

Trusted AI for HTA Submissions: How DistillerSR Fulfills the Requirements of the ISPOR ELEVATE-GenAI Framework

HTA Bodies increasingly expect AI use in HEOR evidence to be documented against ELEVATE-GenAI's 10 Domains. General-Purpose Tools were not built to meet that standard. This Paper maps DistillerSR's Smart Evidence Extraction against all 10 Domains, from model traceability to security and privacy.

Executive Summary

HTA bodies including NICE, CDA, and G-BA have begun issuing guidance on AI use in submissions. The question facing HEOR teams is no longer whether to disclose AI use, but how to document it in a way that meets the standards regulators are applying.

ISPOR's ELEVATE-GenAI framework, published in *Value in Health* in November 2025¹, is the first structured, HEOR-specific reporting standard for studies where large language models (LLMs) play a substantive role in evidence generation, synthesis, or analysis. It is already referenced in NICE's AI position statement, and organizations submitting HEOR evidence to bodies that cite it will be expected to document AI use against its 10 domains.

That requirement creates a practical problem for teams using general-purpose AI tools such as ChatGPT, Claude, and Gemini. These tools were not designed for regulated evidence workflows. They offer no verifiable process: there is no record of which model version generated a given output, under what instructions, or on what date. The same question asked twice may produce different answers. They do not provide source-level citations linking extracted claims to specific passages in source documents. They maintain no audit trail. Independent research published in *JMIR Mental Health* in November 2025² found that nearly 20% of citations generated by GPT-4o in simulated literature reviews were fabricated, rising to 29% in specialized therapeutic areas.

DistillerSR's Smart Evidence Extraction (SEE) was designed for regulated evidence environments. This article maps SEE's capabilities against all 10 ELEVATE-GenAI domains and explains how organizations using SEE are positioned to meet the framework's documentation requirements.

Four design principles underpin SEE's architecture: source citations for every extraction, a full audit trail capturing model version and reviewer decisions, human expert review at every step, and a contractual commitment that customer data is never used to train third-party models. These align directly with what ELEVATE-GenAI requires.

For teams preparing submissions to NICE, CDA, G-BA, or other bodies referencing ELEVATE-GenAI, the practical implication is that using SEE for data extraction produces the documentation the framework calls for at the point of extraction. Model version traceability, source citations, confidence indicators, audit logs, and security controls are built into the workflow rather than assembled after the fact.

What the framework is clear about, and what no tool can substitute for, is the human expert judgment applied to AI outputs. DistillerSR's Expert-in-the-Loop design requires a human decision on every AI suggestion: accept, adjust, or reject. That decision is logged and forms part of the audit record that regulators and HTA bodies expect to see.

The Question Every HEOR Team Is Asking Now

We hear a version of the same question in almost every customer conversation, at every conference, and in every RFP that crosses our desks: can we use generative AI within research workflows and, if so, how do we do so in a way that the outputs will pass regulatory scrutiny?

The concern is legitimate and the stakes are real. HTA bodies including NICE³, the Canadian Drug Agency (CDA)⁴, and G-BA have begun issuing guidance on AI use in submissions. The FDA⁵ has published a risk-based credibility framework for AI in regulatory processes. And the EU AI Act is establishing new obligations for high-risk AI systems across regulated industries.

Against this backdrop, ISPOR's Working Group on Generative AI published the ELEVATE-GenAI framework in November 2025 in Value in Health¹. This represents the first structured, HEOR-specific reporting standard for studies where LLMs play a substantive role in evidence generation, synthesis, or analysis. It is already being cited in guidance from NICE and other HTA bodies as the reference framework for AI disclosure in submissions.

A distinction worth drawing: this is not an argument against AI in evidence synthesis, or against machine learning generally. Deterministic AI has been part of systematic review workflows for years, including our own. The issue is specific to general-purpose LLMs used through a chat interface, where the properties that make them useful for open-ended tasks – variability, fluency, inference beyond the source – are precisely the properties that make them indefensible in a regulatory submission.

The question is not whether to disclose AI use in HEOR submissions. It is how to document it in a way that satisfies regulators, withstands peer review, and holds up under HTA scrutiny. ELEVATE-GenAI provides the guidance. DistillerSR provides the platform on which to implement it.

This article explains what ELEVATE-GenAI requires, why general-purpose AI tools cannot meet those requirements, and how DistillerSR's regulatory-grade AI platform, specifically its Smart Evidence Extraction (SEE) capability, is designed to satisfy each of the framework's 10 domains.

What Is ELEVATE-GenAI and Why Does It Matter?

ELEVATE-GenAI is a 10-domain reporting framework and checklist developed by the ISPOR Working Group on Generative AI, published in Value in Health (November 2025, Vol. 28, Issue 11). Its purpose is to fill the gap left by existing AI reporting guidelines, which do not adequately address the challenges of transparency, reproducibility, factual accuracy, bias, uncertainty, and data governance in the context of HEOR decision-making, regulatory review, or HTA.

