Pricing and rate limits

The token limits include the input + output tokens

Model Tier Price per 1M tokens RPM [1] RPD [2] TPM [3] TPD [4] Llama 8B ”Free” - 60 28,800 40k 5M Llama 8B ”Infinite” 0.08/M ∞ ∞ ∞ ∞

Llama 70B ”Free” - 60 28,800 40k 5M Llama 70B ”Infinite” 0.79/M ∞ ∞ ∞ ∞

Llama 405B ”Free” - 12 6,000 40k 1M Llama 405B ”Infinite” 3.5/M ∞ ∞ ∞ ∞ (1) RPM: Request Per Minute (2) RPD: Request Per Day (3) TPM: Tokens Per Minute (4) TPD: Tokens Per Day

To access “Infinite” tier, pleaseapply here

Rate limit headers

We set the followingx-ratelimit headers to inform you on current rate limits applicable to you.

The following headers are set (values are illustrative):

Header Value Notes retry-after 2 Seconds to wait until retrying* x-ratelimit-limit-requests 28800 Requests per day allowed x-ratelimit-limit-tokens 40000 Tokens per minute allowed x-ratelimit-remaining-requests 123 Requests remaining for the day x-ratelimit-remaining-tokens 1337 Tokens remaining for this minute x-ratelimit-reset-requests 1337s Seconds until the daily rate limit resets x-ratelimit-reset-tokens 1s Seconds until the minute based token limit resets * Theretry-after header is only returned if the response status code is 429 and the request was rate limited