

# DOORM AI API Docs

1. Models & Pricing; .....	2
2. Token & Token Usage; .....	4
3. Rate Limit; .....	5

If you have any questions, please contact email: [service@doorm.ai](mailto:service@doorm.ai)

# 1. 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

MODEL <sup>(1)</sup>		doormai-chat	doormai-reasoner	doormai-special
CONTEXT LENGTH		64K	64K	64K
MAX COT TOKENS <sup>(2)</sup>		-	32K	32K
MAX OUTPUT TOKENS <sup>(3)</sup>		8K	8K	8K
STANDARD PRICE	1M TOKENS INPUT (CACHE HIT) <sup>(4)</sup>	\$0.07	\$0.14	\$0.28
	1M TOKENS INPUT (CACHE MISS)	\$0.27	\$0.55	\$1.10
	1M TOKENS OUTPUT <sup>(5)</sup>	\$1.10	\$2.19	\$4.38
non-profit organization (After review)	1M TOKENS INPUT (CACHE HIT)	Free	Free	Free
	1M TOKENS INPUT (CACHE MISS)	Free	Free	Free
	1M TOKENS OUTPUT	Free	Free	Free

(1) The doormai-chat model points to **text model**. The doormai-reasoner model points to **image model**.The doormai-special model points to **Buddhism Medicine model**, etc.

(2) **CoT (Chain of Thought)** is the reasoning content doormai-reasoner gives before output the final answer.

(3) If max\_tokens is not specified, the default maximum output length is 4K.Please adjust max\_tokens to support longer outputs.

(4) The details of Context Caching.

(5) The output token count of doormai-reasoner includes all tokens from CoT and the final answer, and they are priced equally.

# 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 DOORMAI 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.

## 2. Token & Token Usage;

Tokens are the basic units used by models to represent natural language text, and also the units we use for billing. They can be intuitively understood as 'characters' or 'words'. Typically, a Chinese word, an English word, a number, or a symbol is counted as a token.

Generally, the conversion ratio between tokens in the model and the number of characters is approximately as following:

1 English character    0.3 token.

1 Chinese character    0.6 token.

However, due to the different tokenization methods used by different models, the conversion ratios can vary. The actual number of tokens processed each time is based on the model's return, which you can view from the usage results.

### 3. Rate Limit;

DOORMAI API does NOT constrain user's rate limit. We will try our best to serve every request.

However, please note that when our servers are under high traffic pressure, your requests may take some time to receive a response from the server. During this period, your HTTP request will remain connected, and you may continuously receive contents in the following formats:

Non-streaming requests: Continuously return empty lines  
Streaming requests: Continuously return SSE keep-alive comments (: keep-alive)

These contents do not affect the parsing of the JSON body by the OpenAI SDK. If you are parsing the HTTP responses yourself, please ensure to handle these empty lines or comments appropriately.

If the request is still not completed after 30 minutes, the server will close the connection.