

Nexus OS™ — Endpoint dei Provider

Guida completa per il collegamento ai provider AI — Tier 1, Tier 2 e Tier 3

Per collegare un provider a Nexus, inserisci endpoint e chiave API nel formato: **endpoint;api_key**
Per i modelli locali (Tier 3) basta l'endpoint, nessuna chiave richiesta. Nexus rileva automaticamente i modelli disponibili.

Tier 1 — Provider Dedicati

Questi provider sono gestiti nativamente da Nexus. Per usarli in uno slot Custom con un modello diverso, usa l'endpoint indicato.

Provider	Endpoint	Nota
Google Gemini	https://generativelanguage.googleapis.com	Formato proprietario; nello slot Custom usa via OpenRouter
OpenAI	https://api.openai.com	Nativo OpenAI-compatible
Anthropic Claude	https://api.anthropic.com	Formato proprietario; nello slot Custom usa via OpenRouter
Mistral AI	https://api.mistral.ai	OpenAI-compatible
DeepSeek	https://api.deepseek.com	OpenAI-compatible
Groq	https://api.groq.com/openai	OpenAI-compatible
Grok (xAI)	https://api.x.ai	OpenAI-compatible
Perplexity	https://api.perplexity.ai	OpenAI-compatible

Tier 2 — Provider Cloud (OpenAI Compatible)

Inserisci nel formato: **endpoint;api_key** — Nexus mostrerà la lista modelli disponibili.

Provider	Endpoint	Nota
OpenRouter	https://openrouter.ai/api	Accesso a centinaia di modelli con una sola chiave
Together AI	https://api.together.xyz	Modelli open-source ad alta velocità
Fireworks AI	https://api.fireworks.ai/inference	Inferenza veloce, modelli open-weight
NVIDIA NIM	https://integrate.api.nvidia.com	Modelli ottimizzati su GPU NVIDIA
Cerebras	https://api.cerebras.ai	Inferenza ultra-veloce su hardware dedicato
SambaNova	https://api.sambanova.ai	Modelli open-source su chip proprietari
Lepton AI	https://api.lepton.ai	Serverless, pay-per-token
Cloudflare Workers AI	https://api.cloudflare.com/client/v4/accounts/{ID}/ai	Richiede Account ID nel path
Azure OpenAI	https://{risorsa}.openai.azure.com	Richiede configurazione Azure specifica
Hugging Face	https://api-inference.huggingface.co	Endpoint dedicati per modello

Tier 3 — Modelli Locali

Inserisci solo l'endpoint — nessuna chiave API richiesta. Nexus rileva i modelli installati.

Software	Endpoint	Nota
Ollama	http://localhost:11434	Il più diffuso per modelli locali
LM Studio	http://localhost:1234	Interfaccia grafica, facile da usare
LocalAI	http://localhost:8080	Drop-in replacement OpenAI
llama.cpp (server)	http://localhost:8080	Avviare con --host 0.0.0.0
vLLM	http://localhost:8000	Avviare con --api-key opzionale
Jan	http://localhost:1337	Abilitare il server locale nelle impostazioni
GPT4All	http://localhost:4891	Abilitare il server API nelle impostazioni