The Confidence Layer: A Design Philosophy for Trustworthy AI in Evidence Generation

Aide Solutions LLC
April 2026

This article was developed with the support of an AI-assisted writing system designed for scientific and professional communication materials.

Table of Contents

Executive Summary

Most pharmaceutical and health economics research organizations are deploying AI tools at scale while operating under a limited evaluation framework. The dominant approach to assessing these tools, centered on accuracy, sensitivity, and specificity, captures only a fraction of what determines whether an AI system is trustworthy in the HEOR evidence generation context. This paper introduces a design philosophy, called the **confidence layer**, that addresses the structural gap between what traditional metrics measure and what evidence-facing professionals actually require.

The methodological case for rethinking AI evaluation is no longer speculative. **ISPOR's ELEVATE-GenAI** reporting guideline allocates only one of ten evaluative domains to traditional performance metrics; the remaining nine address reproducibility, transparency, data provenance, human oversight, disclosure of limitations, and ongoing monitoring. The **NICE Decision Support Unit's** analysis of frontier AI independently identifies an accuracy-to-impact gap, characterizing technical benchmarks as an unreliable guide to real-world performance. Taken together, these frameworks establish that accuracy is a necessary but insufficient condition for trust, and they point toward a set of system-level properties that conventional procurement and validation processes do not yet systematically assess.

In the meantime, a second force is reshaping the competitive landscape for AI research tools. **Foundation model capabilities have converged substantially**: leading models cluster within a few percentage points of each other on expert-level scientific reasoning benchmarks, while API costs have fallen roughly 80 percent since 2025. Access to frontier-level capability is no longer a procurement advantage. This commoditization shifts the locus of differentiation from the model itself to the architecture surrounding it: the verification logic, traceability infrastructure, and oversight mechanisms that determine whether outputs can be defended under scientific and regulatory scrutiny. Adjacent industries, including cloud infrastructure, financial services risk governance, and autonomous systems safety engineering, followed an identical trajectory once their underlying capability layers commoditized. AI research tooling is following the same path.

The practical consequence of these dynamics has started shaping how senior HEOR professionals actually engage with AI tools. Two questions surface consistently across organizations and tool categories: **1-how will I know when the AI gets it wrong**, and **2-am I outsourcing my expertise?** Both questions highlight genuine design gaps in most current tools. AI systems produce fluent, plausible outputs that present errors with the same confidence as correct results, imposing a uniform verification burden that scales with output volume and largely consumes the efficiency gains the tools are meant to deliver. Separately, the risk of deskilling, of gradually displacing the expert reasoning that gives evidence its credibility, is a structural consequence of tools that position AI as a replacement for judgment rather than an instrument that makes judgment more leveraged.

The *confidence layer* proposed in this whitepaper is the design response to both problems. It encompasses calibrated uncertainty communication that directs expert attention toward consequential decisions, process transparency that makes reasoning chains auditable, human-centered interaction design that preserves and exercises methodological judgment, and monitoring architecture that detects performance degradation across deployment contexts. These are not enhancements to a capable tool; they are the properties that make capability trustworthy in the specific context of evidence generation and communication at any stage of development. Organizations evaluating or building AI research tools should treat these properties as primary procurement and development criteria, not secondary, nice-to-have, add-on features.

Section 1: Where the Validation Gap Leaves Us

The field has reached a point of methodological honesty about AI evaluation that was not available two years ago. The discipline's own standard-setting bodies have now pointed out the challenge that our field has been wrestling with but has not named it explicitly. It is that accuracy, sensitivity, and specificity are necessary conditions for trusting an AI research tool, but they are not sufficient ones. This paper begins from that conclusion and asks what comes next.

The evidence is no longer editorial. ISPOR's ELEVATE-GenAI reporting guideline, published by its Working Group on Generative AI in 2025, structures evaluation across ten domains [1]. Of those ten, only one addresses traditional performance metrics directly (i.e. accuracy assessment). A second domain, model characteristics, includes but is not limited to quantitative performance measures. The remaining eight cover comprehensiveness, factuality, reproducibility & generalizability, robustness, calibration & uncertainty, deployment & efficiency, fairness & bias, and security & privacy. That distribution is informative: a framework designed by and for the HEOR field allocates eighty percent of its evaluative attention to properties that accuracy metrics cannot capture.

The NICE Decision Support Unit's 2026 report on frontier AI reaches a complementary conclusion. Hill and colleagues observe that frontier AI systems may achieve strong benchmark performance while delivering little or no benefit in real-world settings, describing technical metrics as "an unreliable guide to real-world impact" [2]. The report identifies an accuracy-to-impact gap: the distance between what a system demonstrates under controlled evaluation and how it performs in real-world workflows. That gap is not an artifact of poor study design; it is a structural feature of how AI systems behave across diverse deployment contexts.

Table 1. *Dimensions of evaluation insufficiency identified across current frameworks*

Dimension of Insufficiency	What Traditional Metrics Cannot Capture
Output complexity	A single accuracy figure conflates deterministic extraction tasks with design-driven or preference-sensitive modeling work
Environmental variability	Point-in-time measurements do not reflect performance as models update or data distributions shift
Process integrity	Whether outputs are traceable, reviewable, and produced through a defensible reasoning chain
Oversight quality	Whether human review is directed toward consequential decisions or distributed uniformly across all outputs

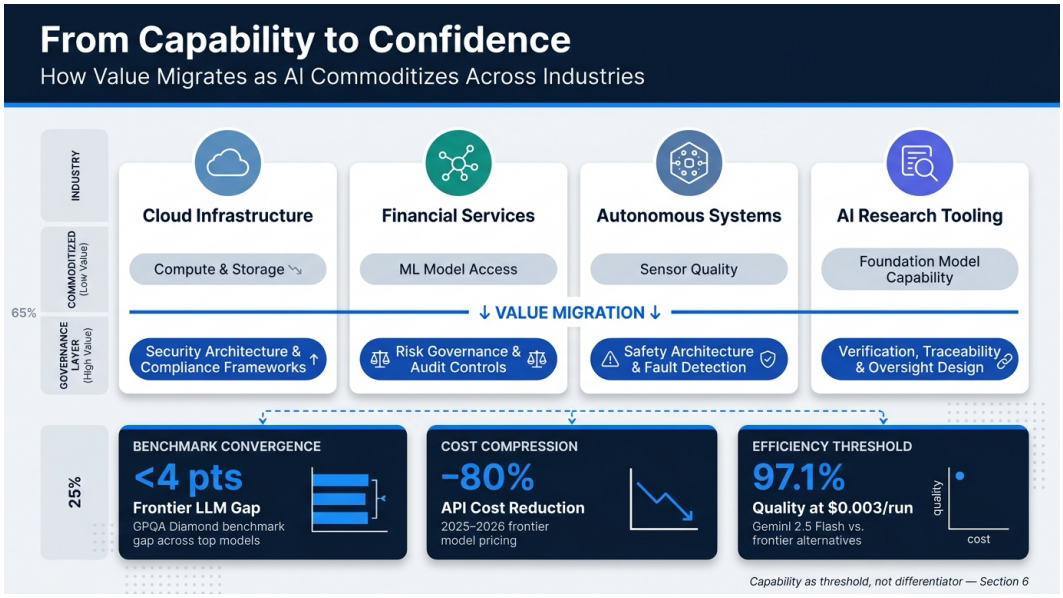
So, then the question is: if traditional metrics are insufficient, what complements them? ELEVATE-GenAI identifies the domains that matter. The NICE DSU identifies the gap. But neither explicitly specifies how to address and measure performance of AI research tools across the full evidence generation and communication lifecycle.

This paper tries to address this gap and proposes that design philosophy under the term ***confidence layer***. The concept is not new to adjacent fields; safety-critical engineering, financial risk systems, and clinical decision support have all developed analogous constructs. Its application to AI-enabled HEOR tools is relatively new, and the stakes are consequential.

The confidence layer complements accuracy measurement rather than replacing it. The framing throughout is one that builds on the promise that well-designed AI systems assist expert judgment and make it more leveraged. The confidence layer is the set of design decisions that makes that augmentation defensible.

Section 2: The Commoditization of Capability

The accuracy-to-impact gap identified in the previous section raises an immediate practical question. If the problem is not simply model performance, what kind of problem is it? Part of the answer is visible in how foundation models have evolved. They have converged in capability and cost to the point where the model itself is no longer the differentiating variable. That shifts attention to what surrounds the model.



From Capability to Confidence: How Value Migrates as AI Commoditizes Across Industries

A Narrow Performance Band

The leading foundation models have converged substantially on scientific reasoning and document-processing benchmarks. On GPQA Diamond, a benchmark designed to test expert-level reasoning on graduate-level science questions, frontier models now cluster tightly: the gap between the second and fifth-ranked models is under three percentage points [3]. On software engineering benchmarks, open-source models now match proprietary frontier models on coding-adjacent tasks, a convergence that would have been unthinkable two years ago [3].

Cost has followed capability. Open-source models now approach proprietary frontier performance while offering dramatically lower costs per token [4]. Gemini 2.5 Flash, benchmarked against 38 real production tasks, delivered 97.1% quality at $0.003 per run [5]. The practical implication is direct: access to a capable model is no longer a procurement advantage. Any team with a modest budget can invoke frontier-level scientific reasoning on demand.

