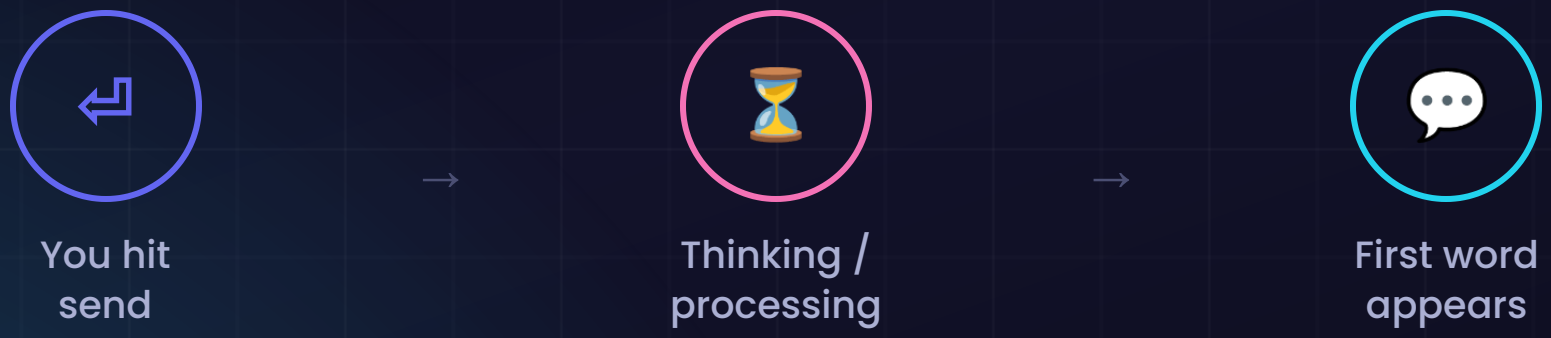
Once the AI starts replying, it generates one token at a time. That rate is called **TPS (tokens per second)** — think of it as how fast someone types out the answer live.



Higher TPS = the reply fills your screen faster, word by word.

The Pause Before It Starts Typing

Latency, or "time to first token" (TTFT), is how long you wait after hitting send — before the very first word shows up. A model can be a fast typer and still have a slow start.



That middle gap is what makes a chat feel instant — or laggy.

What Makes AI Faster or Slower?

Model size

Bigger, smarter models usually think slower than smaller ones

Server demand

Peak hours mean more people sharing the same computing power

Prompt & answer length

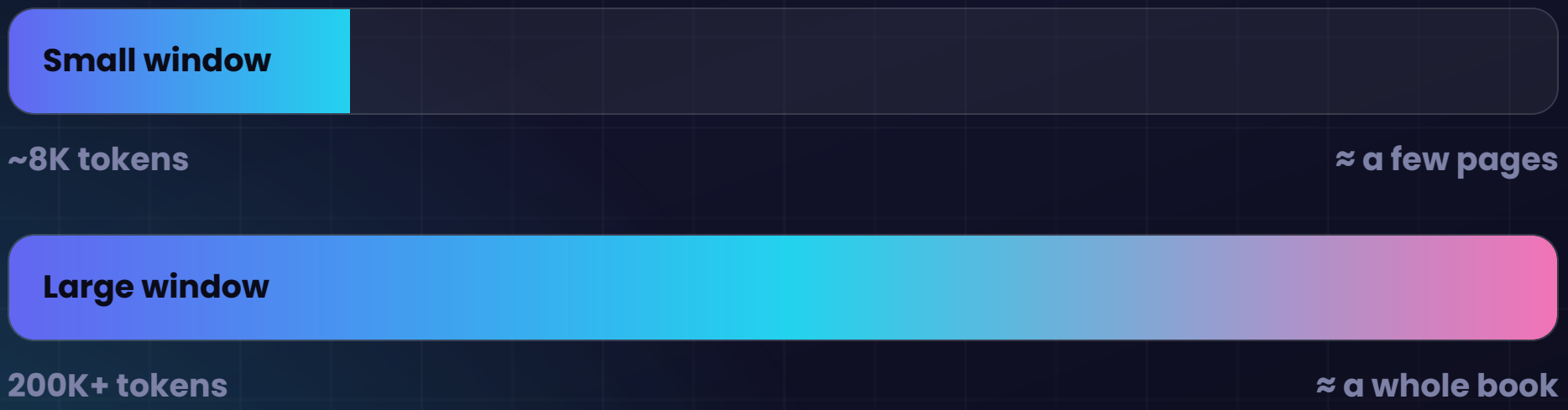
Longer input takes longer to read; longer output takes longer to finish

Hardware & optimization

The chips and software tricks running the model behind the scenes

Context Window = Its Short-Term Memory

The **context window** is the total number of tokens (your messages + its replies) the AI can "see" at once. Go past it, and the oldest parts of the conversation start dropping off.



Why You Sometimes Hit a Wall

Providers cap how many tokens or messages you can send per minute, hour, or day — this is a **usage limit**. Hitting "limit reached" is about your plan's quota, not something you did wrong.

FREE	PAID / PRO	ENTERPRISE
Low	**Higher**	**Highest**
limits / day	limits / day	custom quotas

A Picture Isn't Worth 1,000 Tokens

The old saying doesn't hold up in AI math. Images get sliced into visual patches, and every patch becomes tokens — often **more** than a full page of text.

📝 1,000 words

Plain text compresses well — roughly the same amount in, roughly that many tokens out.

≈ 1,300 tokens

🖼️ 1 photo

A single image can cost as much as several pages of text, depending on size and detail.

≈ 1,000–1,600+ tokens

Tip: **resize or crop images** before uploading — fewer pixels in means fewer tokens (and less cost) out.

Creating an Image Is Priced Differently

That was about images you **upload**. Asking the AI to **generate** one works differently — it's usually billed per image, by resolution or quality tier, not by counting words.

QUALITY / SIZE	TYPICAL COST	≈ TOKEN EQUIVALENT
Small / low quality	$	≈ 300–700 tokens
Standard (e.g. 1024×1024)	$$	≈ 1,000–1,500 tokens
High-res / HD quality	$$$	≈ 4,000+ tokens

Bigger, more detailed, higher-resolution images mean more generation work — and a bigger bill.

Behind Every Reply: Power & Water

Every prompt runs on a data center somewhere — drawing electricity to power the chips, and often water to cool them down. More tokens processed means more energy spent.

⚡ Electricity

Powers the servers doing the "thinking" — longer prompts and answers mean more compute time.

💧 Water

Cools the data centers so chips don't overheat under heavy, nonstop use.

◆ PROMPT GREENER

BE CONCISE

Shorter prompts & answers need less compute

REUSE, DON'T REPEAT

Save good answers instead of re-generating them

RIGHT-SIZE THE MODEL

Use smaller/faster models for simple tasks

TRIM THE EXTRAS

Skip pasting files or images you don't really need

Your Token Cheat Sheet

TOKEN

Small chunk of text — ~1,000 tokens ≈ 750 words

INPUT

Tokens used by what you send in

OUTPUT

Tokens used by what the AI writes back

COST

Input tokens + output tokens, priced separately

TPS

Tokens per second — how fast it "types"

TTFT

Time to first token — how long you wait to start

CONTEXT WINDOW

How much it can "remember" at once

USAGE LIMITS

Caps on tokens/messages per plan

Now You Speak Fluent Token.

Next time you see a token-based price tag on an AI tool, you'll know exactly what's driving the speed — and the bill.

🔁 Repost this to help a colleague understand AI costs

💬 Comment your biggest "wait, tokens are what?" moment

🔔 Follow for more AI, explained simply