

Groq — Modelli API per Nexus OS™

Consigliato per Nexus: `openai/gpt-oss-120b`

Groq non è un creatore di modelli — è un provider di inferenza ultra-veloce che ospita modelli open-weight sul proprio hardware LPU (Language Processing Unit). Il valore di Groq in Nexus non è la qualità del modello, ma la **velocità** (262+ token/secondo, 3-10x più veloce dei provider GPU) e il **costo quasi zero** (\$0.15/\$0.60). GPT-OSS-120B è il modello più capace disponibile su Groq: 117B parametri, reasoning controllabile, open weights (Apache 2.0). Nella Round Table sarà il primo a rispondere, dando all'utente feedback immediato. Per chat singole, PED Manager e task ad alto volume è eccellente. **Nota onesta:** Nella peer review della Round Table, GPT-OSS-120B non compete con i flagship degli altri 7 provider (Gemini 2.5 Pro, Sonnet 5, GPT-5.6 Terra). La qualità è inferiore. Il suo ruolo nel Multi-Chat è complementare: velocità, costo contenuto e prospettiva diversa.

Modelli disponibili via API — Luglio 2026

Prezzi Standard tier, per 1M token, in USD. Groq ospita esclusivamente modelli open-weight.

Modelli principali

Stringa API	Input	Output	Context	Forza principale	Note
<code>openai/gpt-oss-120b</code> ★	\$0.15	\$0.60	131K	Reasoning controllabile, miglior qualità su Groq	Consigliato — 117B parametri MoE (5.1B attivi). Apache 2.0
<code>meta-llama/llama-4-scout-17b-16e-instruct</code>	\$0.11	\$0.34	512K	Context window molto ampia	594 TPS — velocissimo
<code>qwen/qwen3-32b</code>	\$0.29	\$0.59	128K	Buon generalist, ragionamento solido	Open weights

Modelli economici / edge

Stringa API	Input	Output	Context	Forza principale	Note
<code>openai/gpt-oss-20b</code>	\$0.075	\$0.30	131K	Ultra-economico con reasoning	Gira su hardware consumer (16GB RAM)
<code>meta-llama/llama-3.1-8b-instant</code>	\$0.05	\$0.08	128K	Il più veloce: 840 TPS	Task semplicissimi, classificazione

Deprecati

Stringa API	Stato
llama-3.3-70b-versatile	Deprecato 17/06/2026 — Sostituto: openai/gpt-oss-120b

Perché Groq e non un altro provider?

I 7 creatori di modelli frontier con API pubblica accessibile dall'Europa sono già tutti nel Tier 1 di Nexus: Google, OpenAI, Anthropic, Mistral, DeepSeek, xAI e Perplexity. Non esiste un ottavo provider con modello proprietario di qualità comparabile.

Le alternative valutate — Cohere (Command A+), Z.ai/Zhipu (GLM-5), AI21 (Jamba), MiniMax — offrono modelli con qualità inferiore ai 7 flagship già selezionati, a prezzi comparabili o superiori. Microsoft Copilot, Meta e IBM non hanno API pubblica diretta per i loro modelli. Together AI e Fireworks sono inference provider come Groq, senza modelli proprietari.

Groq occupa l'ottavo slot portando qualcosa che nessun altro provider offre: inferenza 3-10x più veloce di qualsiasi alternativa GPU, a costi 10-25x inferiori ai flagship. Nella Round Table non vincerà la peer review sulla qualità, ma arriva per primo e non sbilancia la fattura. Per le altre funzioni di Neural Chat — chat singole, PED Manager, broadcast su task semplici — la velocità e il costo di Groq sono un vantaggio concreto.

Dati verificati sulla documentazione ufficiale Groq — groq.com — Luglio 2026