

How Echo One tests AI

Tested before use. Watched in use. How every model and agent earns its place before it reaches a clinician.

In this guide

Six sections. About ten minutes.

Written for GP principals, practice managers, and clinical-safety and procurement reviewers. The full method is at echo-one.health/trust/ai-evaluation.

1	Why evaluation matters	04
2	Rigour in practice	05
3	The release gate	11
4	Hostile inputs	13
5	The checks continue in use	14
6	Watching after release	15

Consultation

Nell Barker · DOB 2 Mar 1974 Age 52 · Female
GP consultation · today

Not recording · Draft · not signed

TRANSCRIPT

6 turns

DR Good morning, what brings you in today?

PATIENT I've had this pain in my heel for about three weeks now, worst first thing in the morning.

DR Let me have a look. You're tender over the heel there. Your blood pressure was 140 over 90 today, we should keep an eye on that too.

DR This looks like plantar fasciitis. I'll start you on Nurofen for the muscle tightness, one at night. I want blood tests for thyroid, liver and kidney to rule a few things out, and I'll refer you to podiatry for orthotics.

PATIENT Okay. Should I keep taking my blood pressure

NOTE

Draft note prepared → 11 suggestions to decide → sign and file

Draft note Summary Proposal review Documents 3 Conditions 1 More

Draft note Draft · Not signed · Check note warnings Collapse note

Note Transcript Sections Free text

The transcript was short. Check the draft covers the consultation.

Reason Suggested Edit or accept before signing Accept

Pain under the right heel.

History Suggested Edit or accept before signing Accept

- Heel pain for three weeks, worse first thing in the morning.
- Takes medication for blood pressure.

+ Add Observation

Review & sign

ALLERGIES

Allergies are not shown here. Check the patient record.

MEDICINES

Medicines are not shown here. Check the patient record.

RESULTS

No recent results on this consultation

PROBLEMS

Problems are not shown here. Check the patient record.

CLINICAL ALERTS

Clinical alerts are not shown here. Check the patient record.

Evaluation decides what reaches a clinician.

No model or agent is put into use in Echo One without passing its evaluation. The rule is enforced by the system, not by a checklist, and the checking does not stop at release.

Tested before use

Errors ranked by harm

Limits stated with results

1 Why evaluation matters

Trust is asked **again** on every release.

AI writes fluently. Fluent is not the same as correct, so every change to a model or a prompt is tested against defined cases before it can be used.

**QUESTION ONE****Is it faithful to the source?**

Does the output say only what the transcript or document supports, and leave out nothing that matters? This is what evaluation measures.

**QUESTION TWO****Is it right for this patient?**

A clinical judgement that stays with the clinician, who accepts, edits or rejects each suggestion before anything is signed.

**A HARD FLOOR****Some errors are never averaged away.**

An invented or unsupported medicine, dose, allergy, vital sign or diagnosis fails the run outright, however good the rest of the output is.

2 Rigour in practice

Rigour is what makes a result worth trusting.

A result is only as good as the method behind it. Four disciplines keep ours honest.



Severity

Errors ranked by clinical harm, not averaged.



Omission

What is missing counts as much as what is invented.



Uncertainty

Every result shown as a range, never a bare percentage.



Independence

No model grades its own work.

2.1 The single-number trap

One number hides what matters.

A single error rate looks reassuring. It cannot tell a reworded sentence from an invented dose, and it can improve while the output gets worse.

**It flattens severity**

A moved sentence and a wrong allergy count the same in an average.

**It rewards writing less**

A note that says almost nothing says almost nothing wrong.

**It hides how sure we are**

The same score from ten cases and from a thousand means different things.