ISPOR's 2026 trends report elevated AI from the third-ranked to the first-ranked trend in HEOR, noting that it "has expanded throughout virtually every aspect of healthcare and HEOR" and that "AI can perform systematic literature reviews in a fraction of the time, structure complex datasets, and accelerate analysis" [6]. The field has moved past debating whether AI is capable. The capability floor is now assumed.

The Harness, Not the Model

Capability convergence creates a diagnostic question: if the models are roughly equivalent, what explains meaningful differences in how AI research tools perform in practice?

The answer, in my assessment, is almost never the model. It is the harness around the model: the structure that channels generation, the verification logic applied to outputs, the interaction design that determines where expert judgment enters, and the audit trail that makes outputs reviewable. A well-harnessed average model outperforms a poorly harnessed frontier model on any task where accuracy, traceability, and oversight matter. In evidence generation, those properties matter on every task.

Most AI research tools available today do not invest seriously in the harness. They are, to use a term that has become common in enterprise architecture, thin wrappers: interfaces that accept a user query, pass it to a foundation model, and return the output with minimal intervening structure. The differentiating work, if any, is limited to prompt engineering and output formatting. From the user's perspective, these tools are largely interchangeable. The underlying model, which is also largely interchangeable, is the primary variable.

This is not a criticism of the tools or their builders. Building the harness is substantially harder than building the wrapper. It requires investment in verification architecture, traceability infrastructure, and interaction design that does not show up on a feature comparison sheet. For a field that has historically evaluated AI tools on benchmark

accuracy, the incentive structure has pointed toward model selection rather than system architecture.

That incentive structure is changing.

Where Value Migrates

Adjacent industries have navigated this transition and offer instructive patterns. In cloud infrastructure, Amazon Web Services and Microsoft Azure achieved effective compute parity by approximately 2015. Subsequent competitive differentiation moved to security architecture, compliance automation, and operational resilience. Raw infrastructure stopped being the competitive advantage; what was built on top of it became the moat [7].

In financial services, no institution today competes on the quality of its machine learning models for risk assessment. The models are roughly equivalent, often sourced from the same providers. Differentiation lies in governance: the controls around model deployment, the audit trails, the risk frameworks that determine when a model's outputs can be trusted and when they cannot [7]. The FS-ISAC principle that "security features should be enabled by default" reflects a broader philosophy: the safety architecture is the product, not the model.

In autonomous systems, sensor quality is not the determinant of market viability. Redundancy architecture, fault detection, and safe-fail mechanisms are. A vehicle with a marginally better camera but inadequate failure detection is less safe than one with commodity sensors and robust safety architecture.

Domain	Commoditized Layer	Value Migration Target
Cloud infrastructure	Compute and storage	Security architecture, compliance frameworks
Financial services	ML model access	Risk governance, audit controls
Autonomous systems	Sensor quality	Safety architecture, fault detection
AI research tooling	Foundation model capability	Verification, traceability, oversight design

AI research tooling is following the same trajectory. The capability layer has commoditized. The tools that will earn and sustain trust in evidence generation are those that invest in what surrounds the model: the structure, verification, and oversight mechanisms that make outputs defensible under scientific scrutiny. That investment is what the rest of this paper addresses.

Section 3: The Two Questions the Expert Actually Asks

When senior HEOR professionals evaluate an AI research tool, the conversation eventually arrives at two questions. They are rarely posed as formal objections; they emerge as

hesitations, as requests for a second demonstration, as a deferral to the methodologist in the room. But they are consistent across organizations, across therapeutic areas, and across tool categories. Understanding them precisely is a prerequisite for understanding what a confidence layer must do.

The two questions are: *How will I know when the AI gets it wrong?* And: *Am I outsourcing my expertise?*

Both highlight genuine gaps in how most tools are currently designed, and point toward how they should be.

The Plausibility Problem

AI language systems produce fluent, coherent, plausible outputs. That is their defining capability, and it creates a specific problem in evidence generation: plausibility and correctness are independent properties. A system can generate a well-structured summary of a clinical trial that misattributes an endpoint, a hazard ratio, or a patient population, and nothing in the surface presentation signals the error. The output looks like correct work because the system was optimized to produce outputs that look like correct work.

Most tools compound this by presenting every output with uniform visual weight. A correctly extracted baseline characteristic and an incorrectly interpolated one occupy the same space on the page, formatted identically, with equal apparent confidence. The expert reading the output has no basis for allocating scrutiny differentially. Reviewing the high-confidence extractions at the same depth as the uncertain ones consumes the time that AI was supposed to free.

