AI, EXPLAINED SIMPLY

AI Tokens: The Cheat Sheet

Speed, cost & sustainability — everything from the 15-slide explainer, on one page.

TOKENS 101

1,000 tokens ≈ **750** words

AI splits text into small chunks — whole words, pieces of words, or punctuation. Longer / unusual words use more tokens per word.

📥 Input
What you send in. Cheaper to process.

📤 Output
What it writes back. Costs more per token.

PRICING

Providers charge per 1K–1M tokens, with separate input / output rates.

TYPE	EXAMPLE RATE
Input tokens	$0.15 / 1M
Output tokens	$0.60 / 1M

Cost = **(input × rate)** + **(output × rate)**

SPEED & LATENCY

TPS = tokens per second — how fast it "types".

Slow	~15 TPS
Average	~50 TPS
Fast	150+ TPS

TTFT (latency) = the wait before the first word appears — separate from typing speed.

WHAT AFFECTS SPEED

Model size
Bigger models often think slower

Server demand
Peak hours = shared capacity

Prompt / answer length
More tokens = more time

Hardware & optimization
The chips & tricks behind it

MEMORY & LIMITS

Context window = tokens the AI can "see" at once.

~8K
Small window ≈ a few pages

200K+
Large window ≈ a whole book

Usage limits cap tokens/messages per plan:

FREE	PAID	ENTERPRISE
Low	Higher	Highest

IMAGES: UPLOAD VS. GENERATE

📎 Uploading
A photo you attach gets sliced into visual patches — often costlier than a page of text.
≈ 1,000–1,600+ tokens

🎨 Generating
Priced per image by resolution / quality tier, not word count.
≈ 300 – 4,000+ tokens

GENERATE QUALITY	≈ TOKEN EQUIV.
Small / low	300–700
Standard (1024²)	1,000–1,500
HD / high-res	4,000+

GREEN AI: POWER & WATER

⚡ Electricity
Powers the servers "thinking" — more tokens = more compute time.

💧 Water
Cools data centers so chips don't overheat under heavy use.

Prompt greener:

BE CONCISE
Shorter prompts & answers

REUSE, DON'T REPEAT
Save answers, don't re-run

RIGHT-SIZE THE MODEL
Small model for simple tasks

TRIM THE EXTRAS
Skip files/images you don't need