The framework is designed for studies where LLMs play a substantive role; systematic literature reviews, health economic modeling, real-world evidence generation. It covers both the technical characteristics of the AI system used and the quality and trustworthiness of its outputs.

It matters for three reasons:

- **It is already the reference standard.** NICE's position statement on generative AI in HTA submissions explicitly references the ELEVATE-GenAI framework. CDA and G-BA are following suit. Organizations submitting HEOR evidence to these bodies will increasingly be expected to document AI use against this framework.
- **It is a living guideline.** The ISPOR Working Group has committed to versioned updates as generative AI evolves, meaning this is not a one-time checklist but an ongoing governance standard.
- **It shifts the burden of proof.** The framework requires researchers to demonstrate, not merely assert, that their AI-generated evidence meets standards of accuracy, reproducibility, fairness, and security. That is a documentation and audit trail problem, not just a methodology problem.

Why General-Purpose LLMs Cannot Meet This Standard

The ELEVATE-GenAI framework is structured around requirements that general-purpose large language models such as Claude, ChatGPT, Gemini, and similar tools used through a chat interface are fundamentally unable to satisfy in a HEOR or regulatory context.

Consider what the framework requires. Every extraction must be traceable to its source document with specific citations. Model characteristics including version, training data, and update governance must be documented. Outputs must be reproducible: the same document, prompt, and model version must produce the same result. Confidence levels must be calibrated and communicated. Customer data must never be used to train third-party models. All AI activity must be logged in an immutable audit trail.

None of these requirements are met by general purpose LLMs. They operate as black boxes. Their outputs are non-deterministic by design; the same question asked twice may produce different answers. There is no audit trail. Training data provenance is opaque. And critically, many of these tools have been shown to hallucinate; a peer-reviewed study published in JMIR Mental Health in November 2025⁵ found that nearly 20% of citations generated by GPT-4o in simulated literature reviews were completely fabricated, rising to 29% in less familiar or specialized topics.

DistillerSR's Smart Evidence Extraction (SEE) was purpose-built for high-regulated environments. The following section maps DistillerSR's capabilities against each of the 10 ELEVATE-GenAI domains, drawing directly from our internal compliance analysis.

DistillerSR and ELEVATE-GenAI: A Domain-by-Domain Alignment

The table below summarizes DistillerSR’s coverage against each ELEVATE-GenAI domain, followed by a detailed explanation of how each requirement is met.

#	DOMAIN	DISTILLERSR COVERAGE	STATUS
1	Model Characteristics	Model name, version, and inference endpoint logged per extraction as an internal audit record; transparency documented through prospective validation studies; model card available; update governance policy in place	Covered
2	Accuracy Assessment	Precision and recall targets defined; gold-standard test sets used pre-release; go/no-go quality thresholds documented	Covered
3	Comprehensiveness Assessment	Omission rate tracked in production; adversarial edge-case testing; benchmark comparison against published Cochrane SLRs planned	Covered + planned
4	Factuality Verification	Hallucination rate independently measured; every extraction source-cited at page/section level; evidence verification built into prompt design	Covered
5	Reproducibility & Generalizability	Full extraction provenance logged; outputs replicable by re-running same model/prompt/document; regression testing after every update	Covered
6	Robustness Checks	Adversarial test set includes OCR errors, long documents, malformed PDFs, prompt injections; real-world quality variations tested	Covered
7	Fairness & Bias Monitoring	Test sets stratified by customer vertical, study design, and SLR type; extraction quality measured across study designs	Covered
8	Deployment Context & Efficiency	AI usage dashboard in-platform; deployment infrastructure documented and shareable on request; quota and throughput assessed	Covered + planned
9	Calibration & Uncertainty	High/medium/low confidence indicators per field; UI flags low-confidence extractions; calibration study in progress	Covered + calibrating
10	Security & Privacy	GDPR, NIST and SOC2 compliant; customer data encrypted in transit; zero customer data used for model training	Covered

Domain 1: Model Characteristics

The ELEVATE-GenAI framework requires researchers to document the model’s name, version, developer, release date, architecture, training data sources, and access method, and to establish governance over model updates so that silent changes cannot reach production without validation.

DistillerSR logs model name, version, and inference endpoint for every extracted data element, creating an audit record traceable to each output. Due to the rapidly evolving nature of AI models, the models used in the DistillerSR platform undergo continuous updates, refinement and optimization to leverage ongoing performance improvements. DistillerSR’s approach to transparency includes the generation and provision of prospective validation evidence. Extraction performance is documented through structured validation studies available to customers; internal audit records are retained to support regulatory enquiries as needed. Every extraction is traceable not just to the source document but to the specific model version that generated it. A formal model update governance policy requires approval before any new model version is deployed to SEE, preventing undocumented changes from reaching users. An official model card is available that documents architecture, training data sources, intended use, and known limitations.

