**Mistral Api Inference**

**1-Requests per Second**

**Token Limits**

|  |  |  |
| --- | --- | --- |
| **Model** | **Tokens per Minute** | **Tokens per Month** |
| **Open Mistral Models** |  |  |
| open-mistral-7b | 500,000 | 1,000,000,000 |
| open-mixtral-8x7b | 500,000 | 1,000,000,000 |
| open-mixtral-8x22b | 500,000 | 1,000,000,000 |
| open-mistral-nemo | 500,000 | 1,000,000,000 |
| **Mistral Small Models** |  |  |
| mistral-small-2402 | 500,000 | 1,000,000,000 |
| mistral-small-2409 | 500,000 | 1,000,000,000 |
| mistral-small-2501 | 500,000 | 1,000,000,000 |
| **Mistral Medium Model** |  |  |
| mistral-medium | 500,000 | 1,000,000,000 |
| **Mistral Large Models** |  |  |
| mistral-large-2402 | 500,000 | 1,000,000,000 |
| mistral-large-2407 | 500,000 | 1,000,000,000 |
| mistral-large-2411 | 500,000 | 1,000,000,000 |
| **Mistral Saba Model** |  |  |
| mistral-saba-2502 | 500,000 | 1,000,000,000 |
| **Mistral Embedding Model** |  |  |
| mistral-embed | 20,000,000 | 200,000,000,000 |
| **Codestral Models** |  |  |
| codestral-2405 | 500,000 | 1,000,000,000 |
| codestral-2501 | 500,000 | 1,000,000,000 |
| codestral-mamba-2407 | 500,000 | 1,000,000,000 |
| **Pixtral Models** |  |  |
| pixtral-12b-2409 | 500,000 | 1,000,000,000 |
| pixtral-large-2411 | 500,000 | 1,000,000,000 |
| **Ministral Models** |  |  |
| ministral-3b-2410 | 500,000 | 1,000,000,000 |
| ministral-8b-2410 | 500,000 | 1,000,000,000 |
| **Mistral Moderation Model** |  |  |
| mistral-moderation-2411 | 500,000 | 1,000,000,000 |