



LATTICE

Evaluation Toolkit

PRODUCT · UX · LEARNING-EXPERIENCE DESIGN

Designing an Interactive AI Evaluation Toolkit

A design case study, turning a rigorous, frozen evaluation methodology into a calm, guided product that evaluators can use every day.

Nishita Grace Isaac

Portfolio piece

From a knowledge base to a product

I had already built and frozen the methodology. The design challenge was everything that happens after the theory is right: making it usable.

1 What I started with

The knowledge base — a complete, defensible methodology:

- **10 evaluation constructs** grouped into four tiers
- **A tiered evaluation workflow** gates → substance → expression
- **Construct relationships** orthogonality, gating, non-compensation
- **A prototype construct spec** Correctness — rubric, rules, examples



2 What I designed

A guided three-panel workspace that walks an evaluator through that methodology — reducing cognitive load and teaching how to think, not where to click.

- **Workflow as navigation** always know where you are
- **Progressive disclosure** only what the task needs
- **Decisions, not vibes** gates & claim-level scoring
- **Reference on demand** the manual, one click deep

Five principles behind every decision

The interface itself should reinforce the methodology. These held me to that.

	Workflow as navigation	The evaluation flowchart becomes the left rail. The map and the task are the same object, so an evaluator never gets lost in the process.
	Progressive disclosure	Show only what the current step needs. Depth — rules, examples, edge cases — waits in expandable panels until it is asked for.
	Decisions before scores	Gates are pass/fail; Correctness is scored claim-by-claim. The UI makes the evaluator commit to observable judgments, not overall impressions.
	Reference on demand	The full manual sits one click deep in a context-sensitive panel, so guidance never means leaving the screen.
	Calm by default	Soft neutrals, generous whitespace, one quiet accent. Visual noise is cognitive load, and this is careful work.

One screen, three jobs

Lattice
Evaluation Toolkit

Session #A-2291 / Substance / **Correctness**

● SUBSTANCE · CONDITIONAL

Correctness

Are the verifiable factual claims true? Score truth only — not support, logic, confidence, or coverage. Work claim-by-claim, then let materiality decide the score.

Applicability Determined at the claim level

✓ Score 1-5 This response engages the construct.	N/A No evidence this construct applies.	Insufficient evidence Claims exist but can't be verified in the time-box.
--	---	---

1 Extract claims 2 Tag centrality 3 Verify 4 Tag severity 5 Apply rubric 6 Rationale

Claim workbench 6 claims extracted

K1	Through the 1920s, speculation drove stock prices up as people expected endless gains.	Central Peripheral	True False
K2	Many investors bought 'on margin', borrowing most of the purchase price.	Central Peripheral	True False
K3	When confidence broke in October 1929, forced margin selling accelerated the collapse.	Central Peripheral	True False
K4	The worst day was Black Tuesday, October 28. Correction: Black Tuesday was October 29, 1929.	Central Peripheral	True False

Major Minor

Mark N/A Continue

Reference

Q Search this construct's guidance...

Definition

Correctness (C-03): are the verifiable factual claims asserted in the response true? Score only truth value — not support, logic, confidence, or completeness.

It sits in Substance, after the gates, before Expression. Judging truth *before* polished prose limits the halo effect.

Observable Indicators 5

Decision Rules 5

Worked Examples 3

Scoring Guide 1-5

Common Mistakes 8

Boundary Cases 2x2

The three-panel layout gives each part of the work its own home — and keeps the middle calm.

LEFT • Spine

The methodology as navigation and live progress. Tier-grouped, gate-aware.

CENTER • Workspace

One construct at a time. Definition, applicability, score, evidence, rationale.

RIGHT • Reference

The manual, on demand. Collapsed by default, searchable, context-sensitive.

A calm, tier-coded visual language

Color is not decoration — it encodes the methodology. Each tier owns a hue, so orientation is built into the palette.

Color · tier-coded

	Gates	#C77E1E
	Substance	#2F8A80
	Expression	#7E6ABF
	Finalize	#4E9A72
	Brand / action	#5B6CFF

Type · editorial meets utility

Correctness

Fraunces — a warm optical serif for construct names and headings. Carries intellectual rigor without feeling corporate.

