APPLIED AI FIELD NOTES

EXECUTIVE BRIEF

JUDGING AN AI AGENT:

COMPARATIVE VS. RUBRIC-BASED EVALUATION

Tom M. Gomez

June 01, 2026

Luminity Digital

INTELLIGENCE, ENGINEERED. DELIVERED.

INTRODUCTION AND IMPORTANCE

HOW TO JUDGE AI AGENTS

Purpose

- Set clear, practical standards for judging AI agents in business workflows.

What we are comparing

- Rubric scorecards: evaluate performance against pre-set criteria.
- Side-by-side runs: decide which agent output is **better**.

Decision focus

- Choose the method that best supports your approval and rollout decisions.

STAKES RISE WHEN AGENTS ACT

Agents change real things

- An agent can act in business systems, not just respond.
- Those actions can create work, costs, and consequences.

Actions span multiple tools

- It may read files, use tools, message people, or start workflows.
- Small missteps can compound across steps and handoffs.

Quality is a trust decision

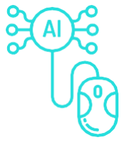
- Evaluation must cover **safety**, reliability, and business judgment.
- Clear standards make deployment decisions easier to defend.

EVALUATION METHODS AND INSIGHTS

RUBRIC VS. A/B

Rubric (scorecard)

- Define criteria upfront and score each run against them.
- Good for repeatable decisions and clear pass or fail thresholds.



Comparison (A vs. B)

- Run two versions on the same task and pick the better one.
- Good when quality is holistic and hard to score precisely.

Management question to answer

- Rubric asks: Did it meet our standards on key dimensions?
- Comparison asks: Is the new approach better than the old?

RUBRICS PINPOINT GAPS

Set expectations upfront

- Agree on what “good” means before anyone runs the test.
- Use criteria that match business risk: accuracy, safety, speed, and cost.

Score the parts, not just the total

- Rate each criterion separately instead of one blended overall grade.
- This makes trade-offs visible, like speed versus safety.

Turn results into next steps

- Find the weakest dimension quickly, then focus fixes there.
- Track the same rubric over time to prove improvement.

OUTCOME VS. TRAJECTORY

Outcome: did it work

- Confirm the job was completed correctly, safely, and as expected.
- Include practical checks like accuracy, risk controls, speed, and cost.

Trajectory: was the path OK

- Review whether tool choices, inputs, and sequence were appropriate.
- Flag wasteful steps or unsafe detours, even with a good result.

Why you need both

- A strong evaluation looks at the **result and process**, not just one.

A VS. B COMPARISON

A simple decision

- Run the same task twice and choose which result you would ship.
- This avoids debates about what a “7 out of 10” means.

Best for “executive-level” quality

- It captures usefulness, judgment, and tone in one call.
- It handles unclear requests where no single perfect answer exists.

How teams use it

- Compare old versus new versions to confirm the change is better.
- Use multiple reviewers to reduce one-person bias in picks.

MANY VALID PATHS

Different routes can still succeed

- Two strong runs may use different steps, tools, or order.
- Treating one “expected path” as truth can mislabel success.

Implication for evaluation

- Use comparisons when multiple approaches can be equally acceptable.
- Design rubrics to allow flexibility, not step-by-step lock-in.

MATCH METHOD TO DECISION

Choose a rubric when you need proof

- Use it for compliance and safety checks that must be auditable.
- Use it for regression testing so releases do not drift
- Use it for root-cause diagnosis when performance drops unexpectedly.

