Models

Supported models

The API currently supports these models:

Llama3.1 8B

- Model id:llama3.1
- orllama3.1:8b
- Full model name:neuralmagic/Meta-Llama-3.1-8B-Instruct-FP8
- Context length: 8192 tokens

Llama3.1 70B

- Model id:llama3.1:70b
- Full model name:neuralmagic/Meta-Llama-3.1-70B-Instruct-quantized.w4a16
- Context length: 128k (131072) tokens

Llama3.1 405B

- Model id:llama3.1:405b
- Full model name:neuralmagic/Meta-Llama-3.1-405B-Instruct-quantized.w4a16
- Context length: 128k (131072) tokens