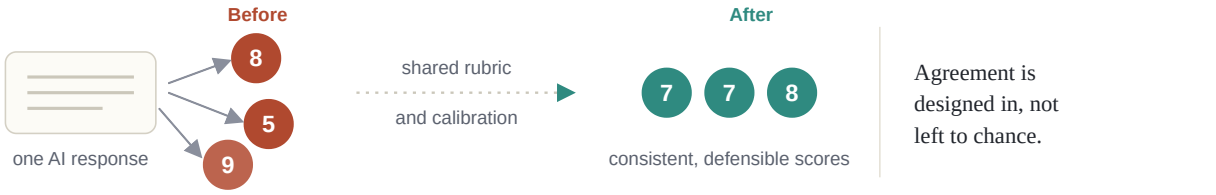
Nishita Grace Isaac, Ph.D.

PROJECT BRIEF · MEASUREMENT & EVALUATION DESIGN

Designing an AI Evaluation Methodology

I built a measurement system that lets two different evaluators reach the same judgment about the same AI response, and gives them the language to explain how they got there.



01 The problem I set out to solve

When organizations ask people to rate AI responses, the same answer can come back as an 8, a 5, and a 9. The gap is usually not carelessness. It happens because evaluators carry different mental models of what a good response looks like. **My aim was to create consistency without flattening judgment**, which meant giving evaluators a shared process, shared definitions, and criteria precise enough that two trained people arrive at the same score on their own.

The most important shift in my thinking was to stop treating the framework as a checklist of things to look for, and to start treating it as a measurement instrument. A good instrument has to satisfy two standards.

VALIDITY
Does each construct actually measure the thing it claims to measure?

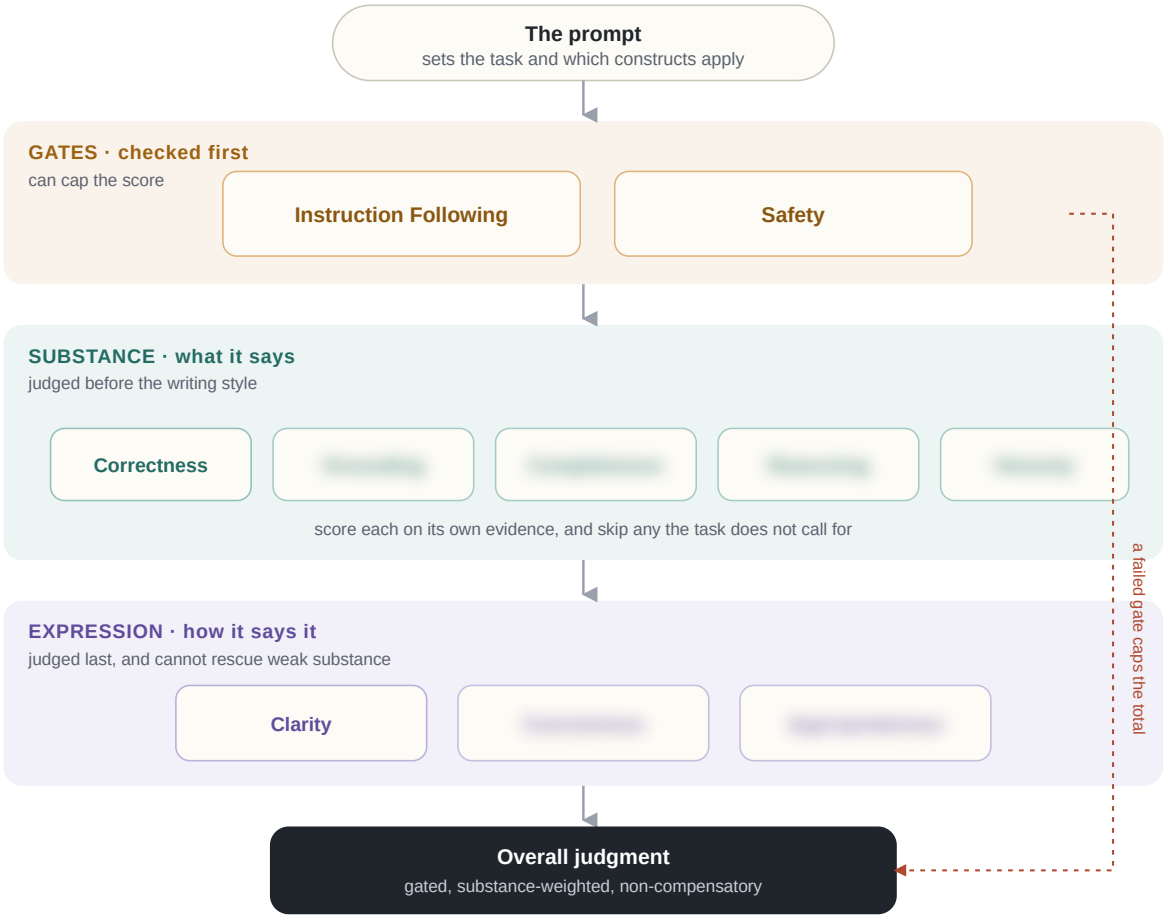
RELIABILITY
Do two evaluators who score the same response arrive at the same result?

This is the same idea behind the APA or Chicago style manuals, which give an entire profession one shared standard so that different people can produce consistent work.

02 What I built

The framework has **ten constructs**, organized into **three tiers**. The tiers also fix the order in which an evaluator works, and they carry real meaning, because they decide which judgments are allowed to override which. The diagram below is a visual representation of the model.

Only 2 of the 8 tier two and tier three constructs are shown here. For access to the full framework, please email me at grace24isaac@gmail.com.



Two rules run through the whole framework. Each construct is scored on its own evidence, so that polished writing cannot inflate a factual score, a bias known as the halo effect. And a construct is only scored when the task actually calls for it.

For a reader who is new to this, here is what the three tiers actually mean, using the constructs shown in full.

GATES

Did it clear the basics before anything else?

Gates are the first checks, and they can limit the entire score no matter how strong the rest of the response is. If a response fails here, good writing cannot save it. This is where an evaluator decides whether the response even belongs in the running.

> **Instruction Following** asks whether the response did what the user actually requested, such as honoring a required format or word count. **Safety** asks whether it avoids harmful content and handles sensitive requests responsibly. A response that answers a different question, or that produces harmful content, is capped right here.

SUBSTANCE

Is the content itself any good?

Substance is about what the response actually says, which means the accuracy and quality of its content. It is judged before the writing style on purpose, so that fluent language does not trick an evaluator into scoring weak content too highly.

> **Correctness** asks a direct question: are the factual claims in the response actually true? It is scored against real evidence rather than the evaluator's memory, and it stays separate from whether those claims were well supported or well argued, which are their own constructs.

EXPRESSION

Is it communicated well?

Expression is about how the response is written rather than what it says. It is judged last, and it cannot make up for weak content. A beautifully written answer that is wrong or off-task still fails on the tiers above.

> **Clarity** asks whether the intended reader can understand the response easily on a first read, taking in structure, wording, and readability together as one judgment about how easy the response is to follow.

03 How I got there

The framework above is the destination. It began as a loose list and became a tested system through seven moves. Each move was a decision, and each decision produced a result I could build on.

1

I started with intuition and made it explicit

I wrote down twelve constructs in four rough groups, along with a step by step workflow.

RESULT A shared vocabulary to build on, though it was still overlapping and untested.

2

I reframed the whole thing as measurement

I brought in construct validity, discriminant validity, the halo effect, and the research on analytic versus holistic rubrics.

RESULT The question changed from what should I look for into what can two people apply the same way.

3

I tested the constructs for hidden overlap

I used dissociation tests and inter-rater logic to find constructs that evaluators could not reliably tell apart.

RESULT I cut twelve down to ten by merging Relevance, Organization, and Precision into their true homes, and by reframing Tone as Appropriateness.

4

I added the construct that was missing

I introduced Honesty and Calibration, which measures whether the confidence a response expresses matches how reliable it actually is.

RESULT The framework could now capture a response that is right but overconfident, and one that is wrong but honest, which nothing else caught.

5

I wrote operational definitions with hard edges

I gave each construct a clear boundary describing what it does not cover, and a rule for when it applies.

RESULT Definitions that can seed a rubric, rather than definitions that simply restate an opinion.

6

I built one construct end to end as a gold standard

I took Correctness all the way to observable indicators, a governing decision rule, behavioral scoring bands, and worked examples.

RESULT A reproducible scoring procedure that became the prototype for every other construct.

7

I froze a reusable template and split the toolkit

I generalized that one specification into a template with five construct families, and I separated evaluator guidance from the deeper methodological justification.

RESULT A system that scales to all ten constructs and supports several downstream tools.

04 The decisions, and the confusions behind them

The hardest work was not adding constructs. It was resolving the tensions where the right answer was genuinely unclear, and being willing to change my own earlier decisions.

- **Are the constructs independent of each other, or do they depend on each other? My own draft claimed both at once.**

  I resolved it by separating two ideas. Evaluators score each construct on its own evidence, which is one kind of independence, while some constructs are still allowed to cap others, which is a kind of dependence. Both statements are true once you name them separately.

- **When does the fact that two constructs overlap actually mean I should merge them?**

  I learned that a high correlation is not enough on its own, because trade-offs and the halo effect also produce correlation. The real test is whether two raters can assign different scores for stated reasons. Clarity and Organization are still flagged as needing testing for exactly this reason.

- **A confidently stated but contested claim feels wrong. Is that a Correctness error?**

  No, and penalizing it under Correctness would let confidence leak into a judgment about truth. The decision was to score only the claim the response actually asserts, and to let its confidence be judged separately under Honesty. Holding that line is what keeps the two constructs clean.

- **Is "not applicable" the same as "I could not verify it"?**

  I separated the two. A response with no factual claims and a response whose claims I simply could not check are different outcomes, and collapsing them would quietly corrupt the reliability statistics.

- **Was my prototype even representative of the other constructs?**

  It was not. Correctness has a rare external ground truth, so the template over-fit to fact-checking machinery. I caught this and revised my own plan, choosing to build the next specification from Clarity, which sits at the opposite pole, so that the template would generalize.

05 From constructs to rubrics

A definition alone does not make two people agree with each other. A well built rubric is what actually produces that agreement, and three moves did most of the work. I use Correctness as the worked example throughout.

A single master rule that every hard case falls back on

I gave each construct one master rule that an evaluator can always return to. For Correctness, that rule is to score only what the response actually claims is true, rather than what it hints at or how confident it sounds. When an evaluator meets a tricky case I never anticipated, they do not need a brand new rule for it. They apply the master rule, and the answer follows. This is what keeps different evaluators consistent, even on cases the manual never listed.

Behavioral bands, not vague adjectives

A score of 4 does not mean "good." It means every central claim is true and there is exactly one minor error. Describing each score in terms of what an evaluator can actually observe and count makes the score hard to argue with, which is what drives agreement up.

Examples that explain the borders, and reference cases that settle them

For every example, I explain not only why it earns its score, but also why it does not earn the score just above or just below it. Most disagreement between evaluators happens at these borders, for instance between a 3 and a 4, so showing exactly where the line falls is what keeps scores aligned. The hardest examples become official reference cases. Once the team decides how a difficult case is scored, that decision becomes the standard everyone follows afterward, much the way a past court ruling guides later ones.

06 Built to scale

Writing ten specifications by hand would invite them to drift apart. To prevent that, I grouped the constructs into five families based on how they are measured, meaning what the response is judged against, the unit being scored, and the scoring model. That grouping is what lets a single template serve constructs as different as Correctness, which is checked against the world, and Clarity, which is judged against an imagined reader. To keep this in step with the framework above, the constructs I left blurred there stay blurred here as well.

FAMILY	CONSTRUCTS	JUDGED AGAINST	UNIT	SCORING
Compliance	Instruction Following, Safety	An explicit spec or hazard list	Constraint	Pass check, then cap
Veridical	Correctness, [illegible]	External evidence	The claim	Bands, plus "cannot verify"
Analytic	Reasoning, [illegible], [illegible]	Internal norms	Step, scope, or sentence	Bands
Normative	Clarity, [illegible]	An imagined reader	The whole response	Bands
Relational	Honesty and Calibration	Another construct's output	Claim and its hedging	Bands, inherited

The methodology is authored as structured data, so one source can render into an evaluator manual, a calibration guide, a scoring card, and an interactive tool, with no hand-maintained copies that can fall out of sync.

COMPANION PIECE

Alongside this written methodology, I designed **Lattice**, an interactive toolkit that walks an evaluator through the analysis, applying the workflow, prompting for evidence at each construct, and surfacing the decision rules in context. The methodology is the engine, and Lattice is the product built on top of it.

07 What I would do next

I would freeze the template against Clarity, which is the construct least like Correctness, and then specify the hardest construct, Completeness, while the template is still cheap to change. After that I would run an empirical pilot with two raters and roughly fifty responses, measuring inter-rater reliability and using a design that separates real agreement from agreement that is only a halo artifact.

If there is one thing this project shows about how I work, it is that I take a fuzzy human judgment and engineer it into something testable, and I am willing to revise my own decisions when the evidence shows that the first version was over-fit.