AI BENCH LAB

The Benchmark Wizard

A complete walkthrough of the Benchmark Wizard — from workload description to live results.

Version 1.0 | March 2026

aibenchlab.com

Getting Started · Professional AI Model Benchmarking for Windows

Contents

A complete walkthrough of the Benchmark Wizard — from workload questionnaire to live results.

---

1. Overview

The Benchmark Wizard is an 8-step guided flow that transforms a description of your AI workload into a reproducible benchmark plan. It operates in two modes:

- **Simple mode** (default): Natural language input, streamlined steps, minimal decisions. Best for getting started quickly.
- **Advanced mode**: Structured questionnaire with full control over every parameter. Best for experienced evaluators who need granular configuration.

Toggle between modes using the **Simple / Advanced** switch at the top of the wizard. Your choice persists across sessions.

---

2. Simple Mode Walkthrough

Step 1: Describe Your Task

Type a natural language description of what the AI model needs to handle:

> *""I need a model that can review pull requests, suggest code improvements, and explain complex functions to junior developers.""*

The wizard sends your description to a two-stage classifier that:

› **Categorizes** your intent (e.g., software, reasoning, safety, agentic)

› **Selects disciplines** — specific evaluation areas within that category, ranked by relevance

The result appears as selectable discipline tags below the input. Each tag shows the discipline name, test count, and estimated run time.

Confidence framing: The wizard is transparent about how well it understood your request:

- **High confidence**: "Based on your description, **Code Generation** is a strong match — it tests whether the model can produce correct, idiomatic code from natural language prompts."
- **Medium confidence**: "My best guess is **Code Debugging**, though **Code Review** could also fit. Do any of these look right?"
- **Low confidence**: "I wasn't confident enough to recommend specific tests. Could you tell me more about what the model will actually be doing day-to-day?"

You can add or remove disciplines manually regardless of the classifier's suggestions.

Example — Alex Benchwright at Acme Agentic Consultants:

Alex types: *"We're deploying an AI coding assistant for our engineering team. It needs to generate code, review PRs, and handle SQL queries."*

The wizard identifies three disciplines: Code Generation, Code Review, and SQL & Database — and shows them as selected tags with test counts.

Step 2: Select Models

Models from all configured providers appear in a unified list. Each model card shows:

- Model name and provider
- Parameter count and context window
- **Match percentage** — how relevant this model is to your selected disciplines (green >70%, yellow 40–70%, red <40%)
- VRAM fit indicator (green = fits, yellow = tight, red = too large)

Select one or more models by checking their boxes. The bundled quick-start model and any locally running models (Ollama, LM Studio) are always available. Cloud models appear if you've configured API keys.

In Simple mode, model selection is organized per-discipline — each discipline card lets you pick up to 3 models.

Step 3: Review and Queue

Review your selections:

- Disciplines and their test counts
- Selected models per discipline
- Estimated total run time

Click **Create & Enqueue** to submit the benchmark to the execution queue.

---

3. Advanced Mode Walkthrough

Advanced mode expands the wizard into 8 full steps with granular control.

Step 1: Structured Questionnaire

Instead of free-text input, you fill out a detailed form:

Field	Options
Primary Intent Type	General, Code, Reasoning, Vision, Agentic, Writing

Field	Options
Output Requirements	Structured JSON, Tool Calling, Code Generation, Multimodal Input, Long Context, High Accuracy
Context Bucket	Short (<4K tokens), Medium (4K–16K), Long (16K–64K), Very Long (64K+)
Deployment Constraints	Local Only, Cloud Only, Privacy Sensitive, Budget Friendly, Open Source Preferred
Latency Priority	Speed, Balanced, Quality
Target Hardware	This system (auto-detected) or different system (manual VRAM/budget entry)

Quick Picks — 8 pre-assembled configurations for common use cases (e.g., Document Q&A, Code Review, Safety Evaluation) that pre-fill the form with a single click.

Step 2: Model Selection

Same unified model list as Simple mode, but with the full filter and scoring engine active. Models are scored against your questionnaire answers, not just discipline matches.

Step 3: Suite and Discipline Selection

Browse all 22 pre-built test suites. Each suite shows:

- Name, description, and category
- Test count and estimated duration
- Tier requirement (Free or Pro)
- Last run date (if previously used)

Select a suite, or let the wizard auto-assemble a plan from your questionnaire responses using the intent block system.

Step 4: Review and Customize Tests

See every test that will run, organized by domain. For each test:

- **Weight**: Required, Recommended, or Optional
- **Score weight**: 1.0–2.0 multiplier
- **Contributing block**: Which intent block included this test and why

You can toggle individual tests on or off. If a test is excluded by one block but required by another, the conflict is displayed with an explanation of how it was resolved (required wins).

Save as Custom Suite (Consultant tier and above): Save your customized test selection as a reusable suite.

Step 5: Session Configuration

Setting	Description
Session Name	Auto-generated or custom. Appears in results and reports.
Contact Name	Optional. Included in PDF reports.
Deterministic Mode	Lock execution order and disable sampling randomness.
Global Seed	Seed value for deterministic mode.
Temperature Override	Override the model's temperature setting for all tests.
Runtime Target	GPU or CPU. Auto-detected hardware shown for reference.

Step 6: Final Review

A summary of every choice: disciplines, suite, tests, models, configuration. Last chance to go back and adjust before submission.

Step 7: Cost Estimator

For cloud models, the cost estimator shows:

- Projected input and output token counts
- Per-model API cost breakdown
- Total estimated cost
- Confidence level (High, Medium, Low)

If the estimated cost exceeds your warning threshold (default $5), a confirmation dialog appears. See Cost Estimator for details.

Step 8: Benchmark Execution

The live execution screen displays:

- **Test progress**: Current test name, X of Y completed
- **Counters**: Pass, fail, and anomaly counts
- **Elapsed time** with running clock

- **GPU telemetry** (NVIDIA): Temperature gauge (color-coded), GPU load, VRAM usage
- **Activity log**: Timestamped messages for each test start, completion, and notable event

Controls:

- **Stop** — Cancel the running benchmark
- **View Full Results** — Appears when execution completes

Thermal protection is always active: benchmarks pause at 80°C and abort at 95°C.

---

4. Intent Block System

Behind the scenes, the wizard assembles benchmark plans using **intent blocks** — pre-built modules that map capabilities to specific tests.

Each intent block defines:

- **What it proves** — the capabilities being tested
- **Which tests** — drawn from the 254-test library, with weights (required/recommended/optional)
- **Exclusions** — tests that would be misleading for this use case
- **Score weights** — multipliers (1.0–2.0) for test importance

When multiple blocks are combined, the assembly engine:

› Deduplicates tests, keeping the highest weight

› Merges score weights (capped at 2.0)

› Resolves conflicts between blocks (required always wins over excluded)

› Generates an audit trail showing which block contributed each test

This system ensures that benchmark plans are repeatable, explainable, and grounded in real test coverage — not guesswork.

---

5. Tips

- **Start with Simple mode.** Switch to Advanced only when you need fine-grained control.
- **Be specific in your description.** "Answer customer questions from our knowledge base" produces better matches than "general chatbot."
- **Check discipline coverage.** If a discipline shows a "thin coverage" warning, results will be directional rather than definitive.
- **Run multiple models.** The real value of benchmarking comes from comparing models side by side.

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.