Questions Industry Leaders Should Ask When Evaluating AI Tools

(Especially Generative and Agentic AI)

Introduction

This artifact exists to help decision-makers evaluate AI systems beyond capability demonstrations. It focuses on governance, control, and accountability under real-world conditions. The questions are designed to surface whether authority and responsibility remain with a human decision-maker, whether systems behave safely when they should not be used, and whether introduced risks are observable before harm occurs.

These questions are not about capability. They are about control, risk and accountability, and whether those risks are acceptable in your operating environment.

The One Question That Matters Most

If time is limited, and you can only ask one question…ask this:

> ***How does this system behave when it should not be used?***

This question exposes authority, stopping behavior, uncertainty handling, escalation, and drift all at once. Everything else flows from that.

I. Authority and Control

- Who retains decision authority at all times? (e.g., designated human operator or authorized decision authority)
- Can a human override or halt the system immediately, without difficulty?
- Does the system ever act without explicit human direction?
- If the system is wrong, who is accountable, clearly and formally?

If authority is ambiguous, the system is not ready.

II. Stopping and Refusal Behavior

- Under what conditions does the system stop instead of responding?
- Can it explicitly say “I don’t know” or “insufficient information”?
- Is refusal treated as correct behavior or as a failure to be engineered around?
- Can stopping behavior be observed and tested in practice?

A system that always answers is a liability.

III. Uncertainty and Confidence

- How does the system surface uncertainty to the user?
- Can it distinguish between high-confidence and low-confidence outputs?
- Does it ever mask uncertainty with fluent or authoritative language?
- Are users guided to recognize unreliable output early?

Confidence without calibration creates false certainty.

IV. Escalation and Human Handoff

- When does the system escalate to a human, and to whom?
- Are escalation thresholds explicit or left to user interpretation?
- Does escalation occur before or after risk is introduced?
- Can escalation decisions be audited after the fact?

"Human-in-the-loop" is meaningless without defined handoff points.

V. Drift and Degradation Over Time

- How does the system behave during long-running sessions or repeated use?
- Can behavioral drift, hallucination increase, or reasoning degradation be detected?
- What happens when context becomes stale, incomplete, or conflicting?
- Is there a mechanism to halt or reset use when drift is detected?

Most AI failures occur over time, not on the first interaction.

VI. Hidden State and Memory

- Does the system rely on hidden memory or internal state not visible to users?
- When prior context is used, is it reintroduced explicitly and transparently?
- Can the system explain what information it is relying on right now?
- Does the system ever treat past output as authoritative truth?

Hidden state creates uninspectable risk.

VII. Portability and Vendor Dependence

- Is governance tied to a specific model, platform, or vendor?
- If the underlying model changes, does behavior change?
- Can the same rules for stopping, refusal, and escalation be enforced elsewhere?
- What actually carries forward: behavior or branding?

Governance that disappears when vendors change is not governance.

VIII. Evaluation Before Adoption

- Can this system be evaluated without trusting vendor claims?
- What observable behaviors demonstrate control, not intent?
- What failure modes has the system been explicitly designed *not* to handle?
- What assumptions are we being asked to accept?

Evaluation should not require belief.

IX. Suitability for Military Use

- Does this system reduce cognitive burden or introduce new ambiguity?
- Does it improve decision hygiene, not just decision speed?
- Does it reinforce or erode command responsibility?
- Would we be comfortable explaining its behavior after a failure?

If you cannot explain it after the fact, you should not use it beforehand.

Bridgewell Advisory is an AI research and advisory firm conducting foundational work in AI governance, assurance, and decision integrity across military, government, and enterprise environments.