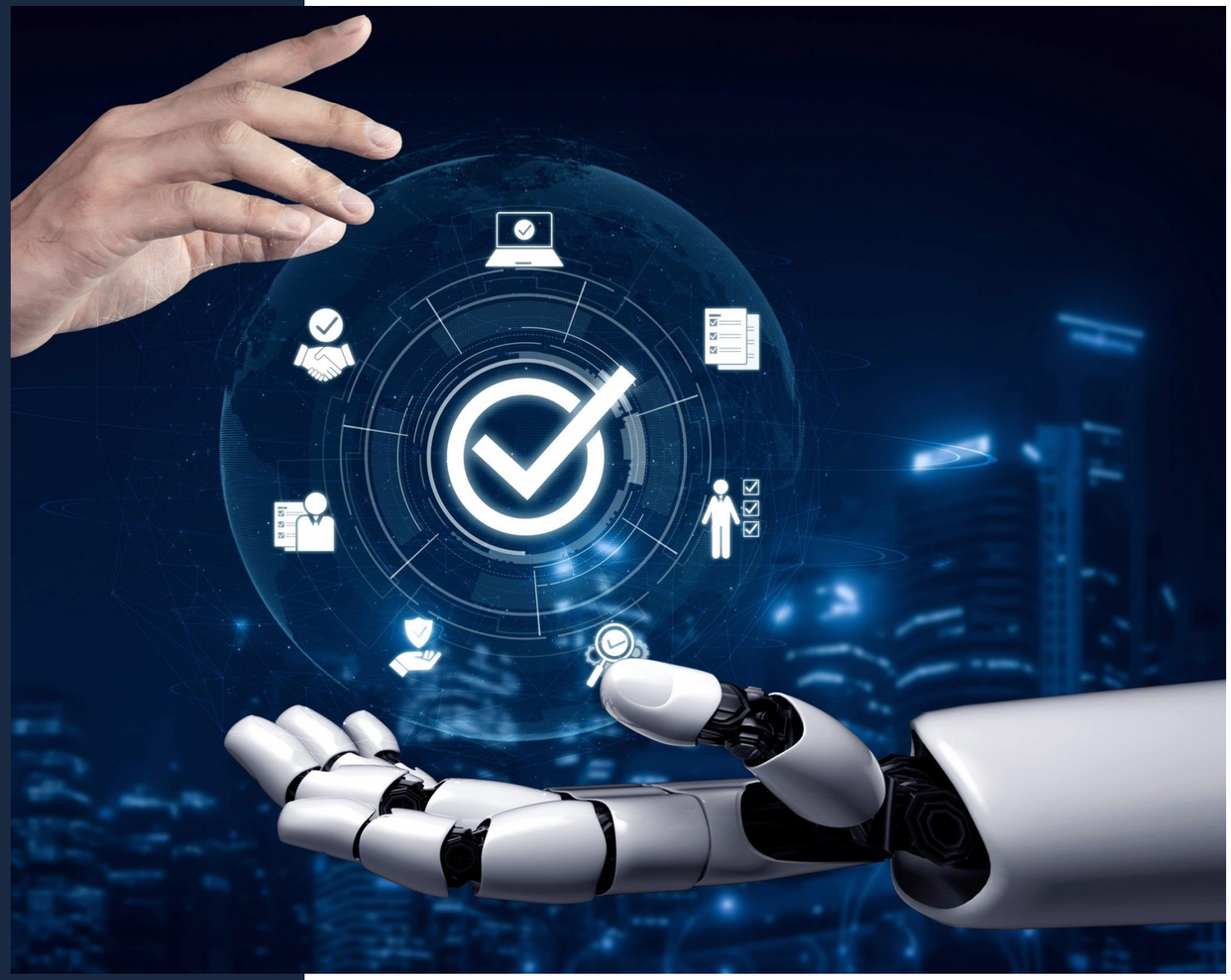
Choose comparison when you need a winner

- Use it for overall quality when criteria are hard to pin down.
- Use it for version decisions: is the new build **better than** the old?

Let the business question drive it

- Start with what you will decide, then pick the evaluation method.

APPROACH & REFERENCES



USE ALL THREE CHECKS

Code checks: Did it work?

- Use pass or fail tests to confirm the job was completed.
- This is your baseline before debating quality or style.

Rubrics: Did it meet standards?

- Score key criteria like safety, cost, speed, and tool choices.
- The breakdown shows what to fix, not just a number.

Comparison: Is it better now?

- Run A versus B on the same tasks and pick a winner.
- Use it to decide if the new version is worth shipping.

REFERENCES

Mapping the empirical evidence. Discovering the patterns that matter for Agentic AI in production.

Read the detailed insights on the Luminity Digital Blog https://www.luminitydigital.com/insights

Subscribe to our Weekly Newsletter | **Follow us on LinkedIn** - https://www.linkedin.com/company/luminity-digital

Sources used for this overview:

- Anthropic guidance on designing evaluations for AI agents.
- Evidently AI on using LLMs to judge outputs consistently.
- Masood, A. (2026) — Rubric-Based Evaluations & LLM-as-a-Judge
- Evaluating AI Agents — Tools for Smarter Performance Analysis
- Google Cloud — Evaluate Gen AI Agents (Vertex AI Trajectory Metrics)
- Ding, L. (2026) — AdaRubric: Task-Adaptive Rubrics for LLM Agent Evaluation (arXiv)

How to use these references

- Treat them as starting points, then tailor to your risk and goals.