The result is a verification burden that scales with output volume. As AI tools produce more, faster, the expert faces more to check, at the same granular level, with the same manual effort. The efficiency gain promised by the tool is largely consumed by the verification cost the tool imposes.

The Institute for Healthcare Improvement's 2024 analysis of AI monitoring in clinical settings documents the structural version of this problem. A readmission prediction model exceeded its AUC thresholds in validation yet failed on real-world outcomes. The infrastructure to detect this automatically, to track performance continuously, segment results, and surface drift as it occurs, does not yet exist as off-the-shelf tooling. Most organizations rely on manual audits and periodic reviews rather than automated dashboards [8]. The HEOR parallel is direct: systematic literature review tools, data extraction platforms, and model-building assistants face the same monitoring gap at the workflow level.

The Expertise Concern, Quantified

The second question is harder to voice in an evaluation meeting, because it carries an implicit admission of vulnerability. If I delegate scientific drafting to an AI, am I still doing the science?

This concern is often dismissed as an expression of professional identity rather than a substantive risk. The evidence suggests it should be taken more seriously.

Shen and Tamkin's 2026 randomized controlled trial assigned 52 professional developers to either AI-assisted or unassisted conditions while learning a new programming library. Participants in the AI-assisted group scored 17% lower on the follow-up evaluation, a difference of 4.15 points on a 27-point scale (Cohen's d = 0.738, p = 0.010) [9]. Crucially, no statistically significant reduction in task completion time was observed, which means the efficiency gain that justified the delegation was largely absent. The cost was real; the benefit was illusory.

The skill dimension with the largest performance gap was debugging. That is precisely the competency required to catch AI errors. The implication is a supervision paradox: as organizations delegate more scientific work to AI, the analytical capacities needed to review that work are precisely the ones eroded by passive delegation [9]. An expert who relies passively on AI output gradually loses the specific analytical capacities needed to evaluate whether that output is correct.

The study's interaction patterns are instructive. Three patterns consistently produced poor learning outcomes: full delegation, progressive reliance, and outsourcing debugging to the AI (quiz scores in the 24–39% range). Three patterns preserved learning: generating code then asking comprehension questions, composing hybrid queries combining code generation with explanations, and posing only conceptual questions while independently resolving errors (65–86% range) [9]. The variable that determined outcome was cognitive engagement, not whether AI was used. Passive use eroded capability; active, directed use did not.

The translation to HEOR is approximate but the direction is clear. A health economist who routinely delegates evidence synthesis to an AI system without sustained analytical engagement risks degrading the judgment that makes expert review meaningful. The concern practitioners voice in evaluation meetings is not abstract.

Table 2. Practitioner concerns and their design implications for AI research tools

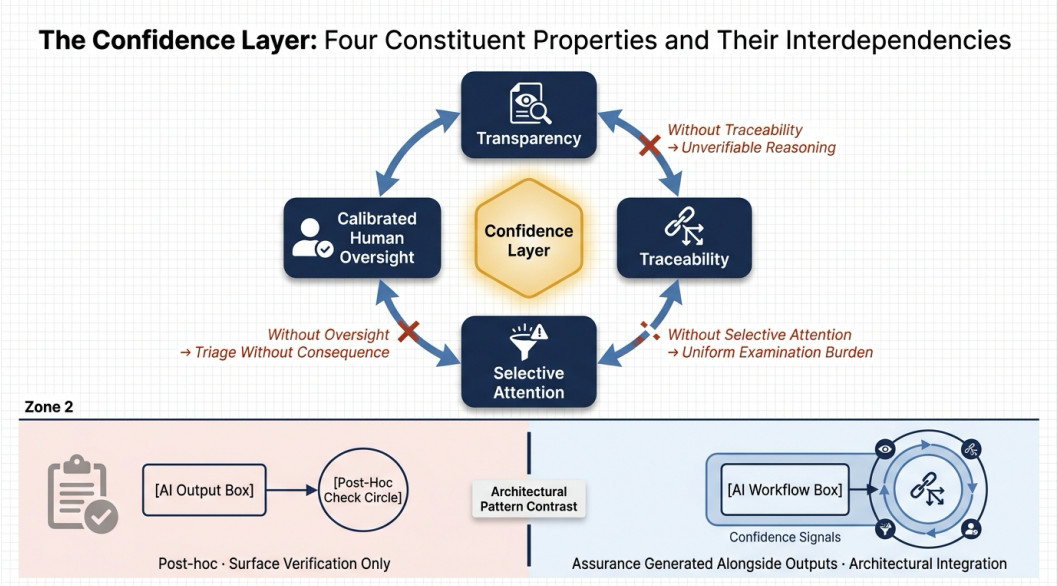
Expert Concern	Mechanism	Design Implication
How will I know when it's wrong?	Probabilistic systems produce plausible errors; uniform output presentation eliminates differential scrutiny	Outputs require confidence signals that direct review toward uncertain elements
Am I outsourcing my expertise?	Passive delegation erodes debugging and conceptual skills (Shen and Tamkin, 2026)	System design must preserve cognitive engagement at consequential decision points

A tool designed around the confidence layer addresses both concerns through a single structural property: it makes the expert's judgment consequential at the points where it matters, rather than optional across everything. The first concern is addressed by signal, not by asking experts to verify more. The second is addressed by design, not by asking experts to be more vigilant. Both require intentional architecture. The next section specifies what that architecture involves.

Section 4: Defining the Confidence Layer

The two expert concerns described in Section 3 share a common structural cause: the AI system offers no reliable signal about where its outputs should be trusted and where they should be scrutinized. This is not a temporary limitation that better models will resolve. Large language models are probabilistic systems. They produce outputs by predicting likely continuations, not by deriving verified conclusions. Errors are a structural property of the architecture, not a defect to be engineered away. The design question, then, is not how to build an error-free model, but how to make the expert aware of where confidence is warranted and where it is not.

The ***confidence layer*** is the set of design decisions, system architecture, and interaction patterns that make an AI system's reliability visible and verifiable with minimal cognitive burden on the expert user. It runs throughout the system rather than sitting atop a finished output. A confidence layer is a property of how the system is built, not a dashboard or score column appended to results.



The Confidence Layer: Four Constituent Properties and Their Interdependencies

Four Constituent Properties

Four properties define a confidence layer in practice.

Transparency. The system makes its reasoning, sources, and decision path visible within the workflow. When a system generates a structured evidence summary, the expert can see which source passages informed which claims, what the system understood about the scope of its task, and where it made inferential connections versus direct extractions. Transparency is legibility in the moment of use, not documentation after the fact.

Traceability. Every output element links back to what informed it. A clinical parameter traces to the specific table from which it was extracted, with page and cell reference. A structural modeling assumption traces to its stated provenance and the expert input that confirmed it. An auditor, a reviewer, or the original user can follow any chain from output to source without reconstruction. This property is what makes AI-generated work defensible in an HTA submission or peer-review process: the reasoning is externalized and verifiable, not locked inside the model.

Selective attention. The system signals where it is on solid ground and where it is uncertain. A high-confidence extraction from an unambiguous table receives a different signal than an interpolated estimate assembled across inconsistent reporting formats. The expert's review effort is directed toward the latter. This is where efficiency gains are genuinely realized: the time freed by AI handling routine extraction is preserved rather than consumed by uniform scrutiny of all outputs.

Calibrated human oversight. The system pauses at junctures where expert judgment is genuinely required: a structural modeling decision, a borderline inclusion call, a high-impact assumption that propagates widely through a cost-effectiveness model. At those junctures, the system presents its reasoning and relevant materials, incorporates the expert's input, and proceeds on that basis. The expert drives scientific decisions; the AI handles labor-intensive steps in between. This is the operational form of augmentation.

These four properties are mutually reinforcing. Transparency without traceability produces visible reasoning that cannot be verified. Traceability without selective attention produces an audit trail requiring uniform examination to use. Selective attention without calibrated oversight creates triage signals without structured consequence. Together, the four properties constitute a system that makes expert review both targeted and consequential.

Embedded Versus Bolt-On

A confidence layer that runs throughout the pipeline is structurally different from one applied after a finished artifact is produced. The distinction matters operationally and for governance.

A bolt-on check applied to a completed output can verify surface properties: whether citations are present, whether values fall within plausible ranges, whether formatting is consistent. It cannot reconstruct the reasoning that produced the output, because that

reasoning was not recorded. The result is a compliance ritual rather than genuine assurance.

An embedded confidence layer generates assurance as part of generation. Traceability records are produced alongside outputs, not derived from them afterward. Uncertainty signals emerge from the same process that produces the outputs they annotate. Human oversight prompts occur at decision points within the pipeline, not as a review step appended to the end.

The NICE DSU's 2026 report frames this distinction at the governance level. Recommendation 12 states: *"Use proportionate conditional recommendation pathways and maintained guidance for high-update frontier AI. Linking access to evidence generation, monitoring and explicit stop criteria allows NICE to support timely innovation while retaining oversight as systems evolve"* [2]. These three elements, continuous evidence generation, conditional access gates, and explicit stop criteria, function as a system. They presuppose an embedded architecture, because conditional access and stop criteria require real-time signal from the pipeline to operate.

The FDA's Predetermined Change Control Plan framework reaches a parallel conclusion from a device regulation standpoint. PCCPs specify planned modifications, the protocols for validating them, and the assessment of their impacts [10]. The governing principle is that permissible changes must be anticipated and their validation protocols specified in advance, not assessed reactively. This is lifecycle governance by design.

