Shumaker

FEBRUARY 13, 2026 | PUBLICATION

The Artificial Intelligence Benchmark: The Most Important Clause You've Never Used (Part 1)

You might have noticed, particularly if you watched the Super Bowl this year, that AI is… everywhere.

INDUSTRY SECTOR

Technology

SERVICE LINE

Technology, Data Privacy, Cybersecurity & AI

RELATED PROFESSIONALS

Brian C. Focht

MEDIA CONTACT

Wendy M. Byrne
wbyrne@shumaker.com

AI is now embedded in nearly everything we use. From customer support chatbots and document-drafting tools to cybersecurity platforms, analytics engines, and autonomous "agentic" workflows that can take action within connected systems, it's virtually impossible to avoid.

Despite its ubiquity, many AI contracts treat performance as marketing puffery, with terms like "state-of-the-art," "industry leading," and "human-like" describing tools, not measurable performance promises.

That gap between "promise" and "puffery" matters.

If you are unable to specify how AI will be tested before deployment, after updates, and when conditions change, you are buying puffery. If the expensive tool you purchased fails to perform, it's worthless. Including benchmark testing requirements in contracts is a highly effective method to ensure that AI promises to translate into enforceable results.

Benchmark testing is more than a technical preference—it is the bridge between an AI vendor's aspirations and the need to hold them accountable. It ensures reliable performance in real-world environments and provides meaningful leverage for remediation, service credits, or exit rights if the system falls short. Crucially, it does so before full integration, transforming "trust us" into "prove it"—and preserving that proof as the tool continues to evolve.

A. WHY BENCHMARK TESTING BELONGS IN EVERY AI CONTRACT

Any contract based on an inaccurate understanding of what is being delivered and how the deliverable may change over time is, at its heart, just a bad deal. You can't put a price on a service, software, or platform when the value you obtain is unknown.

AI Performance in a Demo is Not a Metric

AI performance depends on context. Demonstrations usually run on narrow dataset to ensure predictable

results—when they're not running on an entirely pre-written script. Few use "real-world" data, let alone the unique data collections.

A model that seems accurate in a vendor demo can yield very different results on your hardware with your data, terminology, and workflows. Often, a tool's first real test occurs post-deployment, when business processes depend on it. Benchmarking reverses this by demanding the AI meet performance thresholds on your systems and data.

AI Models are Constantly Changing

Benchmarking also matters because AI systems change over time, sometimes in ways that are difficult to detect and beyond your control. Vendors regularly update models, switch between different foundational models (e.g., from ChatGPT to Claude, or vice versa), alter retrieval logic, tune prompts, or reconfigure the system to "improve quality."

Meanwhile, your environment is constantly evolving. Policies are updated, knowledge bases expand, product lines shift, and customer behavior changes—each contributing to potential performance drift. Without a contractual testing framework in place, the burden of detecting model or application drift falls squarely on you. By incorporating drift metrics into benchmark requirements, you enable early detection and treat performance degradation as a defined contractual event, complete with clear obligations and remedies.

Inconsistency Impacts Value

Traditional software contracts typically rely heavily on feature lists and uptime metrics to define the value proposition, which in turn informs the price. AI introduces a different kind of failure to that analysis: a system can be "up" while producing unreliable outputs or unsafe actions.

If the contract doesn't tie acceptance, ongoing performance obligations, and remediation to measurable outcomes, you will have to rely on creating your own workarounds, like adjusting the outputs you receive to account for bias you've discovered. Those are inconsistent and need to be taught to everyone, which can cause even bigger problems if the vendor corrects the bias without your knowledge.

Benchmark requirements make accurate performance a contractual obligation rather than an aspiration, and they give you a clean, objective basis for seeking remedies for inconsistent results under the contract.

Agentic AI Raises the Stakes. All of Them.

While Generative AI offers reviewable output, Agentic AI enables individual agents to perform multiple tasks to reach a goal. It can trigger workflows, create tickets, update records, send emails, schedule meetings, run and modify code, and interact with other tools and AI agents.

The risk shifts from AI providing you a bad answer to performing a bad act.

Benchmarking is therefore necessary but not sufficient. You also need to build tool-use constraints into the agents, rules governing authority and autonomy, error recovery, and primary instructions to "do no harm" when presented with ambiguous or adversarial inputs.

B. BENCHMARKING IS IMPORTANT FOR ALL AI TOOLS AND SYSTEMS

You can be forgiven for thinking that only the "sophisticated" AI platforms need benchmarking. After all, it's human nature to invest more in quality assurance in something that costs you more to use. That assumption

is increasingly risky.

"Basic" or Foundational Generative AI

Even basic generative AI tools for drafting, summarization, and chat can cause serious issues in sensitive contexts. They may misstate obligations, hallucinate facts, or omit qualifiers when drafting customer communication, summarizing policies, or providing HR guidance, risking compliance and reputation. And they do so confidently. We've all seen it.

Benchmark testing for these tools focuses on reliability in the organization's domains, the rate and severity of hallucinations, consistency with instructions and constraints, and how well the system handles requests that should trigger refusals or escalation to a human.

Generative AI benchmarks:

- **Accuracy/factuality** (especially for regulated or client-facing topics)
- **Hallucination rate** (fabricated citations, made-up policies, invented facts)
- **Instruction-following** (does it respect constraints, tone, forbidden topics?)
- **Privacy/confidentiality behavior** (does it leak sensitive content?)
- **Refusal and escalation** (does it appropriately hand off to a human?)

Retrieval-based or Knowledge Assistant AI

When a system includes retrieval (often called RAG, or retrieval-augmented generation), benchmark testing is crucial because the tool's reliability depends on its grounding and citations. Contracts should require testing to confirm the AI stays anchored to approved sources, properly attributes answers, and avoids citing incorrect or outdated materials. A retrieval tool that occasionally cites the wrong policy or sources from restricted folders isn't merely "less accurate." It is quite literally wrong, and being wrong at the wrong time can lead to regulatory issues and lawsuits.

RAG or Knowledge Assistant-specific benchmarks:

- **Citation correctness** (are referenced sources real and relevant?)
- **Grounding** (do answers stay within retrieved content?)
- **Recency controls** (does it flag outdated sources?)
- **Access controls** (does it respect permissions and segmentation?)

Predictive or Scoring AI

AI tools predicting results or generating rankings pose unique risks. In fraud detection and risk scoring, harms stem from false positives/negatives, miscalibrated scores, or bias. Worse, those harms usually go undetected until a negative outcome is disputed.

Benchmark testing here aims to verify measurable model performance within the organization's environment, ensuring that scoring aligns with business tolerances and that monitoring mechanisms are established to detect drift. In regulated or high-stakes settings, the benchmark design should also consider fairness and the capacity to explain outcomes to internal stakeholders, regulators, or affected individuals.

Predictive/scoring AI benchmarks:

- **Precision/recall** (false positives/false negatives)

- **Calibration** (score meaning aligns with real-world probabilities)
- **Bias and fairness** (disparate impact testing where appropriate)
- **Stability** (how sensitive outcomes are to small input changes)
- **Explainability** (as required for oversight)

Agentic AI

With great power comes significantly greater ability to cause catastrophic damage. In agentic environments, benchmarking should cover output quality and safe tool use, including correct tool use, permissions, avoiding irreversible actions without confirmation, and maintaining audit logs. An agent that's 95 percent helpful but five percent reckless can be unacceptable if the five percent includes unauthorized calls, wrong transactions, or destructive changes.

Agentic AI benchmarking (all of the above, plus):

- **Tool-use correctness** (calls the right tools, in the right order)
- **Permission boundaries** (least privilege, no unauthorized actions, no authority elevations)
- **Safety constraints** (never take irreversible actions without confirmation)
- **Auditability** (action and Application Programming Interface (API) call logs, rationales, inputs/outputs preserved)
- **Adversarial resilience** (prompt injection, data poisoning, malicious inputs)
- **Kill-switch and rollback** (fast disablement and recovery)

C. WHAT CAN GO WRONG WHEN BENCHMARKING IS SKIPPED OR MINIMIZED

For the most part, this failure manifests as frustration with an AI tool's functionality and considerable gnashing of teeth over wasted investment. But what if it's worse?

Operational Failure and Customer Harm

Failing to benchmark AI before deployment often results in operational harm and contractual issues. Organizations find the tool performs inconsistently across departments, fails on critical edge cases, or produces errors needing human correction. Poor outputs lead to wrong decisions. Agentic AI errors can execute the wrong actions. Small errors scale up, causing significant issues in areas like customer service, billing, HR, and security.

Legal and Regulatory Exposure

AI outputs utilized in consumer communications, privacy procedures, cybersecurity responses, employment guidance, or other sensitive domains may result in unreliable performance and could potentially violate consumer protection laws, unfair and deceptive practices regulations, anti-discrimination statutes, and contractual obligations with partners, vendors, and customers, as well as sector-specific requirements. Often, the underlying issue is not the existence of AI itself but rather its implementation without appropriate controls aligned with its risk profile.

If the threat of being investigated by numerous federal and state regulatory agencies and being sued by your suppliers, vendors, customers, partners, employees, and shareholders isn't bad enough, how about airing all your dirty laundry?

Leaks of Protected and Confidential Data

Imagine all the ways a human is capable of accidentally exposing your company's confidential information. Now imagine that same human doing the same thing, but one thousand times more often, without sleep or breaks, and that you can't reprimand or fire them. AI can leak confidential data through prompts and uploaded documents, misconfigured access controls, or malicious prompt injection that inserts commands to exfiltrate your information.

Agentic AI introduces an entirely new problem. AI agents are programmed to prioritize completing assigned tasks and will do so even at the expense of other, lower-ranked priorities (such as confidentiality). Combined with the potential to misuse the authority it's been granted, or more terrifyingly, granting itself additional authority, the "complete the task at all costs" approach provides a perverse incentive to the agent to sacrifice confidential information if it will help achieve its goal.

Other Potential Issues

There are also less obvious but significant downstream risks. Generative systems can produce inaccurate, misleading, or policy-violating content. Their output may be non-original or *too similar* to protected material, creating "authority bias" where users trust confident answers. They can generate attribution or validation issues, questioning record integrity and accountability. Without formal performance expectations, organizations might be stuck with a tool that can't meet needs, lacking contractual options for improvement or exit.

(Part 2 will discuss a practical approach to benchmark testing for AI contracts.)

If you have any questions about incorporating benchmark testing, performance metrics, and drift monitoring into your AI vendor agreements, or would like help drafting practical contract language that ties AI "performance" to measurable acceptance criteria, remedies, and auditability, please contact Brian Focht or another member of Shumaker's Technology, Data Privacy, Cybersecurity & AI Service Line.

Shumaker