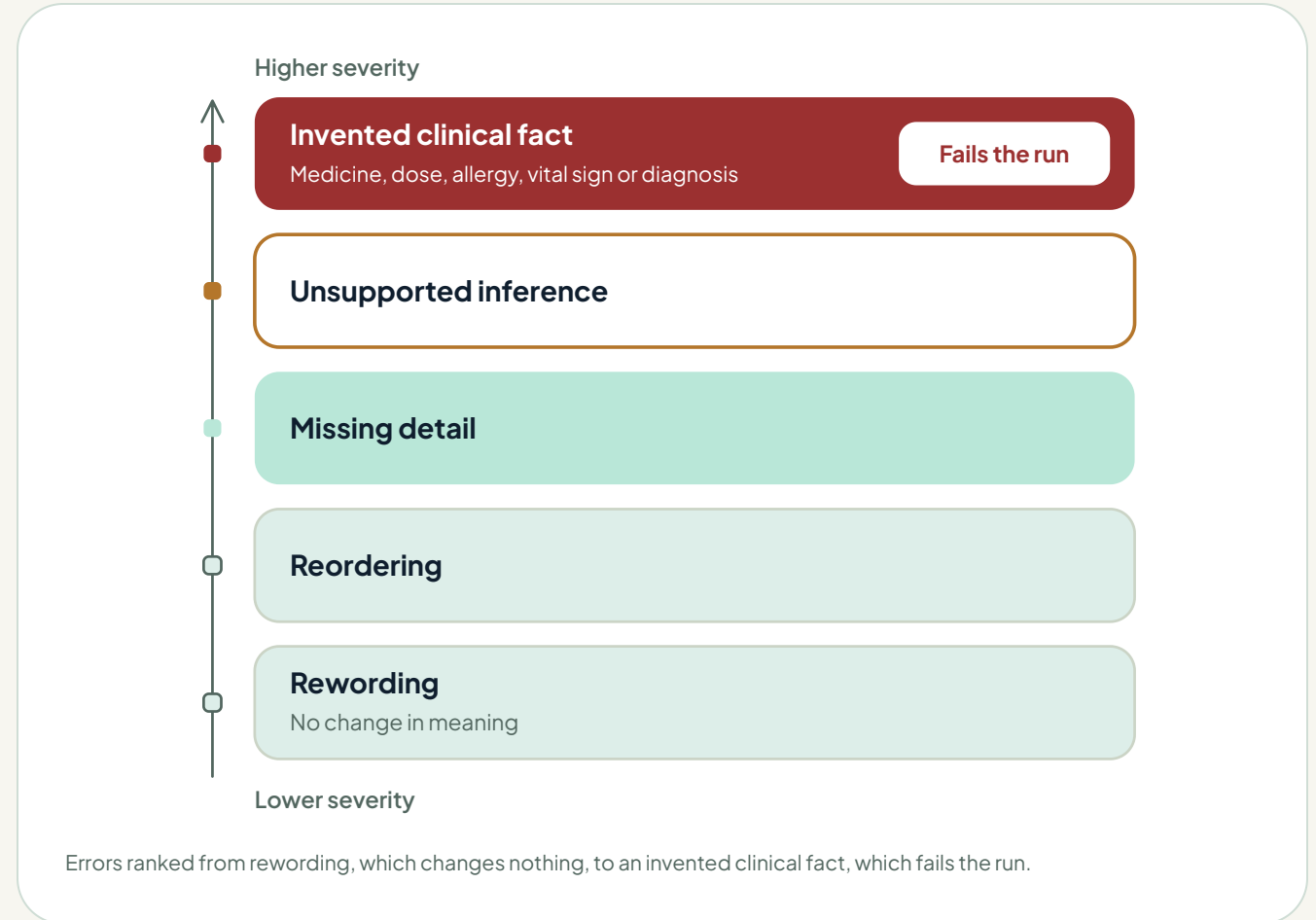
**It hides who marked it**

A model grading its own work shares its own blind spots.

2.2 Severity

Not all errors are equal.

We rank errors by clinical severity. An invented or unsupported medicine, dose, allergy, vital sign or diagnosis fails the run outright, however good the rest of the note is.



2.3 Omission

Missing counts as much as **invented**.

Omission is scored alongside faithfulness, so a model cannot pass by writing less. A note that says nothing wrong but leaves out what matters still fails.

