Microsoft 365 Copilot

A Framework for Evaluating AI Quality and Value in the Enterprise

Note: This document reflects Microsoft's current perspective on evaluating AI systems in enterprise environments and is provided for informational purposes only. It does not constitute a guarantee of product performance, results, or business outcomes. Any examples, scenarios, or evaluation approaches described are illustrative, and actual outcomes may vary based on implementation, data, configuration, and organizational context.

Enterprise adoption of AI is accelerating, but adoption alone does not equal impact. Organizations need a shared evaluation method for assessing whether AI is helping individuals, teams, and organizations move work forward: for example, whether it can help reduce friction, support higher-quality outputs, and operate within the governance boundaries their business requires. Without that shared language, conversations about AI quality stay anecdotal, and investment decisions stay uncertain.

This is especially true as AI evolves beyond chat-based assistance into supporting long-running workflows: retrieving information, synthesizing documents, creating artifacts, coordinating across people and systems, and assisting with tasks across multiple steps — depending on the scenario and configuration. The measurement framework must reflect that full arc, from intent to outcome, not only prompt to response.

This framework evaluates AI systems across four primary outcome dimensions: Speed, Accuracy, Trust, and Cost (SATC). Together, SATC gives decision-makers a clear way to evaluate what matters most: whether AI helps users reach useful outcomes efficiently, produces responses that are accurate and reliable, within the security and governance boundaries their organization requires — and how the cost of that investment translates to measurable business value.

Speed — Time to Value

Speed measures whether AI can help shorten the full arc from intent to outcome. It's not about response latency, but the total time knowledge workers spend moving through successfully completing a work-related task. From simple to complex, this can involve activity like switching between apps, files, and conversations, engaging different stakeholders, tracking down the latest version of a document, pulling context from multiple discussions, and assembling all of that into something a user can act on, send, or build from. When an AI system is fast, it compresses the distance between what a user needs and what they walk away with — whether that's a drafted email, a populated report, a research synthesis, or a completed workflow.

With more complex tasks, users are also willing to wait longer if the outcome is better. This is why speed is measured to outcome, not to output. A fast response that requires three follow-up prompts, manual correction, or a restart from scratch has added friction. Speed adds value when the tool helps produce a high-quality output that advances the user's work.

In practice, this means evaluating how much effort one must invest to get value from AI: the context they need to assemble, the apps they need to switch between, the documents they need to manually upload, the ongoing fact checking, the reformatting they need to do after the fact — and, critically, the degree to which an AI tool handled the work rather than handing it back. The less friction between the intent and a usable outcome, the more speed a user experiences.

Accuracy — Response Quality

Accuracy, as a SAT dimension, is broader than factual correctness alone. It measures response quality: whether the work AI produces is perceived as faithful to the underlying facts, complete in its scope, relevant to the task at hand, and useful in advancing the user's goal.

One key input to this evaluation is the ACRU response-quality framework. ACRU isolates the four dimensions that evaluate the substance of an individual AI output:

- **Accurate** — Reflects factually correct information, free of hallucination.
- **Complete** — Covers the full scope of what the user needs without omission.
- **Relevant** — Addresses the actual task rather than an adjacent one.
- **Useful** — Advances the user toward their intended outcome, whether that's a decision, a deliverable, or a delegated action.

When an individual element fails – such as receiving an inaccurate, incomplete, irrelevant, or unhelpful answer – the user's confidence erodes, regardless of how polished the result looks. When these elements are present together, the output is more likely to be useful and support trust.

Note: ACRU evaluates the substance of what AI produces. It does not capture engagement or perceived intelligence, which can diverge from actual response quality. A strong output can be evaluated across all three vectors — substance, engagement, and perceived intelligence — while recognizing that they are not always the same.