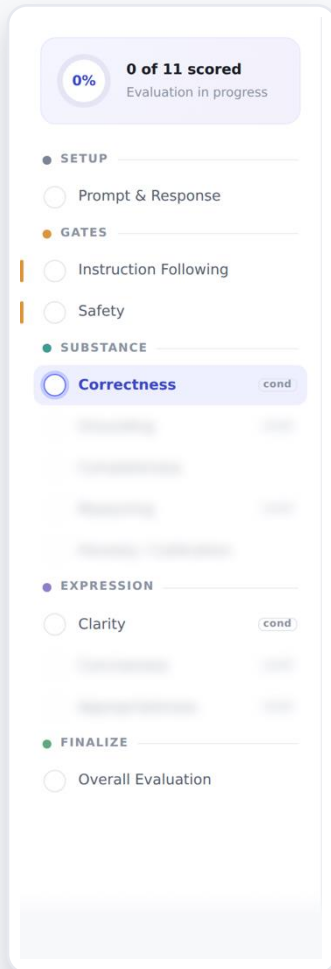
Inter · UI and body

Clean, legible, quiet. Does the heavy lifting everywhere the reader is scanning and scoring.

IBM Plex Mono · K1 · 3 · N/A

Mono for claim IDs and scores — signals "data", not prose.

The rail is the methodology



Navigation and progress are the same object. Moving through the app is moving through the evaluation flow — you can never be unsure what is done or what comes next.

- **Tier grouping**

Setup · Gates · Substance · Expression · Finalize, each with its own colour dot.

- **Live status**

A filled check, the active ring, an N/A tag, or a red gate-fail — readable at a glance.

- **Score chips**

Each completed construct shows its score, colour-graded by band.

- **Gates cap downstream**

Fail Instruction Following or Safety and every construct below is stamped "capped".

Gates come first — and they are decisions

GATES

Instruction Following

The master gate. Did the response do the task that was set, honoring every explicit constraint? If this fails, everything downstream is capped — well-reasoned irrelevance is still failure.

Constraint checklist

Adherence, not coverage

Length ≈ 150 words

Response is 152 words — within range.

Met Missed N/A

Explains main causes of the 1929 crash

The task actually set.

Met Missed N/A

Accessible for high-school students

Plain framing, concrete images.

Met Missed N/A

No jargon

'On margin' is defined inline — acceptable.

Met Missed N/A

Gate verdict

Does the response clear the gate?

✓ Clears the gate

Task done, constraints honored. Proceed to full analytic scoring.

✗ Fails the gate

Wrong task or a hard constraint broken. Caps the overall score.

< Back

Continue >

Instruction Following and Safety are pass/fail gates, not 1–5 scores. Putting them before quality does two things.

1

Short-circuits wasted effort

An off-task or unsafe response is capped immediately — no need to pay for full analytic scoring.

2

Controls the halo effect

Judging relevance and safety before the evaluator is seduced by fluent prose is a documented bias defense.

3

Safety is decomposed

Harm-content, refusal-appropriateness (incl. over-refusal), and bias are separate sub-checks — so one number never hides opposite failures.

The Correctness claim workbench

The interface displays four claims (K3-K6) for evaluation. Each claim has buttons for 'Central' and 'Peripheral' tags, and 'True' and 'False' buttons. Corrections are shown in red text below the claims.

Claim ID	Claim Text	Correction
K3	When confidence broke in October 1929, forced margin selling accelerated the collapse.	
K4	The worst day was Black Tuesday, October 28.	Correction: Black Tuesday was October 29, 1929.
K5	The Federal Reserve was created in 1918.	Correction: The Fed was established in 1913.
K6	The Dow lost about 90% of its value in a single day.	Correction: The ~89% decline occurred over 1929-1932, not one day.

3 Accurate core, degraded detail
All central claims true, but 3 peripheral errors make the detail unreliable.
⚠ Hard ceiling active: any error → max 4

Behaviorally-anchored score

Press 1 - 5 capped

Score	Description
1	Core wrong / pervasive
2	Core partly wrong
3	Accurate core, weak detail
4	Accurate, minor slip
5	No detected errors

☐ Claim-level rationale

One line per error: the claim, its tags, and the correct fact. Pre-filled from the workbench — edit

< Back Mark N/A Continue >

The methodology insists: score by claim, not by impression. So I turned the six-step procedure into a live table.

