Evaluation run comparator

Fictional sample. Requires tailoring and responsible review.

Document control

Field	Value
Document type	Evaluation run comparator
Document reference	12-evaluation-run-comparator
Jurisdiction	United States
Version	2026-10-07.statistics.1
Status	Within entered threshold
Classification	Fictional sample. Requires tailoring.
Prepared by	Customer prepared
Prepared on	October 7, 2026
Sources reviewed	September 24, 2026
Generated	October 7, 2026

Purpose and decision boundary

Compare baseline and candidate evaluation pass rates, show the percentage-point change and test it against a customer-entered review threshold.

Use comparable cases and identical acceptance criteria. The result does not establish model quality or authorize release.

Entered counts and review thresholds

Input	Value	Basis
Baseline cases	500	Enter the number of evaluated cases in the approved baseline run.
Baseline cases passed	460	Enter cases that met the baseline run’s stated acceptance criteria.

Input	Value	Basis
Candidate cases	500	Enter the number of evaluated cases in the candidate run.
Candidate cases passed	451	Enter cases that met the candidate run’s same acceptance criteria.
Maximum allowed regression in percentage points	2	Enter the organization’s review threshold. This is not supplied as a legal or technical standard.

Calculation findings

The candidate pass-rate change is within the entered regression threshold.

Displayed percentages are rounded to two decimal places. Gates use unrounded expressions.

Measure	Result
Baseline pass rate	92.00%
Candidate pass rate	90.20%
Candidate change	-1.80 percentage points
Entered maximum regression	2.00 percentage points

Entered review gates

Gate	Result	Basis
Regression limit	Met	Exact count cross-products against the entered decimal threshold. Inclusive boundary.

Method and precision

- Pass rate = passed cases divided by total cases.

- Change = candidate pass rate minus baseline pass rate.

- Within threshold when change is not lower than the negative entered threshold.

- Counts and ratio gates use exact integer cross-products.

- JSON and the workbook retain unrounded numeric observations. Those are floating-point approximations. Gate decisions use exact count and decimal expressions.

Responsible review

- Use comparable cases and identical acceptance criteria. The result does not establish model quality or authorize release.

- Confirm population completeness, labels, sampling and comparability against retained source evidence. This calculator accepts entered aggregate counts and does not inspect individual records.

- Customer-entered gates do not establish statistical significance, fairness, legal compliance or release approval.

Official sources

Source	Provision	Date	Link
National Institute of Standards and Technology: NIST AI Resource Center	Testing, evaluation, verification and validation resources. This source supports the Evaluation run comparator scope described on this page. It does not determine applicability or approve the customer result.	September 24, 2026	https://airc.nist.gov/

Source	Provision	Date	Link
National Institute of Standards and Technology: Artificial Intelligence Risk Management Framework 1.0	Govern, Map, Measure and Manage. This source supports the Evaluation run comparator scope described on this page. It does not determine applicability or approve the customer result.	September 24, 2026	https://www.nist.gov/itl/ai-risk-management-framework
National Institute of Standards and Technology: Generative Artificial Intelligence Profile	Generative AI risks and suggested actions. This source supports the Evaluation run comparator scope described on this page. It does not determine applicability or approve the customer result.	September 24, 2026	https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf