Neuphlo

TECHNICAL WHITEPAPER

Bring your own model.

How Neuphlo's BYOM architecture lets every workspace choose its own AI — cloud APIs, local models, even your own coding agents — with your keys, your bill, and your data path, never ours.

No model lock-in

No token markup

Local-model ready

neuphlo.com · 2026

01 — THE PRINCIPLE

§ Why BYOM

Most SaaS products with "AI inside" make the model choice for you — one vendor, one price, your prompts routed through their account. Neuphlo takes the opposite stance: **the platform is model-agnostic, and the model is yours.**

BYOM — bring your own model — means you connect the AI provider you already trust, on your own API key. Neuphlo orchestrates agents, search and automations on top, but it never sits between you and your model commercially: no token markup, no usage caps imposed by us, no silent vendor switch. If a better or cheaper model ships tomorrow, you point your workspace at it and everything — agents, semantic search, automations — follows.

9	1	0
provider types supported	abstraction — every feature works with every provider	markup on your tokens

What BYOM buys you

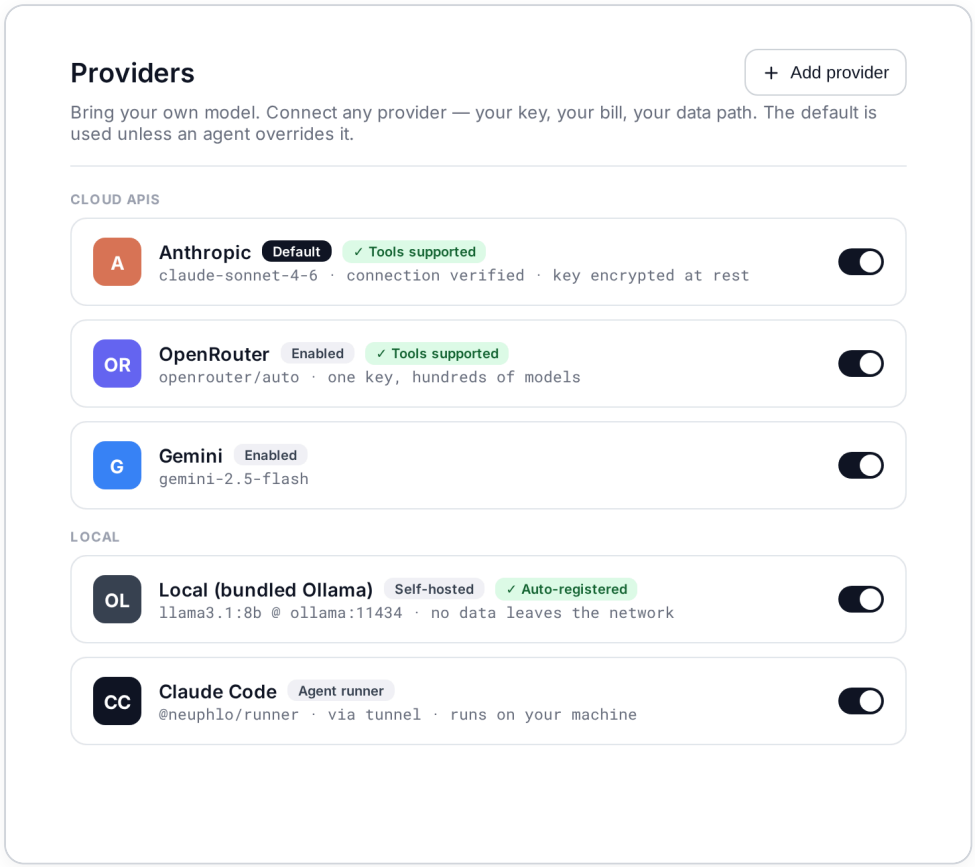
- **Choice.** Anthropic, OpenAI, Gemini, or hundreds of models through OpenRouter — switch any time without touching your workflows.
- **Cost control.** You pay the model provider directly at their prices, and see exactly what each feature consumes.
- **Privacy.** Run a local model and your prompts and documents never leave your own hardware — the strongest privacy stance possible.
- **Power-user leverage.** Already pay for Claude Code, Codex or Cursor? Plug the agent runtime you own straight into Neuphlo as a provider.

> **This document** covers the provider menu, how a connection is made and verified, the local-model and agent-runner paths, per-agent model choice, and the guardrails that keep any model accountable. Screenshots are taken directly from the product.

02 — THE MENU

§ One screen, every kind of AI

Providers live in `Preferences` › `AI` › `Providers` and fall into two groups: cloud APIs you reach with a key, and local providers that run on hardware you control.