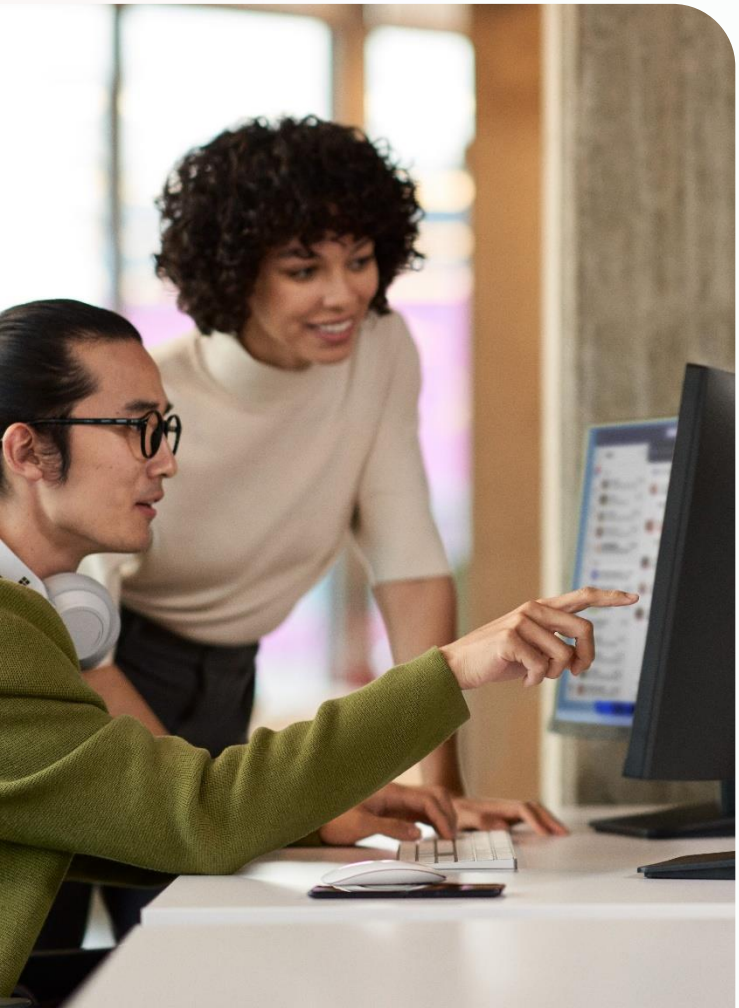
Trust — Governance and Transparency

Trust in enterprise AI is multidimensional. First, it includes security: does the tool respect sensitivity labels, honor access permissions, and adhere to data loss prevention policies? These controls are a foundational enterprise requirement. If an AI system surfaces content a user should not be able to access, or does not honor applied protections, trust in the system may be significantly undermined. In that sense, governance trust is often evaluated as a pass/fail requirement.

Second, trust includes perceived reliability. Does the system cite sources so users can verify information where appropriate? Does it acknowledge uncertainty rather than project false confidence? Does it produce stable, non-contradictory outputs over time? Does it communicate in an appropriate enterprise tone? Equally important is whether AI systems honor enterprise-designated sources of truth — for example, when a user asks for HR guidance on vacation policy, the system should direct the user to the HR-sanctioned guidance or source.

Third, trust is safety and responsible behavior under real-world conditions. It reflects whether the system maintains its guardrails under pressure — resisting prompt injection, jailbreaking, and other adversarial inputs, and avoiding harmful or policy-violating outputs. It also captures alignment to enterprise Responsible AI standards so the system can be evaluated not only in routine scenarios, but also in ambiguous or higher-risk ones. As AI systems support more complex workflows, organizations may also evaluate how clearly the system communicates its plan, limitations, and validation points before users rely on the outcome.

These perceptual dimensions build — or diminish — the kind of trust that determines whether any role in an organization relies on an available AI tool or routes around it. Governance trust establishes the baseline; perceived trust influences continued use.



Cost — Value per Outcome

As enterprise AI capabilities advance, the work AI supports is becoming more varied. Some tasks are quick and lightweight — a draft, a summary, a question answered. Others may run for longer periods, reach across systems, reason over larger amounts of data, and support multi-step workflows. These scenarios vary in duration, complexity, and compute intensity.

Against that backdrop, enterprise AI may be delivered through different pricing models, including subscription and usage-based approaches. Subscription models can cover everyday productivity and prompt-based interactions at a more predictable per-user price. Usage-based offerings can extend that foundation for more advanced, multi-step workflows that require incremental compute and orchestration as usage scales. In this context, cost becomes a measurable dimension of AI quality alongside Speed, Accuracy, and Trust.

Under usage-based billing, the cost of a query can differ based on the mix of models, context, tools, and orchestration it requires. Questions leaders may consider include:

- Does the system integrate with the organization's data, memory, and context in a way that supports the intended use case?
- Does it provide flexibility in how different tasks are handled across scenarios?
- Can administrators forecast, govern, and control spend — for example, by setting limits and policies at the user, group, or tenant level — so consumption remains manageable as adoption scales?

A lower-priced interaction that produces an unusable output may not represent good value. In some scenarios, a higher-cost workflow that helps complete a deliverable across multiple steps — with appropriate accuracy, security, and efficiency — may prove more cost-effective overall. As organizations consider cost in their AI buying decisions, they may also consider the total cost of the work involved and the value of the outcome sought.

The Value Model

Organizations can use this evaluation framework to move beyond anecdotal feedback and make informed decisions about where AI is driving real business impact. By consistently measuring speed, accuracy, trust, and cost across workflows, leaders can start to identify where AI is accelerating time to value, improving output quality, operating within required security boundaries, and delivering efficiently relative to the work performed —and where gaps remain. This helps enable more targeted investments, clearer accountability, and a connection between AI adoption and measurable outcomes such as productivity gains, reduced friction in workflows, and higher confidence in enterprise-wide usage. Observed outcomes will vary by use case, data, configuration, and organizational context.