
Re: API List and Pricing

From Harshit Gupta <Harshit.Gupta@anu.edu.au>

Date Tue 25/11/2025 3:37 PM

To Shunichi Ishihara <Shunichi.Ishihara@anu.edu.au>

Dear Shun,

Thank you for your email and for looking through the list.

Below is a brief explanation for why Llama, Copilot and Perplexity were not included in the initial comparison:

1. **Llama**

The Llama API is currently in a restricted, waitlist-based rollout, and in practice access is primarily geared towards US-based developers. Because of this limited and uncertain availability, I focused the current analysis on models with more generally accessible APIs that we can rely on for reproducible experiments without additional access constraints.

2. **Copilot (Microsoft 365 Copilot / Copilot Chat API)**

Copilot's Chat API is relatively new, in a preview stage and primarily designed for interactive, grounded enterprise use rather than controlled offline generation. For our setting, two aspects are less suitable:

- The API tightly couples responses to enterprise web search grounding, which makes it harder to control prompts and topics in a reproducible way across many generations.
- It is not optimised for long, batch-style experiments and can be prone to timeouts for longer-running tasks.

Given these constraints, Copilot seemed less appropriate as a primary model for systematic style imitation experiments.

3. **Perplexity**

Perplexity was omitted due to oversight on my part. I have now looked at its pricing (e.g. Sonar / Sonar Pro / Sonar Reasoning tiers: roughly \$1–3 per million input tokens, \$1–15 per million output tokens depending on the tier) and can add it to the cost table if you would like. Functionally, Perplexity is heavily optimised for retrieval-augmented web search and sits on top of underlying foundation models, so for our purposes it may be more natural to evaluate those base models directly, but I am happy to include Perplexity explicitly in the comparison.