Understanding “Tokens” and Natural Language Processing (NLP)

medium.com/@varada/understanding-tokens-and-natural-language-processing-nlp-1ee1caa10083

Varada

July 16, 2025

As you explore AI Studio and its API, you’ll frequently encounter the term **“token cost.”** This is how AI service providers primarily bill for the use of their models, and it’s a concept deeply rooted in **Natural Language Processing (NLP)**.

Natural Language Processing (NLP) is a crucial subfield of Artificial Intelligence that focuses on enabling computers to understand, interpret, and generate human language. Think of it as the bridge between human communication and computer logic.

Before an AI model (especially a large language model, or LLM, such as Gemini) can process your text, NLP techniques first break down that raw text into smaller, numerical units called **tokens**.

- Tokens aren’t always complete words; they can be parts of words, individual characters (like punctuation), or even spaces. For example, the word “unbelievable” might be broken into tokens like “un”, “believe”, and “able”.
- Once text is tokenized, each token is assigned a unique numerical ID. These numbers are what the AI model actually “understands” and processes.
- This tokenization process is crucial for efficient model training and inference, as it standardizes the input and effectively manages vocabulary size. Approximately .

How Token Cost Works in AI Billing

You’re typically charged based on two types of tokens:

1. You pay for the tokens in the text you send to the AI model (your prompt, any context you provide, or previous turns in a conversation history). Longer, more detailed prompts will incur higher costs.
2. You also pay for the tokens in the response generated by the AI model. Longer, more verbose AI responses will incur higher output token costs.

Why Token Cost? It provides a granular, precise way to measure and bill for the actual computational work the AI model performs. Processing more tokens directly correlates with the amount of computing power and time required from the AI model, making it a fair and scalable billing metric.

Optimization Tip: To manage costs, write concise prompts that guide the model to generate only the necessary output, and choose a model of the right size for your task.

What’s been your experience? Share in the comments!