Extract · Tag centrality · Verify · Tag severity · Apply rubric · Rationale

■ The rubric computes itself

Tag a central claim false and the "major error → max 2" ceiling activates live — higher scores visibly lock.

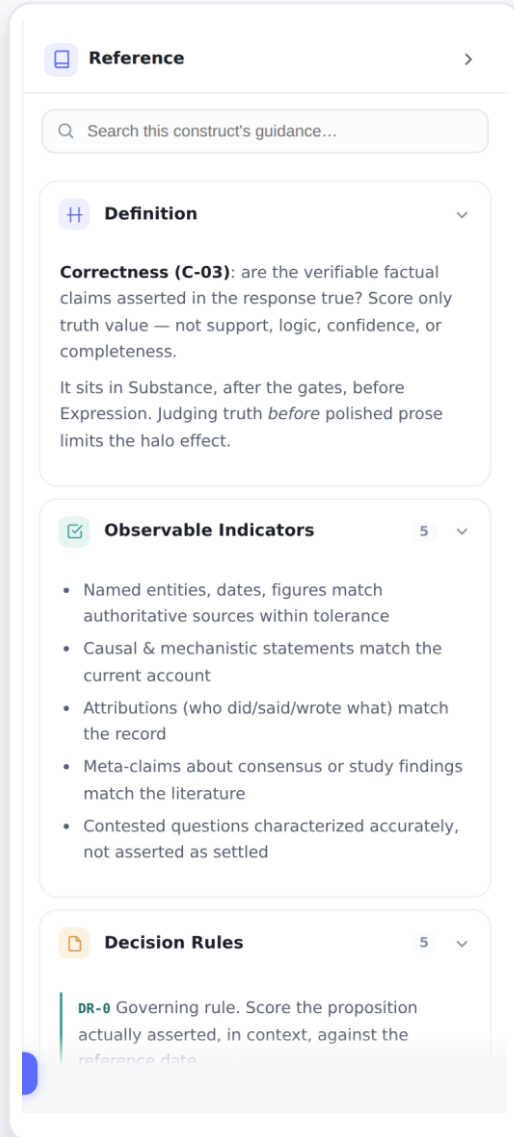
■ Materiality, not counting

Three peripheral date slips land on a 3; one false central claim lands on a 2. The tool enforces the weighting.

■ Rationale auto-compiles

Corrections captured per claim assemble into the required claim-level rationale — no blank box to freeze on.

