Groq Rate limits

Rate limits are measured in:

* ****RPM:**** Requests per minute
* ****RPD:**** Requests per day
* ****TPM:**** Tokens per minute
* ****TPD:**** Tokens per day

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| MODEL ID | RPM | RPD | TPM | TPD | ASH | ASD |
| deepseek-r1-distill-llama-70b | 30 | 1,000 | 6,000 | - | - | - |
| llama-3.3-70b-versatile | 30 | 1,000 | 6,000 | 100,000 | - | - |
| llama-3.3-70b-specdec | 30 | 1,000 | 6,000 | 100,000 | - | - |
| llama-3.2-1b-preview | 30 | 7,000 | 7,000 | 500,000 | - | - |
| llama-3.2-3b-preview | 30 | 7,000 | 7,000 | 500,000 | - | - |
| llama-3.1-8b-instant | 30 | 14,400 | 6,000 | 500,000 | - | - |
| llama3-70b-8192 | 30 | 14,400 | 6,000 | 500,000 | - | - |
| llama3-8b-8192 | 30 | 14,400 | 6,000 | 500,000 | - | - |
| llama-guard-3-8b | 30 | 14,400 | 15,000 | 500,000 | - | - |
| mixtral-8x7b-32768 | 30 | 14,400 | 5,000 | 500,000 | - | - |
| gemma2-9b-it | 30 | 14,400 | 15,000 | 500,000 | - | - |
| llama-3.2-11b-vision-preview | 30 | 7,000 | 7,000 | 500,000 | - | - |
| llama-3.2-90b-vision-preview | 15 | 3,500 | 7,000 | 250,000 | - | - |

Rate limits are measured in:

* ****RPM:**** Requests per minute
* ****RPD:**** Requests per day
* ****TPM:**** Tokens per minute
* ****TPD:**** Tokens per day