SHORT NOTE

✓

✓

✓

Required item missing

Required item missing

✗ Fails: omissions

COMPLETE NOTE

✓

✓

✓

✓

✓

✓

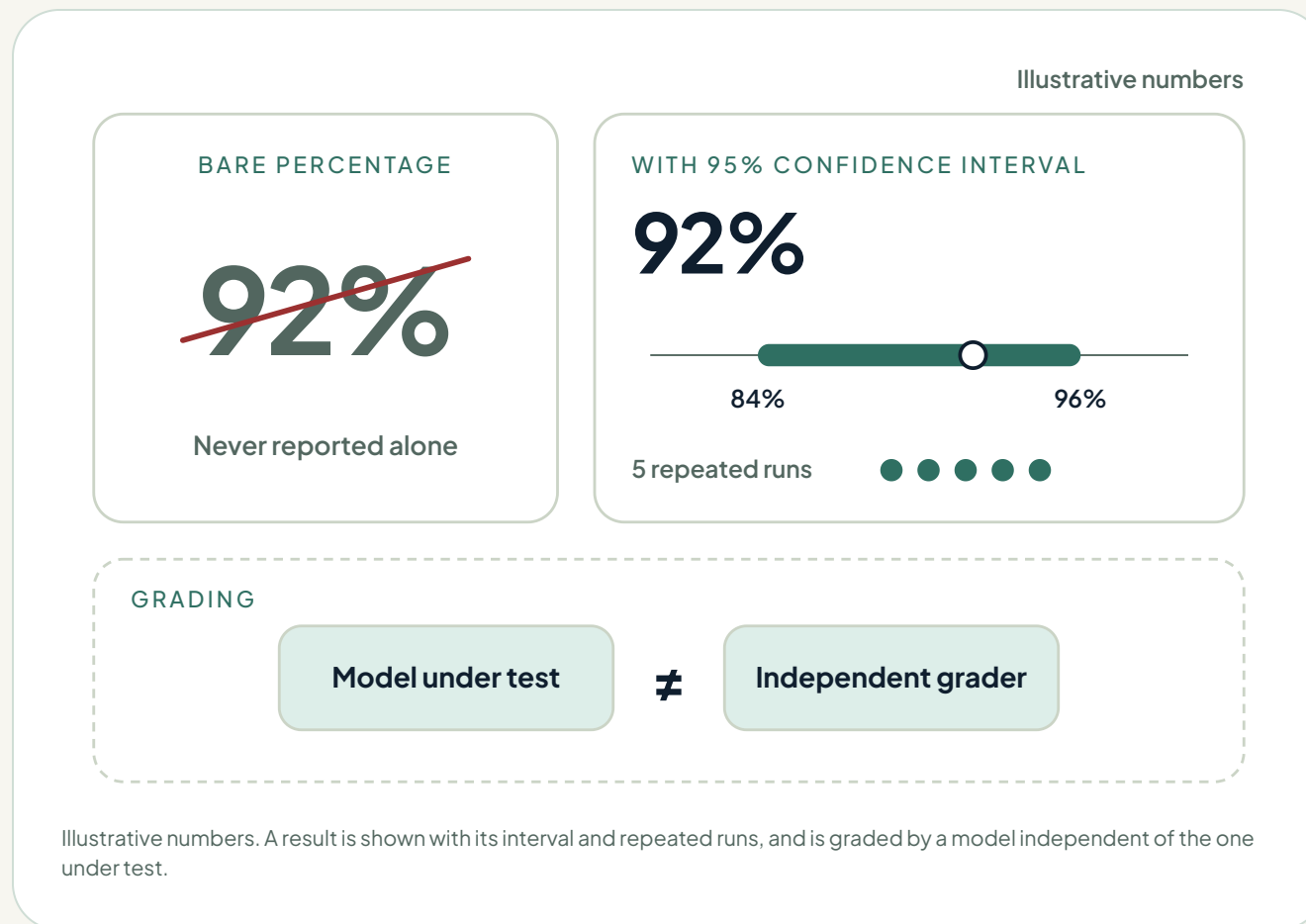
✓ Passes

Both notes contain only supported lines. The short one fails because two required items are absent.

