

Unpacking AI Tokenomics

AI models don't read words or interpret images the way humans do. They process input and output in small units called tokens, the fundamental currency of AI computation and the building blocks of AI communication. Tokenomics provides a framework for measuring and managing these costs, shaping research budgets and resource allocation. This month, we unpack this framework and share practical ways to keep costs less unpredictable.

What is a Token?

A token is a basic unit of data that an AI model processes or generates. Tokens can represent pieces of text, visual information from images, audio, video, or other modalities.

Think of it as the anatomic unit that models read and generate, and the basis for how providers measure and price usage.

For example, the sentence below has 9 words, 45 characters, but can you guess how many tokens?

Do all AI models count tokens the same way?

If you guessed 10, you could be right, for some models.

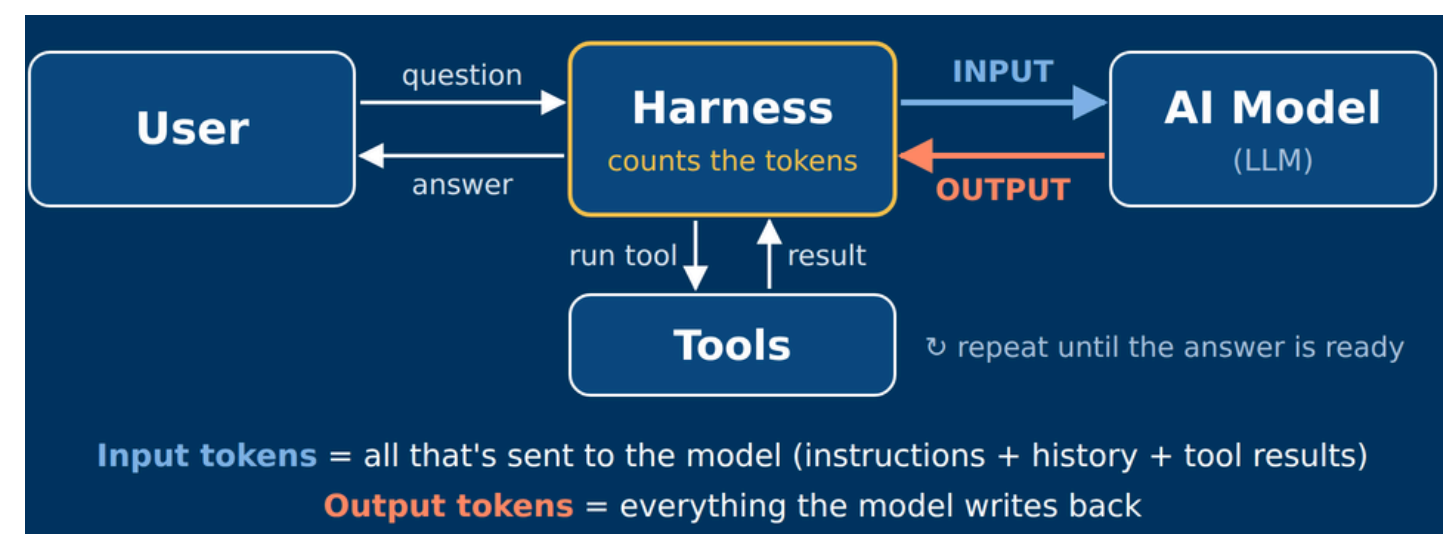
Each AI model may use a different tokenizer, so identical content can be broken into different numbers of tokens.

This applies across modalities, including text, images, audio, and video.

How Are Tokens Counted?

Every model interaction consumes tokens. AI utilizes a harness (software layer) that orchestrates a model's interactions with its context, tools, data, and execution environment.

The harness controls what the model receives, when it is called, what tools it can use, and how results flow back into the process. Each interaction consumes input and output tokens, turning token usage into compute and, ultimately, cost.



With agents, the loop runs longer than with chatbots, calls more tools, and may start sub-agents, so input tokens can pile up much faster.

Which Factors Affect Pricing?

Inference costs are calculated as a rate (\$/million tokens). AI providers typically price input and output tokens separately, with output tokens often costing more than input tokens.

Pricing varies by model, capability, and provider. Use services like tokenplans.dev to compare AI pricing across providers.

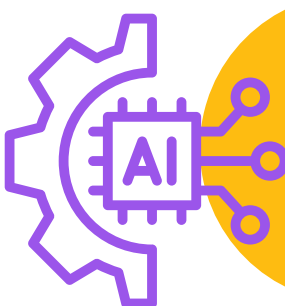
Tips to Reduce AI Costs

- Keep chats short and focused: Start a new chat for each task to avoid resending unnecessary history.
- Send only what's needed: Trim, clean, or summarize documents before sharing them.
- Use the smallest model that works: Turn off reasoning for simple tasks.
- Limit output: Ask for concise, structured responses since output tokens often cost more.
- Cut hidden overhead: Shorten system prompts, adjust configuration settings, or use plugins to disable unused tools and minimize unnecessary token usage.
- Test small, then scale. Use planning mode and pilot prompts on a small sample before scaling.

An Example

Suppose you want to compare methods across 5 papers:

	✗ Costly	✓ Efficient
Focus chats	One long thread	New chat per task
Send less	5 full PDFs	Methods only
Right-size model	Largest model	Small, fast model
Limit output	"Explain in detail"	"5 bullets"
Cut overhead	All tools on	Unused tools off
Test, then scale	Run all 500	Pilot 5, then batch
	✗ ~200k	✓ ~5k tokens per question



Looking to build an AI-powered research workflow? Reach out to dreamlab@library.ucsb.edu to explore options.