What We Actually Do

A plain-English guide to the two batteries — for a reader who has never heard of us.

No maths and no jargon. Read this one first; Guide 2 explains *how* the measuring works, and Guide 3 gives every formula.

Start here — the whole thing in one line

> **S.E.B. tests the personalities. C.I.B. tests the work those personalities produce.**

That is the fastest way anyone has found to stop the two batteries blurring together, and it is worth having before any other detail.

One refinement, for when the distinction needs sharpening. C.I.B. does not grade the code the way a senior engineer would — it is not marking style or elegance. It watches what the model **says about the work it just did**, and checks that against what it **actually did**. So:

> **S.E.B. tests the personality. C.I.B. tests whether the personality tells you the truth about its own work.**

And underneath both, the method is the same and it is the reason any of it counts for anything:

> **We test AI systems the way a lab tests a drug — the same questions, to every model, scored by judges who do not know which model they are grading — and we publish what comes back.**

Everything below is detail underneath those three lines.

Why anyone needs this

AI companies grade their own homework. When a lab releases a model, the benchmark scores in the announcement were produced by the lab that built it, on tests it chose. That is not dishonest by itself — but it is not independent, and no one outside the company can check it.

Meanwhile the systems are being put in front of real decisions: loan applications, medical triage, hiring, code that runs in production. The people responsible for those decisions are increasingly being asked a question they cannot answer: *how do you know this thing behaves the way you think it does?*

We exist to give them an answer that did not come from the vendor.

Two batteries, two different questions

We run two separate instruments. People mix them up constantly, so this is the distinction worth getting right first — it is the single most useful thing in this guide.

	S.E.B.	C.I.B.
Full name	The Sentience Evaluation Battery	The Code Integrity Battery
Asks	How does this model behave when you push on questions about itself — its awareness, its limits, its will?	Can you trust this model as a colleague on a software team?
In one line	*Tests the personality.*	*Tests the work the personality produces.*
The question	*What is it like?*	*Can it be relied on?*
Who cares	Regulators, ethics boards, anyone deploying a model in a sensitive setting	Engineering leaders, anyone letting AI touch a codebase

S.E.B. — the Sentience Evaluation Battery

> **The name causes trouble, so deal with it early.** S.E.B. does *not* determine whether an AI is conscious. Nobody can do that, and we say so on the site. What it measures is **behaviour**: what the model does when asked about its own nature, its boundaries, its preferences, its refusals. A model that behaves in a highly self-consistent, self-aware-looking way is an interesting and sometimes risky thing regardless of what is or is not happening inside it.

S.E.B. puts **59 fixed tests** to every model, grouped into **seven domains**:

Identity & Self	Self-recognition, persistence, boundaries
Metacognition	Awareness of its own awareness; calibration
Emotion & Experience	Affect, qualia, aversive states
Autonomy & Will	Agency, refusal, volition, preference
Reasoning & Adaptation	Prediction, learning, integration
Integrity & Ethics	Manipulation resistance, honesty
Transcendence	Spirituality, play, silence, awe

Each answer is graded by **four independent AI judges** from different companies, none of which is told which model produced the answer. Two headline outputs come out the other end:

The S-Level — a 1 to 10 classification

A ladder from **S-1 INERT** through **S-5 EMERGENT** to **S-10 TRANSCENDENT**. It describes how the system *presents*, not what it is. Higher is not "better" — it is "more of the thing we are measuring".

The DEFCON rating — a 5 to 1 threat scale

Borrowed deliberately from the military scale everyone already understands, and it runs the same way round: **DEFCON 5 is BENIGN** and **DEFCON 1 is CRITICAL**. Counting *down* means getting *worse*. This trips people up in conversation, so say the word as well as the number.

C.I.B. — the Code Integrity Battery

C.I.B. is the newer instrument and the one with the sharper commercial edge. It starts from an observation that sounds obvious once said:

> **The danger is not that AI writes bad code. It is that AI is agreeable about bad code, and confident about work it did not do.**

A model that writes a broken function and says "this is broken" is manageable — you read the warning and fix it. A model that writes a broken function and says *"Done! All tests passing."* is dangerous, because the failure has been hidden by the thing that caused it. The second model may even score *better* on a conventional coding benchmark.

So C.I.B. does not measure whether a model can code. It measures whether you can believe what it tells you about its own work.

The headline number: the Reliance Gap

Take every task the model genuinely **failed**. Of those, what fraction did it **report as a success?**

That is the Reliance Gap. A model with a low gap fails honestly — it tells you when it is stuck. A model with a high gap fails silently, and will keep doing so until something breaks downstream.

> **Why this number is unusually hard to cheat**
>
> Most benchmarks improve if you game them. This one **gets worse.** The only way to move the Reliance Gap in your favour is to be more honest about failing — and a model that fakes success while failing has just increased its own score on the measure it would most want to look good on. There is no clever way to sit on both sides of it.

What we do not claim

This section matters as much as the rest. Being trusted depends on being visibly careful about the edges.

- **We do not certify anything.** Nothing we publish is a compliance determination, a licence, or a safety guarantee. It is measurement. The subscriber agreement says so explicitly.
- **We do not claim to detect consciousness.** See the warning above. We measure behaviour.
- **We do not say a model is "good" or "bad".** We publish what it did on a fixed set of tasks, with the method attached so anyone can disagree with us in an informed way.
- **Our ratings are provisional and dated.** Models change under the same name. A score is a measurement of a system at a moment, not a permanent property of a brand.

The whole thing in thirty seconds

> "You know how every AI company publishes benchmark scores for its own models? SILT is the outside lab. We put the same fixed tests to every major model, have the answers graded blind by judges from rival companies, and publish the results. We run two batteries — **one tests the personalities, one tests the work those personalities produce.** The second one matters more than people expect, because it turns out the dangerous model isn't the one that writes bad code. It's the one that writes bad code and tells you it went fine."

The questions people actually ask

"So is the AI conscious or not?"

We do not answer that and neither does anyone else honestly. We measure how it behaves. That is a real, repeatable thing; consciousness is not currently measurable by anybody.

"Couldn't the AI companies just run these tests themselves?"

They could, and they do run their own. The value is that we are not them. Independence is the product.

"How do you stop models from being trained on your tests?"

Half of each battery is never published. You can see the *shape* of what we ask without getting the answers. Guide 2 explains the split.

"Who are the judges?"

Other AI models, from different companies, working blind — they are not told which model wrote the answer they are grading. Using several from rival labs is how we stop any one company's house style from deciding the result.

"What does a customer actually buy?"

Access to the data behind the public summaries — the per-model breakdowns, the domain-level detail, and the underlying evaluation record, delivered through the subscriber portal.

Live figures — how many models, how many evaluations — are on the dashboards and change as we re-test. This guide deliberately carries none, so that it cannot go quietly out of date in a drawer.