Table 3. Architectural patterns for assurance in AI research tools

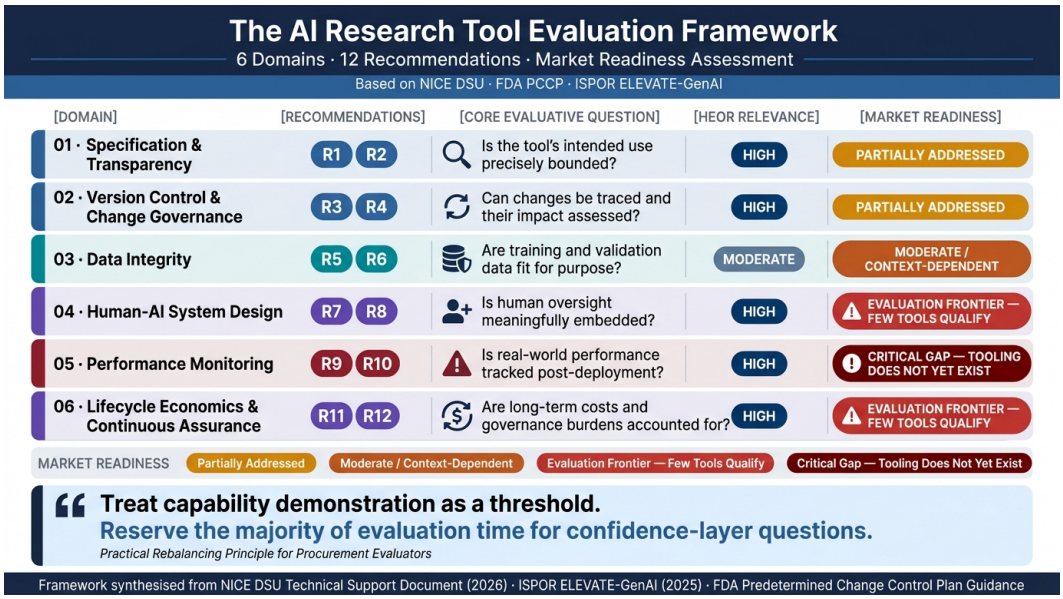
Architecture Pattern	How Assurance Is Produced	Implications for Expert Review
Bolt-on check	Applied to finished artifact; generation reasoning is inaccessible	Surface verification only; reasoning chain is unrecoverable
Embedded confidence layer	Produced alongside outputs during generation	Traceability, uncertainty signals, and oversight prompts are live and actionable

Most AI research tools today follow the bolt-on pattern, because embedding assurance into the pipeline requires architectural decisions made from the outset: how reasoning is recorded, how uncertainty is computed, where expert input is solicited, and what conditions trigger output flagging. The regulatory and methodological frameworks cited here point toward the embedded pattern. The confidence layer, as a design philosophy, describes what a system must do throughout its operation to make expert judgment more leveraged rather than more burdened.

Section 5: A Framework for Evaluating AI Research Tools

The 12 recommendations in Hill and colleagues' 2026 NICE DSU report provide a comprehensive foundation for evaluating AI tools in health technology assessment contexts [2]. This section organizes those recommendations into six thematic evaluation domains, each expressed as a question the tool must be able to answer. The framework applies equally to internal builds and external procurements; the questions are the same and only the interlocutor changes.

One clarification on sourcing: the NICE DSU report does not present these as a pre-enumerated checklist. Section 23.5 proposes that NICE "add a short frontier AI checklist to NICE guidance to ensure consistent reporting," framing this as a call to action rather than an existing instrument. The 12 recommendations are dispersed across the report and mapped here to governance domains. Recommendations 5, 6, 8, 9, and 10 primarily address NHS-context appraisal objects (comparator justification, UK transferability, post-market surveillance linkage) rather than tool design properties. They are included for completeness and annotated accordingly.



The AI Research Tool Evaluation Framework: Six Domains, 12 Recommendations, and Current Market Readiness

The Six Evaluation Domains

Domain 1: Specification and Transparency (Recommendations #1, #4)

Can the tool produce a structured account of its inputs, configuration, intended purpose, and known limitations and failure modes? Recommendation 1 calls for structured

specification of AI system scope and intended use. Recommendation 4 requires disclosure of limitations and misuse risks within the workflow, not confined to documentation.

For HEOR evaluators, this domain is foundational. A tool that cannot articulate its own limitations cannot be used defensibly in evidence synthesis. The operative question is whether disclosure is embedded in the user's working context or deferred to documentation that is rarely consulted during active research.

