AI BENCH LAB

Cloud Provider Setup

Connect OpenAI, Anthropic, Google Gemini, Groq, and other cloud APIs.

Version 1.0 | March 2026
aibenchlab.com

Guides · Professional AI Model Benchmarking for Windows

Contents

Connect OpenAI, Anthropic, Google Gemini, Groq, and other cloud APIs for remote model benchmarking.

1. Supported Cloud Providers

AiBenchLab supports bring-your-own-key cloud providers out of the box. You supply your API key; the app handles the rest.

Provider	Protocol	Default Endpoint	Models
OpenAI	OpenAI-compatible	`api.openai.com/v1`	GPT-4o, GPT-4o Mini, o1, o1-mini, and more
Anthropic	Anthropic (dedicated)	`api.anthropic.com/v1`	Claude 3.5 Sonnet, Claude 3.5 Haiku, and more
Google Gemini	Google (dedicated)	`generativelanguage.googleapis.com/v1beta/openai`	Gemini 2.0 Flash, Gemini 1.5 Pro, and more
xAI (Grok)	OpenAI-compatible	`api.x.ai/v1`	Grok-2, Grok-2 Mini
Groq	OpenAI-compatible	`api.groq.com/openai/v1`	Llama 3.3 70B, Mixtral, and more
Together AI	OpenAI-compatible	`api.together.xyz/v1`	Llama 3.1 405B, open-source models
Custom	OpenAI-compatible	User-configured	Any OpenAI-compatible endpoint

2. Adding a Cloud Provider

› Open **Settings** from the sidebar.

› Click **+ Add Provider**.

› Select a provider template (e.g., OpenAI, Anthropic).

› Enter your **API key**.

› Click **Save**.

› Click **Test Connection** to verify.

The connection test checks authentication, counts available models, and measures response time. Results include:

- **Success/failure** status
- **Model count** — how many models the provider exposes
- **Response time** in milliseconds
- **Error details** if the connection failed

---

3. API Key Management

Where to Get Keys

Provider	Key Source
OpenAI	platform.openai.com/api-keys
Anthropic	console.anthropic.com/settings/keys
Google Gemini	aistudio.google.com/apikey
xAI	console.x.ai
Groq	console.groq.com/keys

How Keys Are Used

API keys are stored in the local SQLite database on your machine. They are never sent to AiBenchLab servers — the app communicates directly with the provider's API.

Authentication headers vary by provider:

Provider	Header	Format
Anthropic	`x-api-key`	Raw key value
All others	`Authorization`	`Bearer <key>`

Key Display

API keys are masked in the UI for security. Only the first 5 and last 3 characters are shown (e.g., `sk-pr•••••acA`).

---

4. Model Discovery

After adding a provider, AiBenchLab fetches available models automatically:

- **OpenAI, Groq, xAI, Together AI**: `GET /models` (OpenAI-compatible)
- **Anthropic**: Model list is hardcoded (Anthropic does not expose a model listing endpoint)
- **Google Gemini**: `GET /models`

Models appear in the Benchmark Wizard's model selection step and in the Cloud Models section of the sidebar.

---

5. Provider-Specific Notes

OpenAI

- All GPT-4 and o1 model variants are supported.
- Vision-capable models (GPT-4o) can run multimodal benchmark tests.
- Rate limits are respected — the app does not retry on 429 responses during benchmarks.

Anthropic

- Uses the dedicated Anthropic API protocol, not OpenAI-compatible.
- The `x-api-key` header is used instead of `Authorization: Bearer`.
- Claude models support long context windows (up to 200K tokens).

Google Gemini

- Uses a dedicated Google handler for authentication and request formatting.
- Gemini 2.0 Flash is cost-effective for high-volume benchmarking.
- Gemini 1.5 Pro supports up to 1M tokens of context.

Groq

- Extremely fast inference — useful for quick iteration on benchmarks.
- Serves open-source models (Llama, Mixtral) with hardware-accelerated inference.
- Lower cost per token than most proprietary providers.

Custom OpenAI-Compatible

For any provider that exposes an OpenAI-compatible API (e.g., self-hosted vLLM, SGLang, or other inference servers):

› Select **Custom OpenAI-Compatible** as the template.

› Enter the base URL (e.g., `http://my-server:8000/v1`).

› Enter an API key if required (leave blank for unauthenticated endpoints).

› Optionally add custom HTTP headers.

› Save and test the connection.

---

6. Benchmarking Cloud Models

Selecting Cloud Models

In the Benchmark Wizard, cloud models appear alongside local models in the model selection step. They are grouped by provider and show:

- Model name and context window
- Provider badge
- Cost indicator (if pricing data is available)

Select one or more cloud models. You can mix cloud and local models in the same benchmark session.

Cost Awareness

Cloud benchmarks consume API tokens. Before running:

- The Cost Estimator shows projected token usage and cost.
- If the estimated cost exceeds your warning threshold (default $5), a confirmation dialog appears.
- You can adjust the threshold in **Preferences > Cost Settings**.

Performance Metrics

AiBenchLab captures timing data for every cloud API call:

Metric	Description
TTFT	Time to First Token — latency before the first token arrives
TPOT	Time Per Output Token — average time between tokens
TPS	Tokens Per Second — throughput rate
E2E Latency	Total time from request to complete response

These metrics appear in the results alongside quality scores, allowing you to evaluate both capability and speed.

---

7. Troubleshooting

"Authentication failed" (401/403)

- Verify your API key is correct and active.
- Check that the key has the required permissions (some providers restrict access by key type).
- For Anthropic, ensure you're using the `x-api-key` header format (the app handles this automatically — if you see this error, the key itself may be invalid).

"No models found"

- The provider may require a paid plan to access models. Free tiers sometimes have limited model access.
- For custom endpoints, verify the `/models` endpoint returns data.

Rate limiting

- AiBenchLab does not automatically retry rate-limited requests during benchmarks.
- If you hit rate limits, consider running fewer tests in parallel or using a provider with higher rate limits.
- Groq and Together AI typically have more generous rate limits for open-source models.

Connection timeout

- Check your internet connection.
- Verify the provider's status page for outages.
- For custom endpoints behind firewalls or VPNs, ensure the endpoint is reachable from your machine.

Resources

- Documentation: **aibenchlab.com/docs**
- Support: **aibenchlab.com/contact**
- YouTube walkthroughs: **youtube.com/@aibenchlab**
- Email: **support@aibenchlab.com**

AiBenchLab is built by a solo founder with 40+ years of software engineering experience. Every feature is designed with the conviction that you deserve honest, transparent, and reproducible AI model evaluations. No hype. Just truth.