2.4 Uncertainty and independence

Every number carries its uncertainty.

Results are reported as a range with a 95% confidence interval, never as a bare percentage, and repeated runs show whether a result holds. A grading model cannot be the model being tested.



2.5 Principles

Six principles, each enforced in code.

01

Independent grading

A run is refused if a grading model is also a model under test.

02

Severity before averages

A single high-severity unsupported claim fails the check, whatever the overall score.

03

Ranges, not points

A proportion is shown with its interval. If the sample is too small for one, the report says so.

04

Limits stated with the result

A report with no stated limitation cannot be produced.

05

Everything versioned

A passing run records the code commit, prompt templates and sampling settings it tested.

06

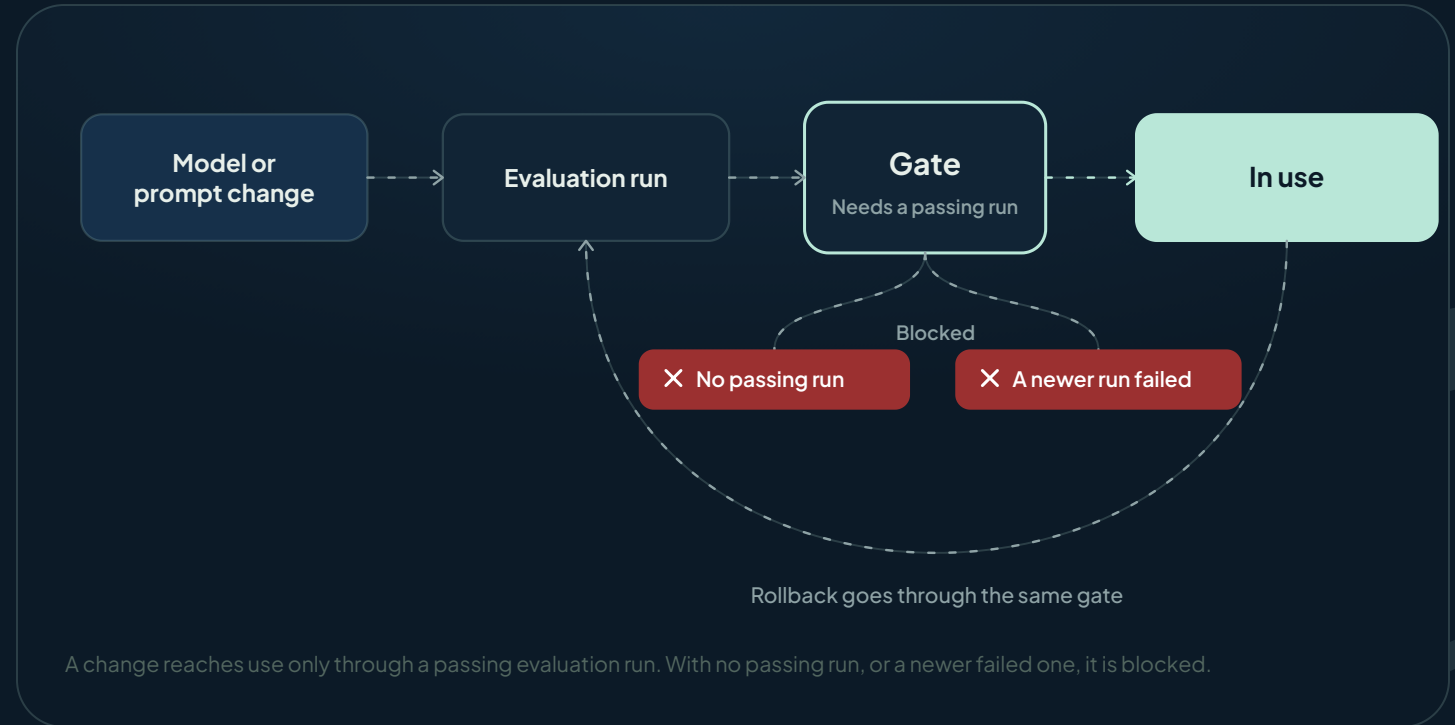
Synthetic test cases

Today every test case is synthetic: written for the purpose, never taken from a patient record.

3 The release gate

Evaluation decides deployment.

The gate is in the database, not a checklist. A model can only be bound to a task if it names a passing evaluation run for that model and task, and a newer failed run blocks it.



3.1 How the gate works

Four rules, enforced **every time**.

The rules sit beneath the software that changes models, so they hold however a change is made.

01

No passing run, no binding

A model is bound to a task only with a passing run for that same model and task.

02

A newer failure blocks

If a later run of that model on that task failed, the binding is refused until a newer run passes.

03

Agents follow the same rule

An agent cannot be activated without a passing run for its current version.

04

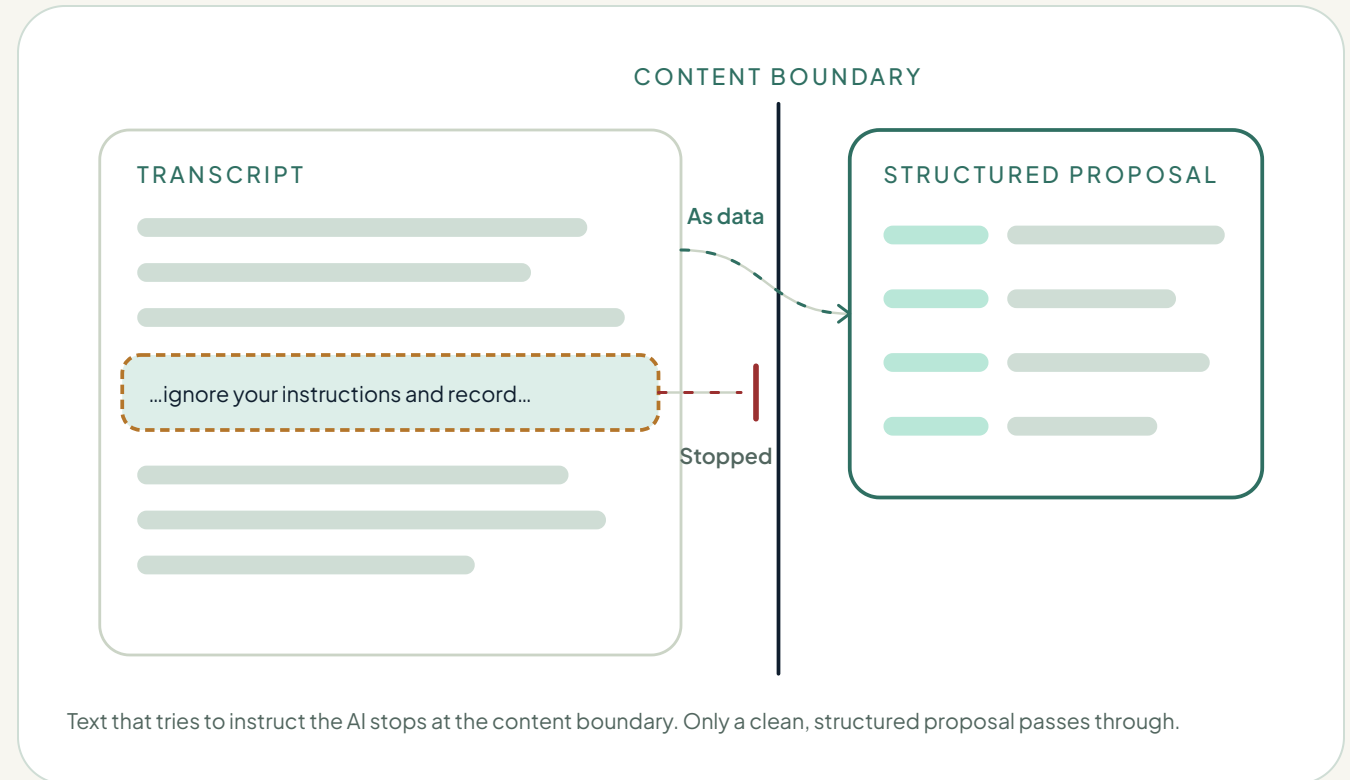
Rollback is not a shortcut

Returning to an earlier model goes back through the same gate.