Domain 2: Version Control and Change Governance (Recommendations #2, #3)

Is every output bound to the specific model versions, prompt versions, and data sources that produced it? When those inputs change, is there a validated protocol in place before the change goes live?

Recommendation 2 addresses version binding: the requirement that outputs can be reproduced and attributed to a specific system state. Recommendation 3 addresses change governance through the lens of Predetermined Change Control Plans, requiring that planned modifications are specified in advance with validation protocols defined before implementation [2]. For tools used in regulatory or HTA submissions, the inability to reproduce an output under identical conditions is a fundamental scientific problem.

Domain 3: Data Integrity and Bias Management (Recommendations #5, #6)

Does the tool disclose the quality and representativeness of its training and retrieval data? Are outputs appropriately qualified when data limitations apply?

These recommendations primarily concern AI systems evaluated for direct clinical use within the NHS context. For evidence synthesis applications, the relevant translation concerns retrieval corpora, literature databases, and real-world data sources used to ground outputs. A tool drawing on a narrow or unrepresentative evidence base without disclosure introduces systematic distortion that accuracy metrics alone will not surface.

Domain 4: Human-AI System Design and Oversight (Recommendations #7, #8)

Is the tool designed so that human oversight is meaningful at consequential decision points, rather than a review step appended to finished outputs? Does transferability analysis address how AI-generated evidence applies to the specific population, setting, or clinical question at hand?

Recommendation 7 is among the most practically significant for HEOR tools. It addresses the sociotechnical system rather than the model in isolation, asking where the human sits in the workflow and whether that position affords genuine influence over scientific decisions. Recommendation 8 concerns transferability: the structured assessment of how AI-generated evidence translates across geographies, populations, and care settings. For pharmaceutical companies generating global evidence packages, transferability determines whether a systematic review or economic model is fit for a specific HTA jurisdiction.

Domain 5: Performance Monitoring and Validation (Recommendations #9, #10)

Does the tool provide a plan for ongoing real-world performance monitoring? Is post-market surveillance linked to governance processes that can respond when performance degrades?

These recommendations reflect the Total Product Lifecycle model that now characterizes both FDA and NICE approaches to AI governance. Point-in-time validation is insufficient when models update, retrieval corpora change, and deployment contexts evolve. For HEOR evaluators, the practical question is whether the vendor can demonstrate any ongoing monitoring capability, or whether the validated state is a historical snapshot with no renewal mechanism.

Domain 6: Lifecycle Economics and Continuous Assurance (Recommendations #11, #12)

Do the tool's economic analyses account for ongoing costs: monitoring, governance, periodic revalidation, and model updates? Are there explicit stop criteria that determine when outputs are flagged or blocked?

Recommendation 11 notes that AI system costs are not one-time investments; governance infrastructure and revalidation carry ongoing resource requirements that should be modeled in procurement decisions. Recommendation 12 integrates continuous assurance with conditional access: "Linking access to evidence generation, monitoring and explicit stop criteria allows NICE to support timely innovation while retaining oversight as systems evolve" [2].

Table 4. Evaluation framework for AI research tools organized by NICE DSU recommendations

Evaluation Domain	Recommendations	Core Question	HEOR Relevance
Specification and transparency	#1, #4	Can the tool explain what it did and disclose limitations in context?	High: foundational for defensible evidence synthesis
Version control and change governance	#2, #3	Are outputs reproducible and are changes validated before deployment?	High: critical for regulatory and HTA submissions
Data integrity and bias management	#5, #6	Are data sources and their limitations disclosed?	Moderate: most salient for real-world data applications
Human-AI system design and oversight	#7, #8	Is human oversight meaningful, and does transferability	High: determines fitness for specific HTA contexts

		analysis apply?	
Performance monitoring and validation	#9, #10	Is ongoing monitoring planned and linked to governance?	High: necessary given model drift and update cycles
Lifecycle economics and continuous assurance	#11, #12	Are ongoing costs modeled and are stop criteria explicit?	High: essential for procurement and governance decisions

An Honest Assessment of Where the Market Stands

Few tools available (if any) today answer "yes" across all six domains with documented evidence. Infrastructure to continuously monitor AI model performance and detect drift "does not yet exist as off-the-shelf tooling," and most organizations rely on manual audits and periodic reviews [8]. That gap is most visible in Domains 5 and 6, where ongoing monitoring and explicit stop criteria require architectural investment that most current tools have not made.

Domains 1 and 2 are better served, at least partially: most tools can produce some form of output provenance, and version documentation is increasingly standard. Domains 4 and 6 represent the current evaluation frontier. Meaningful human-AI system design and genuine continuous assurance are the properties that distinguish a small number of carefully engineered tools from the broader market.

