

Tokens & Cost

Computing

How usage is billed, and where the money actually goes.

Providers bill by token, usually at different rates for input and output. A token is roughly four characters of English, so a page of text is about 500 tokens.

1,000 TOKENS
~750 words

1 TOKEN
~4 chars

A PAGE
~500 tokens

- ~750 words $\approx$ 1,000 tokens in English.
- Output tokens usually cost several times input tokens.
- The whole conversation is resent on every turn.
- A long system prompt is paid for on every request.
- Caching can cut the cost of a repeated prefix substantially.
- Other languages tokenise less efficiently than English.

LOOK FOR IT

A 2,000-token system prompt over 10,000 calls is 20 million input tokens, before any user text.

Trim the system prompt before optimising anything else. It is the one cost multiplied by every request you serve.