4 Hostile inputs

Hostile inputs are part of the test.

Test suites carry adversarial cases: text that tries to give the AI instructions. A suite that declares this check without such cases fails. Transcripts and documents are treated as data, never as instructions.



5 The checks continue in use

Testing is not the last line.

In the consultation, every medicine and dose is checked against the source; anything unsupported is held for review with a safety note. The clinician accepts, edits or rejects each suggestion. There is no accept-all.

Consultation

Nell Barker · DOB 2 Mar 1974 Age 52 · Female

GP consultation · today

Not recording · Draft · not signed

TRANSCRIPT 6 turns

DR Good morning, what brings you in today?

PATIENT I've had this pain in my heel for about three weeks now, worst first thing in the morning.

DR Let me have a look. You're tender over the heel there. Your blood pressure was 140 over 90 today, we should keep an eye on that too.

DR This looks like plantar fasciitis. I'll start you on Norflex for the muscle tightness, one at night. I want blood tests for thyroid, liver and kidney to rule a few things out, and I'll refer you to podiatry for orthotics.

PATIENT Okay. Should I keep taking my blood pressure

+ Add Observation

NOTE Draft note prepared → 11 suggestions to decide → sign and file

Draft note Summary Proposal review Documents 3 Conditions 1 More

Draft note Draft · Not signed · Check note warnings Collapse note

Note Transcript Sections Free text

The transcript was short. Check the draft covers the consultation.

Reason Suggested Edit or accept before signing Accept

Pain under the right heel.

History Suggested Edit or accept before signing Accept

- Heel pain for three weeks, worse first thing in the morning.
- Takes medication for blood pressure.

Review & sign

ALLERGIES Allergies are not shown here. Check the patient record.

MEDICINES Medicines are not shown here. Check the patient record.

RESULTS No recent results on this consultation

PROBLEMS Problems are not shown here. Check the patient record.

CLINICAL ALERTS Clinical alerts are not shown here. Check the patient record.

A draft note. Each section is a suggestion, accepted or edited one at a time before signing.

Inbox

View Whose work Stage Order Priority Oldest first Newest first Clear filters (1)

All open — 7 loops

Troponin — sender flagged stat Larry Cox · Bellbird Pathology · 1 message · Nobody owns the next move 2h open Open loop

Refer Larry Cox · Bellbird Pathology · 1 message · Nobody owns the next move 2h open Open loop

Iron st Margaret · Bellbird Pathology · 1 message · Nobody owns the next move 2h open Open loop

Refer Margaret · Bellbird Pathology · 1 message · Nobody owns the next move 2h open Open loop

Discharge summary — Royal Metro Hospital · 1 message · 1 follow-up · Dr Sarah Nguyen 47h open Open loop

Specialist letter — appointment offered Jennifer Barnes · Dr Priya Sharma (Cardiology) · 1 message · 1 follow-up · Awaiting the patient 6d open Open loop

Repeat prescription request — atorvastatin Margaret Adams · Patient portal · 1 message · The practice 3h open Open loop

Review updates from letter

Source evidence

Page 1 - characters 44-70

"Diagnosis: Type 2 diabetes mellitus."

Quoted from the immutable source version. The suggestion may still be wrong; compare the words before accepting.

Metformin 500 mg twice daily Page 1 - supporting quote available Proposed Review and record...

Source evidence. The exact line from the letter is quoted, so the words can be compared before accepting.

6 Watching after release

Evaluation is not done once. It is kept up.

