Quick Start Models & Pricing

Models & Pricing

The prices listed below are in unites of per 1M tokens. A token, the smallest unit of text that the model recognizes, can be a word, a number, or even a punctuation mark. We will bill based on the total number of input and output tokens by the model.

Pricing Details

USD CNY

MODEL[1]	CONTEXT LENGTH	MAX OUTPUT TOKENS[2]	INPUT PRICE (CACHE HIT)[3]	INPUT PRICE (CACHE MISS)	OUTPUT PRICE
deepseek-chat	64K	8K	~~$0.07/ 1M tokens~~[4] $0.014 / 1M tokens	~~$0.27/ 1M tokens~~[4] $0.14 / 1M tokens	~~$1.10/ 1M tokens~~[4] $0.28/1M tokens

- (1) **The `deepseek-chat` model has been upgraded to DeepSeek-V3.**
- (2) If max_tokens is not specified, the default maximum output length is 4K. Please adjust `max_tokens` to support longer outputs.
- (3) Please check this article for the details of Context Caching.
- (4) The form shows the the original price and the discounted price. **From now until 2025-02-08 16:00 (UTC), all users can enjoy the discounted prices of DeepSeek API.** After that, it will recover to full price.

Deduction Rules

The expense = number of tokens × price. The corresponding fees will be directly deducted from your topped-up balance or granted balance, with a preference for using the granted balance first when both balances are available.

Product prices may vary and DeepSeek reserves the right to adjust them. We recommend topping up based on your actual usage and regularly checking this page for the most recent pricing information.