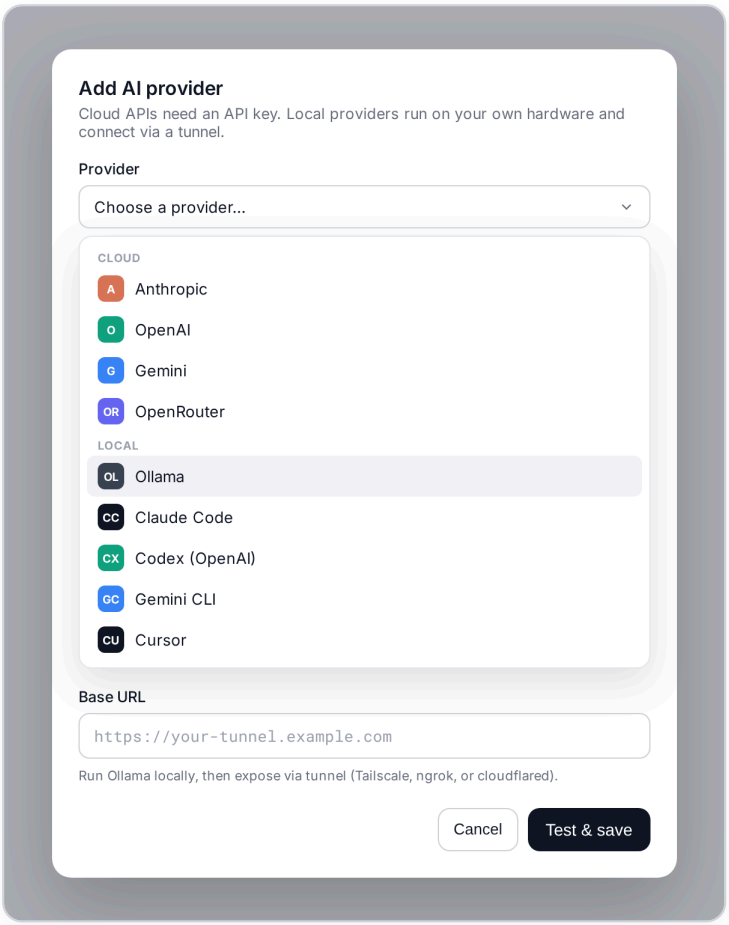
The provider list. Cloud APIs and local providers side by side. One is marked default; agents can override it individually.

PROVIDER	TYPE	NOTES
Anthropic	Cloud	Claude family, native API.
OpenAI	Cloud	GPT family, plus embedding models.
Gemini	Cloud	Google's models via OpenAI-compatible endpoint.
OpenRouter	Cloud	One key, hundreds of models from many vendors.
Ollama	Local	Any open-weights model on your own machine or server.
Claude Code · Codex · Gemini CLI · Cursor	Local	Your own agent runtimes, connected via the Neuphlo runner.

03 — CONNECTING

§ Connect, encrypt, verify

Adding a provider is a one-minute job — but what happens underneath is what makes BYOM trustworthy.



Adding a provider. Cloud providers take an API key; local providers take a base URL, with setup hints inline.

Three guarantees on every connection

Keys are encrypted at rest. API keys are encrypted with a server-side encryption key before they're stored, and decrypted only at the moment a request is built — they're never returned to the browser after saving.

Connections are tested, not assumed. A test call verifies the provider responds — and goes one step further with a tool-call probe that checks whether the model can actually drive Neuphlo's agent loop. A model that chats but can't call tools is flagged before it ever fails silently in an agent run.

Sensible models only. Model lists are filtered per provider — embedding, speech, image and moderation models are excluded from chat pickers automatically, so nobody accidentally points an agent at a model that can't do the job.