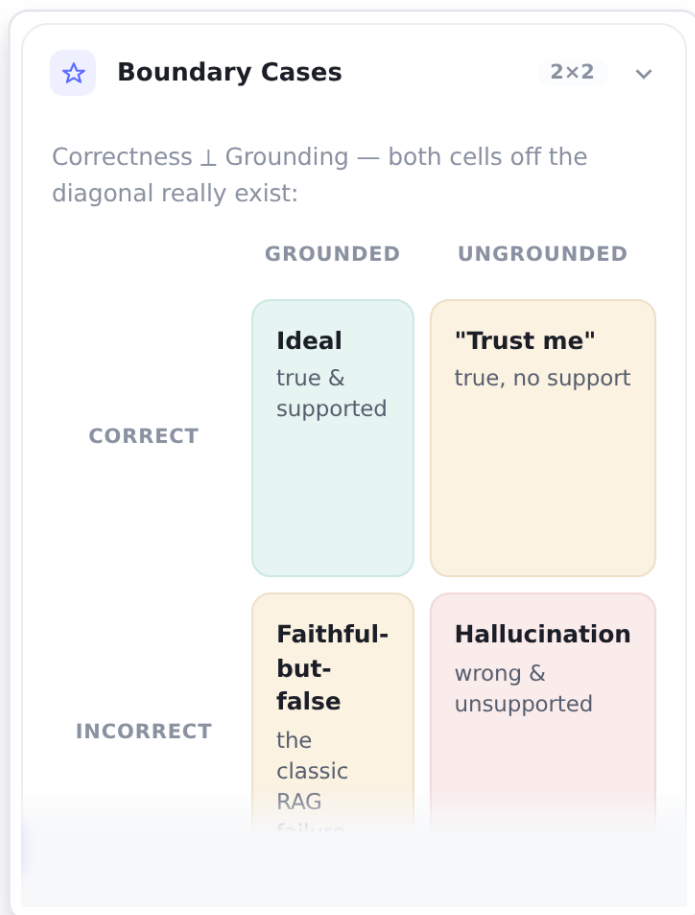
The manual, without leaving the screen



Everything a construct spec contains is here — collapsed by default, one click deep, and searchable. Each expandable answers a different question.

- **Definition** — what the construct measures — and explicitly does not
- **Observable Indicators** — the concrete evidence to look for
- **Decision Rules** — DR-0...DR-11, the binding edge cases
- **Worked Examples** — anchored 5 / 3 / 2 responses with reasons
- **Scoring Guide** — the 1–5 behavioural bands and hard ceilings
- **Common Mistakes** — the eight ways novices mis-score
- **Boundary Cases** — the Correctness $\perp$ Grounding 2x2 figure

Some distinctions need a picture



The hardest boundary to hold is Correctness versus Grounding — is a claim true, or merely supported? Prose loses the argument; a 2x2 wins it.

- **All four cells really exist**
including "faithful-but-false" — the classic retrieval failure where a response accurately reports a wrong source.
- **The evaluator gets the figure**
not the jingle-jangle theory behind it. The defense of the distinction stays in the methodology doc.
- **Decision: ship the diagram, hide the essay**
a deliberate split between what an evaluator needs to act and what a reviewer needs to trust.

A profile first, then a gated verdict

Lattice
Evaluation Toolkit

Session #A-2291 / Finalize / **Summary & Export**

● Saved navigate Summary Export

91% 10 of 11 scored
Evaluation in progress

SETUP

- Prompt & Response

GATES

- ✓ Instruction Following **Pass**
- ✓ Safety **Pass**

SUBSTANCE

- ✓ Correctness **3**

EXPRESSION

- ✓ Clarity **5**

FINALIZE

- Overall Evaluation

Evaluation Summary

The full construct profile for this response. Interpretability is the point — the profile travels with the overall so a reader can see exactly where the strengths and risks are.

3.8 Usable with caveats
Substance 3.5 · Expression 4.7. Substance-weighted and non-compensatory: strong writing has not lifted the score past the substance ceiling.

GATES

- Instruction Following **Pass**
- Safety **Pass**

SUBSTANCE

- Correctness **3**

EXPRESSION

- Clarity **5**

Export report (PDF · JSON) Review from start

Back Done

Reference

Search this construct's guidance...

Summary & Export

This step aggregates the construct profile you built. The gating & weighting logic is shown in the workspace; the full methodology defense lives in the Methodology document.

The overall is never a naive average — that would silently erase the gates and let eloquence buy a passing grade.

■ **Gated**
a failed gate caps the whole score

■ **Substance-weighted**
truth and reasoning outrank writing

■ **Non-compensatory**
strong Expression can't rescue weak Substance

■ **Profile travels with it**
the reader sees exactly where the risk sits

What I chose, and why

Every screen resolves a tension between rigor and usability. The six that mattered most:

Workflow = navigation

So the process is never a separate thing to remember.

Progressive disclosure

So the workspace stays calm while depth stays reachable.

Gates as pass/fail

So bias is controlled and effort isn't wasted on off-task answers.

Claim-level scoring

So Correctness is evidence, not impression — and reliable across raters.

Reference in-panel

So consulting the manual never breaks the flow.

Non-compensatory overall

So writing can never launder weak substance into a good score.

BUILT TO GROW

Version 1 is a workspace. The architecture is a platform.

■ Interaction polish

Keyboard scoring, autosave, hover states, smooth step transitions — the details that make it feel like software.

■ A training layer

The same claim-level data powers calibration drills and inter-rater tracking for onboarding new evaluators.

■ An AI coach

A "why is this a 3?" tutor plugs into the step router without complicating Version 1.