ForexTerminals — Free AI Models & the AI Chatbox

Every free AI model behind your in-game AI Chat, and how to use them — written from the live catalog at freellmapi.co/models · Game version 4.06

The AI Chatbox — Where to Find It

- **Open it:** press Ctrl+1 (the "AI Chat" slot on the hotbar's second row) or click the AI Chat button on the left icon bar. The Playground opens in a browser at http://127.0.0.1:3001/playground, running locally on your own machine.
- **Type and send:** type a message and press Enter (Shift+Enter adds a new line). Press the stop button to cut a long answer short, or regenerate to ask again. Hover any reply to copy it.
- **Conversations:** the left sidebar keeps your chats — start a new one any time and reopen old ones later.
- **Attach files:** add images (PNG, JPEG, WebP, GIF) or text files (TXT, MD, CSV, JSON, LOG) to a message.
- **Voice dictation:** click the microphone and talk instead of typing (Windows shortcut: Win + H).
- **System prompt:** in the right settings rail, give the model a persona — e.g. "You are a medieval tavern keeper." The dot lights up when one is active.
- **Sampling controls:** Temperature, Top P and Max Tokens — leave them off for provider defaults, or switch on only the ones you want.
- **Web search:** toggle it on and the AI looks things up before answering, listing its sources.

What Powers It — FreeLLMAPI

- **One key, dozens of providers:** FreeLLMAPI is a free, open-source AI router that runs on your machine. It puts every free provider tier in its catalog behind a single OpenAI-compatible endpoint — the router sends each request to whichever provider serves the model free.
- **The live catalog (from freellmapi.co/models):** 487 models across 52 providers, every one of them $0 through FreeLLMAPI — with 509 free endpoints listed on the models page. The deepest catalogs right now: ElectronHub (91 models), Cloudflare Workers AI (53), Inferbase (42), AI Horde (31) and ACLIDE (29).
- **How it stays free:** it only calls each provider's own free tier, using keys you add yourself. Inference costs nothing and your keys never leave your machine. No credit card is needed — most providers issue a free key without one.
- **The router never gets stuck:** every provider meters its own free tier (requests/tokens per minute and day). The router counts that usage per key and automatically fails over to the next provider when one is rate-limited or out of daily quota.
- **OpenAI-compatible:** point any OpenAI SDK at http://localhost:3001/v1 with your unified key (shown on the Keys page). The router also speaks the Anthropic Messages surface and Gemini's native API, and every response carries an X-Routed-Via header naming exactly which provider and model served it.

- **Personal use:** built for personal use only — each provider's Terms of Service apply, and the proxy should not be exposed publicly.

The Free Model Types

Type	What It Does
Chat	Text conversation — answering questions, writing, coding, roleplay (the biggest group by far)
Image	Image generation — e.g. Stable Diffusion XL & SDXL Lightning (Pollinations), Moondream vision models
Audio	Speech — Gemini Live models, Whisper transcription (Large v3, Turbo), Voxtral, Typhoon realtime ASR, Speechify Simba
Embeddings	Convert text into vectors for search & similarity — e.g. SEA-LION and Qwen3 Embedding models
Search	Web-search tool providers — Tavily, Exa, Firecrawl, Linkup, Brave Search, SerpSearch (power the Playground's web-search toggle)
Decision	Specialized choice/extraction models for structured decisions

Models Worth Knowing (examples live on the catalog today)

- **Big-context chat:** GLM 5.3 (1M context), Gemini 3.8 Flash / 3.7 / 3.6 / 3.5 Flash (1.0M context, with vision), DeepSeek V4 Flash & Pro (1.0M), Qwen3.8 Flash (1M), MiniMax M3 (1M) — huge context windows for long conversations and big documents.
- **Everyday workhorses:** GPT-OSS 120B & 20B (131K, served by 6–8 providers each — very reliable fallbacks), GLM-4.7 Flash (200K, 5 providers), Qwen3.8 27B (262K, 7 providers), Gemma 4, Typhoon, North Mini Code for code.
- **More than one home:** models listed as "3 providers" or "8 providers" are served by several free providers at once — that is why the router almost always finds one that is not rate-limited.
- **The Auto model:** in the Playground, the default "Auto (fallback chain)" sends your message to the best available model and hops to the next when one is out of quota — you do not even need to pick. "Fusion (multi-model synthesis)" asks several models and a judge synthesizes the best answer.
- **Browse & sort:** on freellmapi.co/models you can filter by All providers / All types and sort Smartest first, Fastest first, Biggest context, or A to Z; each model page lists its exact free-tier limits, context window and provider quirks.

Getting More Models — The Keys Page

- **Works instantly:** some providers are "no-key" — enable them with one click and Auto mode works immediately.
- **Unlock the full catalog:** open the Keys page in the Playground, click a provider (Google AI Studio, Groq, Mistral, NVIDIA NIM, OpenRouter, Cohere…), follow its sign-up link to grab a free API key, and paste it. Keys stay encrypted and never leave your machine.

- **Checklist:** the Keys page shows "X of Y providers configured" plus the keyless providers you can enable at once.
- **Catalog updates:** your router pulls a signed catalog twice a day. If a provider drops a free model, it is retired automatically at the next sync (or instantly on an end-of-life error) and your request fails over to the next model in the chain. The free install pulls a monthly snapshot; Premium ($19/yr) gets the live feed the same day.

Quick Recipe for Friends

- **1. Press Ctrl+1 in the game** — the AI Chat Playground opens in your browser.
- **2. Just start typing** — Auto mode routes you to a free model with no setup at all.
- **3. Want a specific brain?** open the model selector and search — or enable no-key providers / add a free key on the Keys page to see all 487 models.
- **4. Give it a persona** with a system prompt ("medieval scribe!"), attach a screenshot, or dictate with the microphone.
- **5. Rate-limited mid-chat?** do nothing — the router already hops to the next free provider. Errors are rare and temporary.

Catalog facts from freellmapi.co/models (live, updated today at time of reading) · Chatbox features from the game's Playground UI · November 23, 2026 · v4.06