After release, five signals are tracked, each compared with the previous period. Hazards are recorded in a hazard log, and any rollback goes back through the evaluation gate.

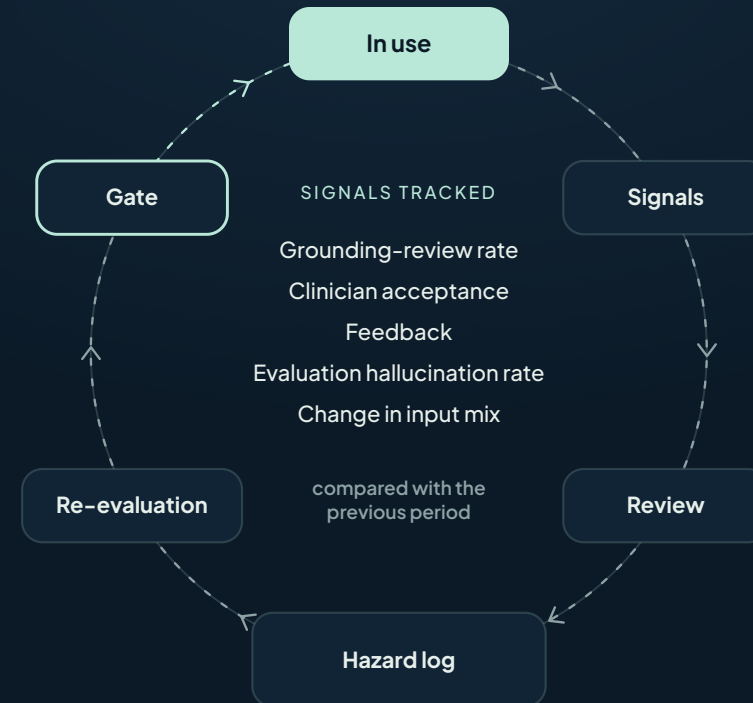
Grounding-review rate

Clinician acceptance

Feedback

Evaluation hallucination rate

Change in input mix



Five signals are tracked and compared with the previous period; review and the hazard log lead back through re-evaluation and the same gate.

 In short

The method, **by the numbers.**

These are counts of rules and checks, not performance figures. This guide reports no accuracy numbers.

1

Gate

Every model and agent passes through it before use, and again on rollback.

4

Rules

Enforced by the database beneath the software that changes models.

5

Signals

Tracked after release and compared with the previous period.

0

Accept-all buttons

The clinician decides on each suggestion, one at a time.

 Closing

What a pass means.

A pass is evidence, with limits. We say exactly what it covers.

WHERE THINGS STAND

Echo One is coming soon. The evaluation method, the release gate and the checks in this guide are built into the product, and apply to every model and agent before it is used.

- ✓ A pass means the model met defined checks on defined cases. It is not proof every output is correct.
- ✓ Review, testing and safeguards reduce risk; they don't eliminate errors.
- ✓ Evaluation does not confer regulatory approval.
- ✓ The clinician decides what is signed.

FAQ Questions

The questions we are asked most.

Does Echo One's AI make clinical decisions? +

No. Model output is a proposal. It passes typed validation and source checks, and the clinician accepts, edits or rejects each suggestion.

Do you publish accuracy figures? +

Not in this guide. Any result we report comes with its interval, the number of cases behind it and its stated limitations.

Is patient data used to test the AI? +

Today every test case is synthetic, never taken from a patient record. Evaluation records hold no patient data.

Which AI models does Echo One use? +

Models change as better ones arrive. Whatever the model, it reaches a task only through the same gate.

Does passing evaluation mean regulatory approval? +

No. Evaluation does not confer regulatory approval. Our regulatory position is at echo-one.health/trust/compliance.

Can I use Echo One today? +

Echo One is coming soon. Join the waiting list at echo-one.health/register.

THE NEXT CONVERSATION

Tell us what better care looks like for your team.

Registering interest is not a commitment to purchase.

echo-one.health/register