AI BENCH LAB

Quick Start Guide

Download, install, and run your first benchmark in under 5 minutes.

Version 1.0 | March 2026

aibenchlab.com

Getting Started · Professional AI Model Benchmarking for Windows

Contents

1. System Requirements
2. Step 1: Download and Install
3. Step 2: First Launch Setup
4. Step 3: Run Your First Benchmark
5. How the Built-In Server Works
6. What's Next

Download, install, and run your first benchmark in under 5 minutes.

1. System Requirements

Component	Minimum	Recommended
OS	Windows 10+, macOS 10.15+, Linux (Ubuntu 20.04+)	Windows 11, macOS 13+
RAM	8 GB	16 GB+
GPU	None (CPU inference supported)	NVIDIA GPU with 6 GB+ VRAM
Storage	3 GB (app + required components)	Additional space for local AI models

2. Step 1: Download and Install

Windows

› Download `AiBenchLab-Setup.exe` from the Download page.

› Run the installer. Accept the EULA when prompted.

› Choose per-user or system-wide installation.

› Launch AiBenchLab from the Start Menu or Desktop shortcut.

macOS

› Download `AiBenchLab.dmg` from the Download page.

› Open the DMG and drag AiBenchLab to Applications.

› Right-click > Open on first launch (macOS Gatekeeper prompt).

Linux

› Download `AiBenchLab.AppImage` from the Download page.

› Make it executable: `chmod +x AiBenchLab.AppImage`

› Run: `./AiBenchLab.AppImage`

Debian/Ubuntu users can also use the `.deb` package. Fedora/RHEL users can use the `.rpm` package.

3. Step 2: First Launch Setup

When you launch AiBenchLab for the first time, the Setup screen downloads three required components:

Component	What It Is	Download Size
llama.cpp runtime	Built-in inference server for running local models	~141 MB (CUDA) or ~55 MB (Vulkan)
Judge model	Gemma 2 2B IT — scores model responses during benchmarks	~1.5 GB
Quick-start model	Llama 3.2 1B Instruct — a bundled model so you can benchmark immediately	~1.3 GB

The setup screen shows real-time download progress with speed and ETA. GPU detection runs automatically — if you have an NVIDIA GPU, the CUDA backend is selected. AMD and Intel GPUs use the Vulkan backend.

Components are stored locally:

- **Windows**: `%LOCALAPPDATA%\AiBenchLab\components\`
- **macOS**: `~/Library/Application Support/AiBenchLab/components/`
- **Linux**: `~/.local/share/aibenchlab/components/`

Once all three components are installed, the app redirects you to the Benchmark Wizard.

---

4. Step 3: Run Your First Benchmark

Open the Wizard

Click **New Run** in the sidebar. The Benchmark Wizard opens in Simple mode — a guided flow that gets you running in a few clicks.

Describe Your Task

Type a short description of what you need the AI model to do. For example:

> *""Answer customer support questions about our product documentation""*

The wizard interprets your intent and recommends relevant test disciplines. You can also skip this step and manually select disciplines from the grid.

Select a Model

The bundled quick-start model (Llama 3.2 1B Instruct) is already available. Select it by checking the box. If you have Ollama or LM Studio running, their models appear here too.

Review and Launch

Review the selected disciplines and test count. Click **Create & Enqueue** to start.

Watch Progress

The execution screen shows:

- Current test name and progress percentage
- Pass/fail counters updating in real time
- GPU temperature and VRAM usage (NVIDIA GPUs)
- Elapsed time

Benchmarks pause automatically if GPU temperature exceeds 80°C and abort at 95°C.

View Results

When the benchmark completes, click **View Full Results**. You'll see:

- **Overall Score** (0–100) — a composite measure of model capability
- **Per-domain breakdowns** — how the model performed in each tested area
- **Individual test results** — score, pass/fail status, and dimension-level detail

---

5. How the Built-In Server Works

AiBenchLab ships with its own llama.cpp inference server. You do not need Ollama, LM Studio, or any external tool to run benchmarks.

The built-in server:

- Launches automatically when you select a local GGUF model
- Runs an OpenAI-compatible API on auto-assigned ports (8900–9999)
- Supports GPU acceleration (CUDA for NVIDIA, Vulkan for AMD/Intel)
- Falls back to CPU if no GPU backend is detected

A separate **Judge Server** runs on port 8899 during benchmarks. It uses the Gemma 2 2B model on CPU so it doesn't compete for GPU VRAM with the model under test.

---

6. What's Next

- **Connect more providers**: Add Ollama, LM Studio, or cloud APIs for more models to test.
- **Browse the Model Catalog**: Over 51,000 models indexed from HuggingFace, searchable and filterable by size, license, and GPU fit.
- **Explore test suites**: 22 pre-built suites covering production readiness, role-specific evaluation, and app-specific benchmarks. Two suites (Quick Compare and Customer-Facing Chat) are available on the free Trial tier.
- **Understand scoring**: Read Understanding Your Score to learn how the composite Score, domain breakdowns, and pass thresholds work.

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.