

AI BENCH LAB

Quick Start Guide

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

Version 1.1 | April 2026

aibenchlab.com

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	6 GB (app + required components)	Additional space for local AI models

2. Step 1: Download and Install

Windows

1. Download `AiBenchLab-Setup.msi` from the Download page.
2. Run the installer. Accept the EULA when prompted.
3. Choose per-user or system-wide installation.
4. Launch AiBenchLab from the Start Menu or Desktop shortcut.

macOS

1. Download `AiBenchLab.dmg` from the Download page.
2. Open the DMG and drag AiBenchLab to Applications.
3. Right-click > **Open** on first launch (macOS Gatekeeper prompt).

Linux

1. Download `AiBenchLab.AppImage` from the Download page.
2. Make it executable: `chmod +x AiBenchLab.AppImage`
3. 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 3 4B Instruct Q8_0 — scores model responses during benchmarks	~4.1 GB
Quick-start model	SmolLM2-135M — a bundled model so you can benchmark immediately	~101 MB

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.

Tell Us What You Need (Intent & Model Selection)

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's classifier interprets your description and recommends relevant test disciplines as selectable tags. Models from all configured providers appear in the same step, scored and ranked against your intent — pick one or more.

You can also start with a **Quick Pick** (e.g., "Agent Builder", "Code Assistant") to skip the description and load preset disciplines instantly.

Select Tests

Review the tests that will run based on your selected disciplines. In Simple mode, the test plan is auto-assembled. In Advanced mode you can customize individual tests, add tests from other domains, and save the result as a custom suite (Consultant tier and above).

Review & Queue

Set session name and run parameters. Pick a test suite — Trial users get Quick Compare and Customer-Facing Chat; more suites unlock with Pro. Confirm everything on the final review screen and click **Create & Enqueue**.

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 **Gemma 3 4B Instruct Q8_0** for scoring model responses. The judge server runs on its own bundled, version-locked llama.cpp instance to ensure scoring consistency across releases.

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 Trial tier.
 - **Understand scoring:** Read *Understanding Your Score* to learn how the composite Score, domain breakdowns, and pass thresholds work.
-

All data stays on your machine. AiBenchLab collects no telemetry — your benchmarks are yours alone.

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.