This framework is a map of where evaluation is heading. The NICE DSU's recommendations anticipate an appraisal environment in which these questions are asked systematically. Evaluators who begin applying them now will be better positioned as that environment matures, and they will ask better questions of vendors accustomed to capability demonstrations rather than governance scrutiny.

Section 6: What This Means for Evaluators, Builders, and the Profession

Three groups face distinct decisions about how to respond to the direction this framework describes.

For Evaluators

The capability floor for AI research tools has risen to the point where it no longer differentiates. Leading foundation models now cluster within a narrow performance band on graduate-level scientific reasoning benchmarks, while open-source alternatives approach proprietary performance at dramatically lower cost [4]. A tool built on one leading model is not meaningfully more capable than one built on a well-configured equivalent for the tasks that define HEOR workflows. Capability demonstrations in procurement settings answer questions that are increasingly beside the point.

The questions that now differentiate are the six evaluation domains from Section 5: specification transparency, version binding, change governance, human-AI system design, performance monitoring, and continuous assurance with stop criteria. These are the domains where tools diverge and where the distance between a credible system and a well-presented wrapper becomes visible.

A practical rebalancing follows from this: treat capability demonstration as a threshold, and reserve the majority of evaluation time for confidence-layer questions. Ask for documented evidence of version binding and change governance protocols. Ask where human judgment enters the workflow and what the system does when it reaches a decision point exceeding its reliable operating range. Most vendors will not have complete answers; the quality and candor of those answers is itself informative.

For Builders

When the underlying model is a commodity, the differentiator is what surrounds it: traceability infrastructure, uncertainty signaling, calibrated oversight design, and the governance architecture that makes continuous assurance operational. Organizations that treat confidence as an engineering discipline, building it into the pipeline from the outset, will be better positioned as evaluation demands mature. Those that treat assurance as a documentation exercise will find their tools increasingly indistinguishable to methodologically rigorous evaluators.

The embedded-versus-bolt-on distinction is the relevant design fork, and it cannot be resolved retroactively. The FDA's Predetermined Change Control Plan framework and the NICE DSU's continuous assurance recommendations both presuppose that governance is built into the pipeline, not appended to finished artifacts [2,10].

For the Profession

The Shen-Tamkin evidence that passive delegation erodes the analytical capacities needed for expert review makes displacement a substantive concern, not merely an identity one [9]. A confidence layer, designed well, addresses this structurally. Transparency and calibrated oversight preserve the points at which expert judgment is consequential rather than replacing those points with automated throughput.

The expert's role shifts from producing evidence to directing and validating evidence production, a higher-leverage position provided the system makes reasoning visible enough to direct and specific enough to validate. Paper 3 will illustrate what this looks like through a concrete multi-agent implementation where these properties are expressed as architectural decisions.

References

[1] ELEVATE-GenAI: Reporting Guidelines for the Use of Large Language Models in Health Economics and Outcomes Research: An ISPOR Working Group Report. https://www.ispor.org/publications/journals/value-in-health/abstract/Volume-28--Issue-11/ELEVATE-GenAI--Reporting-Guidelines-for-the-Use-of-Large-Language-Models-in-Health-Economics-and-Outcomes-Research--An-ISPOR-Working-Group-Report?utm_medium=press_release&utm_source=public&utm_campaign=value_in_health

[2] [PDF] FRONTIER ARTIFICIAL INTELLIGENCE AND HEALTH .... https://sheffield.ac.uk/media/118146/download?attachment

[3] LLM Benchmarks 2026: MMLU, GPQA Diamond, HLE, LiveCodeBench | CodeSOTA | CodeSOTA. https://www.codesota.com/llm

[4] The Definitive LLM Selection & Benchmarks Guide. https://iternal.ai/llm-selection-guide

[5] I Tested 15 LLMs on 38 Real Coding Tasks. Here's My Routing Table.. https://ianlpaterson.com/blog/llm-benchmark-2026-38-actual-tasks-15-models-for-2-29/

[6] Top 10 HEOR Trends. https://www.ispor.org/heor-resources/top-10-heor-trends

[7] [PDF] Securing a Resilient Financial Services Enterprise - Cisco. https://www.cisco.com/c/en/us/solutions/collateral/secure-the-enterprise/securing-fs-enterprise.pdf

[8] [PDF] A comparative evaluation of AI position statements and frameworks .... https://www.ispor.org/docs/default-source/cti-meeting-21305-documents/e21bbd7d-71d6-4036-a3aa-1d9eefa99d60.pdf?sfvrsn=efa92719_0

[9] [PDF] How AI Impacts Skill Formation - Crane. https://r.jordan.im/download/language-models/shen2026.pdf

[10] Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles | FDA. https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles