

THE BLACK BOX.

Creative variance is not a leaderboard. It is an invitation to make judgments inspectable.

A practical field guide to blind comparison, evaluation rubrics, prompt histories, failure archives and the limits of what a demonstration can claim.

ORIGINAL SAHILVERSE FIELD EDITION / 26 OCTOBER 2026

Creative project author: Sahil Tayal. This publication accompanies blackbox.sahiltayal.online. Every sample model identity and output shipped with the exhibition is illustrative, not an attributed model run. No live API calls or measured benchmark results are represented.

01 / What you are looking at

The exhibition holds one shared prompt and four independently authored, illustrative responses. Every specimen uses the same nominal task, but the responses exhibit different design priorities. The website allows you to judge those specimens, edit them, add new ones, preserve changes and export your local record.

Five things the demonstration does NOT establish

It does not establish that any real AI system wrote an entry; that any model performs better than another; that a subjective rubric has external validity; that the number of browser votes equals the number of people; or that a written feature description proves working code. These boundaries are central to the experiment, not fine print.

A useful distinction

Creative variance is difference among outputs. **Reliability** concerns performance under defined and repeated conditions. **Validity** concerns whether a test measures what it says it measures.

Reproducibility requires accessible conditions and artifacts. A striking example can demonstrate a concept while proving almost nothing about the distribution of model abilities.

02 / The six-lens rubric

Score each criterion from 1 (little evidence of meeting it) to 5 (strong evidence in the artifact). Score what is observable, not what you hope exists. The interface calculates a plain arithmetic mean for display; equal weighting is a convenience, not a scientific model.

Novelty	Generic pattern	One distinctive idea	New mechanism with purpose
Coherence	Internal conflict	Mostly consistent	Parts reinforce one another
Usability	Action unclear	Usable with friction	Clear, complete journey
Reproducibility	Conditions absent	Partial instructions	Repeatable inputs and steps
Accessibility	No alternatives	Partial access routes	Keyboard/text/motion accounted for
Functionality	Unimplemented claim	Partial working proof	Feature tested and working

03 / Guided experiment: Blind the brand

Question: Does seeing an identity label change your expectations of the output? This single-person demonstration cannot isolate brand effects, but it can reveal your own decision process.

Protocol: Read the common prompt and constraints. Activate Blind identities. Assess each output using the six rubric lenses. Write one sentence supporting every extreme score (1 or 5). Only then reveal labels. Because writing style and title remain visible, this is label-blinding, not complete concealment.

Reflection: Where did you replace evidence with a brand story? Did you score the same design criteria consistently? What additional controls would a proper study need (random assignment, independent raters, multiple prompts, pre-registered outcomes)?

04 / Guided experiment: Change one constraint

Question: How does an output look when the definition of success changes?

Protocol: Commit a new prompt version that explicitly requires an accessible keyboard-only path and a stop condition after three actions. Do not pretend existing entries have regenerated. Instead, annotate each existing specimen against the new version, revise one manually, and compare its scores before and after. Export the dataset to preserve your record.

Reflection: If a response becomes less visually spectacular but more usable, is it worse or better? State the value judgment instead of smuggling it into a score.

05 / Guided experiment: The failure ledger

Question: What changes when we stop hiding failed promises?

Protocol: Identify a specimen that mentions an implementation without showing executable evidence. Archive the gap under Nonfunctional feature or Unverifiable provenance. Write a clear observation, not an accusation. Suggest one test that would close the gap.

Reflection: Distinguish a design idea, a prototype, and a tested deployed feature. Remember that a charming description is not a working application.

06 / A failure archive worth keeping

A useful failure note names the expected behavior, the observed behavior, relevant conditions, and what evidence would change the assessment. It should neither conceal errors nor exaggerate what was tested.

Example record A: Accessibility

Expected: A visitor can complete the journey using only a keyboard. **Observed:** The text describes clickable cards but does not supply keyboard semantics or a tested build. **Conclusion:** Accessibility is not established. **Next verification:** Run actual tab-order and keyboard-operability tests on an implemented artifact.

Example record B: Provenance

Expected: A real model run is traceable. **Observed:** A provider name appears with no run identifier, date, model configuration or raw response archive. **Conclusion:** Do not attribute the sample to that provider. **Next verification:** Collect consented logs of real requests and document parameters and processing.

Example record C: Reproducibility

Expected: The result can be repeated. **Observed:** The prompt is given, but retrieval, sampling settings, tools and runtime versions are missing. **Conclusion:** Replication conditions are incomplete. **Next verification:** Record fixed inputs, dependency versions, timestamps, model build identifiers and a clear test procedure.

07 / Suggested field notebook

Prompt version: _____ Constraints: _____

Specimen / anonymous ID: _____ Date: _____

Observed strength: _____

Observed failure: _____

Novelty: ____ Coherence: ____ Usability: ____ Reproducibility: ____ Accessibility: ____ Functionality: ____

Which score has the weakest supporting evidence? _____

What would change my judgment? _____

08 / Interpret results without myth-making

The output score is a description of one evaluator's judgment on one artifact under one declared rubric. It is not a population statistic. An average hides the dimensions in which outputs differ: a concept with high novelty and low accessibility may share the same mean as a practical but conventional one. Preserve dimension-level results and reasoning.

If studying real generators later, design the evaluation before collecting results. Use multiple prompts sampled from a defined task domain; several independent judges; a plan for disagreement; a blinded and counterbalanced presentation order; honest logging of model versions, sampling parameters, retries and exclusions; and functional tests rather than prose-only assertions. Report uncertainty and limitations.

Human stakes

Quality is contextual. A playful art experiment and a public service tool have different failure costs. Novelty may be an artistic virtue; accessibility and reliability are obligations when people rely on a service. A single universal score can erase those tensions.

Privacy and local votes

This edition stores votes, prompt history, and notes only in browser localStorage. A browser vote is not a unique human participant. A shared device, repeated clicks, clearing local data, or importing saved data can change the tally. Do not publish it as a public preference poll. JSON export gives you portability but also responsibility: review records before sharing them, especially if you enter personal information.

09 / Further reading and intellectual lineage

NIST AI Risk Management Framework (AI RMF 1.0, 2023): governance, measurement and responsible AI risk management. <https://www.nist.gov/itl/ai-risk-management-framework>

Mitchell et al. (2019), Model Cards for Model Reporting: standardized documentation of intended use, performance context and limitations. <https://arxiv.org/abs/1810.03993>

Geburu et al. (2021), Datasheets for Datasets: documentation of dataset motivation, composition, collection and use. <https://arxiv.org/abs/1803.09010>

W3C WCAG 2.2 (2023): accessible interface success criteria, including keyboard operation and focus appearance. <https://www.w3.org/TR/WCAG22/>

ISO/IEC 25010 quality model: a framework for thinking about software product quality characteristics. Consult the standard for formal definitions.

THE TAKEAWAY: A black box becomes more accountable not when we attach a winner's ribbon, but when we expose the conditions of its judgment.