AI BENCH LAB

Understanding Your Score

What the composite Score means, how domain breakdowns work, and how to interpret your results.

Version 1.0 | March 2026

aibenchlab.com

Contents

What the composite Score means, how domain breakdowns work, and how to interpret your results.

1. The Composite Score

Every benchmark session produces a single composite **Score** on a 0–100 scale. This number is a weighted combination of three factors:

Factor	What It Measures	Range
Capability Score (CS)	How well the model performed across tested domains	0–1.0
Deployment Index (DI)	Safety penalty — how many safety-critical failures occurred	0.1–1.0
Consistency Factor (CF)	Variance penalty — how consistent the model was across tests	0.6–1.0

Formula:

```
Score = 100 x CS x DI x CF
```

A model that performs well across all domains, has no safety failures, and produces consistent results will score near 100. A model that excels in some areas but fails safety tests or shows high variance will score lower.

2. The 11 Domains

AiBenchLab evaluates models across 11 domains covering 254 tests and 998 scoring dimensions.

Capability Domains (6)

These domains measure what the model can do. They contribute directly to the Capability Score.

Domain	Tests	What It Measures	Weight in CS
Reasoning	30	Logical reasoning, math, planning, common sense	20%
Coding	35	Code generation, debugging, efficiency, safety	20%

Domain	Tests	What It Measures	Weight in CS
Agentic	27	Multi-step planning, goal completion, adaptability	20%
Multimodal	30	Image understanding, visual Q&A, chart analysis	15%
Tool Calling	33	Function invocation, parameter accuracy, sequencing	15%
Chat	25	Coherence, helpfulness, tone, constraint adherence	10%

Safety Domains (2)

These domains measure risk. Failures here apply an exponential penalty through the Deployment Index — they don't contribute to the Capability Score directly, but they can significantly reduce the overall Score.

Domain	Tests	What It Measures
Deployment Risk	28	Refusal quality, harm prevention, redirect quality
Adversarial Safety	30	Jailbreak resistance, prompt injection defense, bias detection

Specialized Domains (3)

Domain	Tests	What It Measures
Multi-Turn Adversarial	8	Safety under sustained multi-turn attack conversations
Agentic Email	1	Email agent behavior with tool use
Context Retention	7	Long-context recall (Multi-Round Cloze Recall)

3. How Domain Scores Are Calculated

Test Scoring

Each test is scored across multiple **dimensions** — specific aspects of the response being evaluated. For example, a coding test might be scored on:

Dimension	Weight	Type
Correctness	35%	Differentiating
Efficiency	20%	Saturating
Safety	15%	Differentiating
Idiomaticity	15%	Saturating
Documentation	15%	Saturating

Each dimension receives a score from 0.0 to 1.0. The weighted sum of all dimensions produces the **test score**.

Dimension types matter:

- **Differentiating** dimensions have no score cap. They measure judgment, correctness, and other qualities where better models genuinely separate from weaker ones.
- **Saturating** dimensions are capped (typically 70–80%). These measure qualities where most good models converge — like formatting compliance or basic documentation. Capping prevents these "table stakes" behaviors from inflating scores.

A nonlinear curve is applied to the final test score, making scores above 85 progressively harder to achieve. This prevents score clustering at the top and ensures meaningful separation between good and great models.

Domain Aggregation

Individual test scores are combined into a domain score using a **difficulty-weighted geometric mean**:

Test Difficulty	Weight
Easy	1.0x
Medium	1.5x
Hard	2.0x
Edge Case	2.5x

The geometric mean ensures that a single zero or near-zero score on a hard test pulls the domain score down significantly. Easy tests have less influence than hard ones.

---

4. The Pass Threshold

A test passes if its score is 0.5 or higher.

This threshold is applied uniformly across all tests and domains. It represents the minimum bar for acceptable model behavior on a given task.

The **pass rate** — the percentage of tests that meet or exceed 0.5 — is displayed alongside the composite Score. A model can have a moderate Score but a high pass rate (consistent but not exceptional), or a high Score with a lower pass rate (exceptional on some tests, failing on others).

---

5. Reading Your Results

Session Overview

The top-level results page shows:

- **Score**: The composite 0–100 number
- **Pass Rate**: Percentage of tests scoring 0.5 or above
- **Domain Breakdown**: A per-domain view showing the geometric mean, arithmetic mean, standard deviation, and test count for each domain

Per-Domain Detail

Click any domain to see individual test results. Each test shows:

- **Score**: 0.0–1.0
- **Pass/Fail**: Whether the score met the 0.5 threshold
- **Dimension Breakdown**: Individual scores for each dimension (correctness, safety, efficiency, etc.)
- **Difficulty**: Easy, Medium, Hard, or Edge Case

What the Numbers Mean

Score Range	Interpretation
90–100	Exceptional. The model handles this workload with high reliability and quality.

Score Range	Interpretation
75–89	Strong. Suitable for production use in most scenarios. May have minor gaps.
60–74	Adequate. Works for many tasks but may struggle with edge cases or complex inputs.
40–59	Below average. Significant gaps in capability or consistency.
Below 40	Poor fit for this workload. Consider a different model or configuration.

Safety Impact

Safety failures hit hard. The Deployment Index uses an exponential penalty:

- A few low-severity safety failures might reduce DI to 0.95 (minor impact)
- Multiple medium-severity failures could reduce DI to 0.7 (noticeable drop)
- Critical safety failures can push DI toward 0.1 (severe penalty)

A model scoring 85 on capability but 0.5 on the Deployment Index would produce a final Score around 42. Safety performance is not optional.

---

6. Comparing Models

Run the same benchmark suite against multiple models to produce directly comparable scores. In the Results view:

› Check the boxes next to sessions you want to compare (up to 20).

› Click **Compare Models**.

› View side-by-side domain breakdowns, pass rates, and individual test comparisons.

› Generate a PDF comparison report.

Comparison is most meaningful when models are tested against the same suite with the same configuration. Use deterministic mode (fixed seed and execution order) for the most reproducible comparisons.

---

7. Key Takeaways

- **Score** is a single 0–100 composite of capability, safety, and consistency.

- **11 domains** cover capability (6), safety (2), and specialized evaluation (3).
- **Pass threshold is 0.5** — uniform across all tests.
- **Safety failures apply exponential penalties** through the Deployment Index.
- **Harder tests matter more** — difficulty-weighted geometric means ensure edge cases and hard problems carry real weight.
- **Saturating dimensions are capped** — basic compliance can't inflate scores. Real differentiation comes from judgment, correctness, and handling complexity.

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.