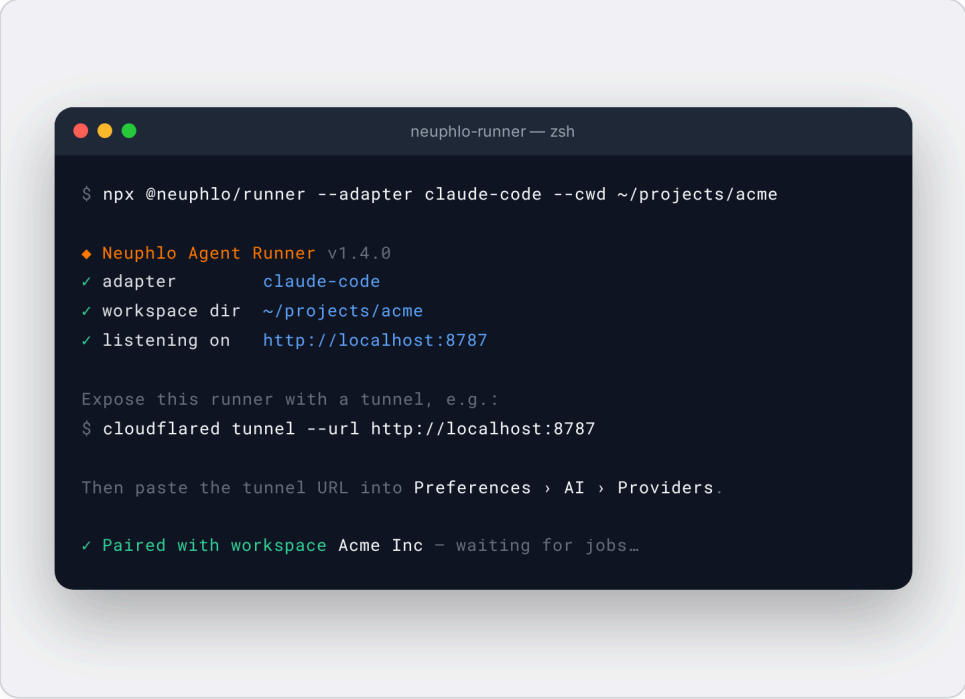
04 — THE LOCAL PATH

§ Local models & agent runners

BYOM's strongest form is a model that never leaves your network. Neuphlo supports two local paths: open-weights models via Ollama, and your own coding-agent runtimes via the Neuphlo runner.

Batteries included on self-host. A self-hosted Neuphlo deployment ships with an Ollama container in the stack, and the installer pulls a starter model (such as `llama3.1:8b`) in the background. It's auto-registered as a provider on workspace creation — AI features work out of the box, with no key, no cloud, and no data leaving the machine.

Your agents as providers. If you already pay for Claude Code, Codex, Gemini CLI or Cursor, the runner turns that subscription into a Neuphlo provider — one command on your machine, exposed through the tunnel of your choice:



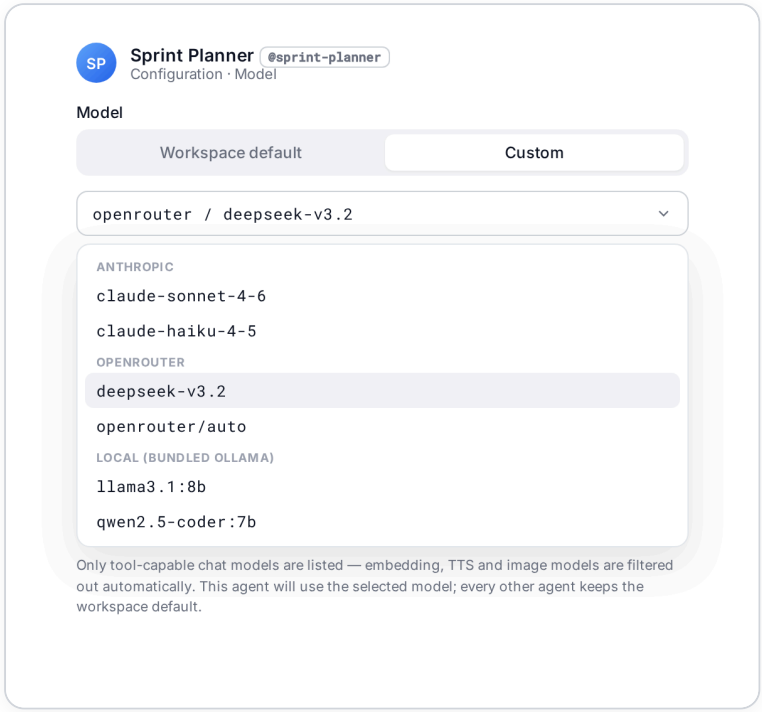
The Neuphlo runner. One command starts an adapter for your local agent runtime; a tunnel (Tailscale, ngrok, cloudflared) makes it reachable, and the URL is pasted into Providers.

> **Why this matters:** the intelligence you've already licensed — and the hardware you already own — become first-class citizens in your workspace. Neuphlo orchestrates; you decide where the thinking happens.

05 — GRANULAR CHOICE

§ The right model for each job

BYOM isn't one global switch. The workspace sets a default, and every agent can override it — so a heavyweight model handles strategy while a small, cheap one does the routine triage.



Per-agent model choice. Each agent either inherits the workspace default or picks a specific model — grouped by provider, filtered to tool-capable chat models.

How teams use it

- **Tiered by stakes.** A frontier model for the orchestrator and customer-facing writing; a local `llama3.1` for internal summarisation and triage.
- **Tiered by sensitivity.** Agents that touch confidential material pinned to the local provider; everything else on a cloud API.
- **Tiered by cost.** High-volume automations on an inexpensive OpenRouter model, with the bill visible in your own provider account.

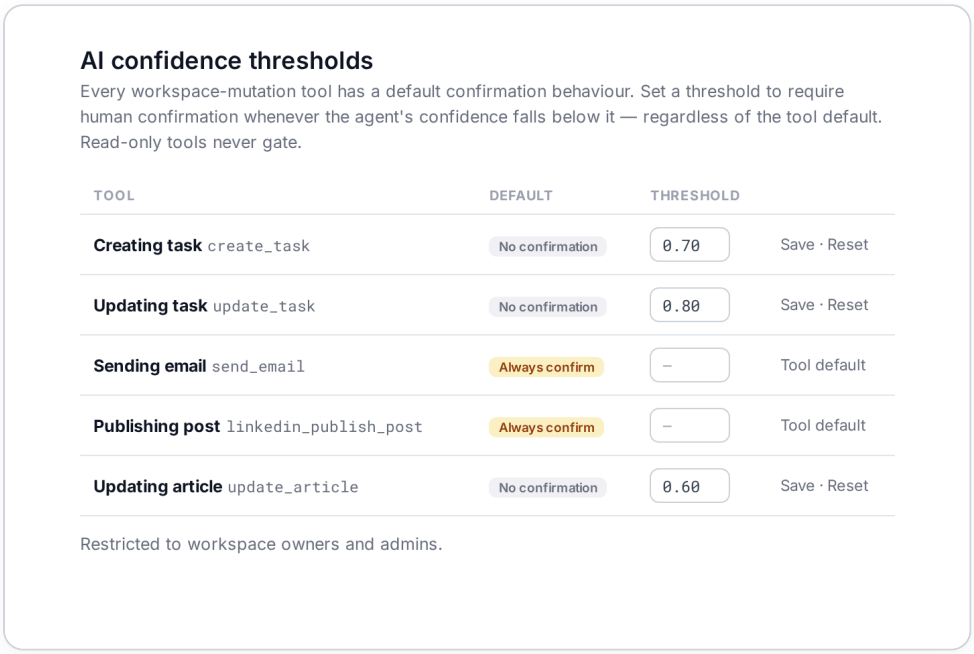
Embeddings get the same treatment: the provider config carries a separate **default embedding model**, so semantic search can run on a different (and far cheaper) model than chat.

06 — TRUST ANY MODEL

§ Guardrails that don't care whose model it is

When the model is interchangeable, safety can't live in the model. Neuphlo's guardrails sit in the platform — so they hold no matter what you plug in.

Every agent action gets a **confidence score** computed from observable signals — whether the action is read-only, whether recent steps succeeded, how deep into the run it is — rather than trusting the model's self-assessment. Workspace owners then set per-tool **thresholds**: below the bar, the action requires human confirmation, whatever the tool's default.



Preferences › AI › Thresholds. Mutation tools carry defaults (some always confirm); owners tighten them per tool with a numeric confidence bar. Read-only tools never gate.

And because the bill is yours, the meter is too: every completion and embedding is logged per workspace, provider, feature and user — tokens and duration included — so you can see exactly which feature, agent or person is consuming what, on which model.

07 — THE BIGGER PICTURE

§ BYOM is a sovereignty feature

Model choice is usually framed as a convenience. For European and privacy-conscious organisations it's a compliance posture: **where the model runs determines where your data flows.** BYOM makes that an explicit, auditable decision instead of a vendor default — and combined with Neuphlo's self-hosting, the entire stack, model included, can run inside your own jurisdiction.

- **Cloud, chosen by you** — your DPA with your provider, your region settings, your keys.
- **Local, owned by you** — open-weights models on your hardware, nothing outbound.
- **Mixed, governed by you** — per-agent model pinning plus platform-side confirmation gates and usage logs.

§ Getting started

PATH	TIME TO FIRST AI FEATURE
Cloud API key	Paste a key in `Preferences › AI › Providers` — about a minute.
Self-host bundled Ollama	Zero — auto-registered when the workspace is created.
Your own agent runtime	One `npx @neuphlo/runner` command plus a tunnel URL.

> **In short:** Neuphlo supplies the workspace, the agents and the guardrails. You supply the intelligence — whichever model, wherever it runs, on whoever's bill you prefer. That's BYOM.

Neuphlo Learn more → **neuphlo.com**