Action in progress: model card information will be published in a

customer-facing document to facilitate ELEVATE-GenAI reporting for customers submitting to HTA bodies.

Domain 2: Accuracy Assessment

The framework requires accuracy to be evaluated against human benchmarks or gold-standard datasets, with documented metrics appropriate to the task (e.g., precision, recall, F1 score) and structured human evaluation for free-text generation tasks.

DistillerSR defines minimum precision and recall thresholds per document type and extraction field, measured against labeled gold-standard test sets before each release. Go/no-go release decisions are based on objective performance criteria, not subjective assessment. This means every SEE release that reaches customers has been validated against a defined accuracy standard, not just tested. Validation is not confined to internal testing. Select customers participate directly in the release cycle, testing extraction performance and accuracy on their own document sets before general availability. This gives us performance signals from real regulatory and HEOR workflows – different document types, different therapeutic areas, different review protocols – rather than from test data alone.

Action in progress: HEOR-specific benchmarks are being developed to complement the internal gold standards, providing domain-native

accuracy measures relevant to systematic reviews and economic modeling contexts.

Domain 3: Comprehensiveness Assessment

Comprehensiveness evaluates whether AI-generated outputs address all required elements of the assigned task, a particular concern in systematic reviews where missed evidence can directly affect safety or efficacy conclusions.

DistillerSR tracks omission rates in production: measuring critical fields missed entirely during extraction, monitored alongside precision to ensure balanced performance. An adversarial test set is specifically designed to catch systematic omissions; the edge cases that precision and recall metrics can miss.

Planned enhancement: a benchmark comparison against published Cochrane and PRISMA-compliant SLRs provides a HEOR-native definition of comprehensiveness, testing SEE extracts the same key evidence as domain-standard systematic reviews.

Domain 4: Factuality Verification

This is the domain most directly relevant to the hallucination concern. The framework requires hallucination rates to be measured independently and reported, with every AI-generated claim traceable to a specific source passage.

This is where DistillerSR's architecture diverges most sharply from general-purpose LLMs. Every SEE extraction includes a page-level, section-level, or full-text passage reference, enabling researchers to verify every claim against the source material before it is accepted. Source citations are not optional. Evidence verification is built into the prompt architecture, anchoring responses to document text rather than model priors. Hallucination rates are measured through independent validation evaluating extractions containing claims absent from source documents. An in-product flagging mechanism allows users to report wrong or missing extractions, creating a real-world quality signal from production.

In the context of HTA submissions, this domain is the most consequential. A system where every extracted data point is linked to its exact location in the source document is not simply a product feature. It is the foundational requirement for defensible AI-assisted evidence synthesis.

Domain 5: Reproducibility and Generalizability

The framework requires that LLM-assisted research can be independently reproduced (same model, same prompt, same document) and that performance generalizes across different study designs, publication periods, and geographies.

DistillerSR's audit logging captures full extraction provenance: document ID, timestamp, model version, prompt version, and additional metadata. This creates the conditions for independent reproduction that the framework requires. A standing regression test set is evaluated after every model or prompt update, detecting performance degradation before it reaches users. Generalizability assessment tests whether extraction quality holds across study designs, publication years, and languages, and identifies document types where performance degrades (such as older scanned PDFs or non-English text).

Domain 6: Robustness Checks

Robustness evaluation tests whether the AI system performs reliably under conditions that differ from an ideal scenario, mimicking the real-world document quality variations that any production system will encounter.

DistillerSR's adversarial test set explicitly simulates these conditions: OCR errors, excessively long documents, malformed PDFs, unrelated research questions, and prompt injection attempts. This is not academic stress-testing; it reflects the actual document landscape in systematic literature review workflows, where PDFs from heterogeneous sources, varying quality levels, and historical publications are the norm.

Domain 7: Fairness and Bias Monitoring

Fairness evaluation ensures that AI performance does not systematically vary across different study types, populations, or research domains in ways that could introduce bias into evidence synthesis.

DistillerSR's test sets are stratified by customer vertical (pharmaceutical, medical devices, HEOR), study design, and SLR type. Extraction quality is measured across these strata to ensure consistent performance. This matters in HEOR because systematic bias toward certain study designs (e.g., RCTs over observational studies) could skew evidence bodies in ways that affect HTA conclusions.

Domain 8: Deployment Context and Efficiency

This domain requires transparency about how AI is deployed in practice; what users can monitor, what infrastructure underpins the service, and how costs and throughput are managed.

DistillerSR provides an in-platform AI Dashboard that allows teams to track documents processed and monitor AI usage across their workflow. Deployment infrastructure is documented and available on request. Quota management and throughput capacity have been assessed. Additional deployment transparency information has been added to customer-facing documentation.

Domain 9: Calibration and Uncertainty

Calibration evaluates whether an AI system's expressed confidence in its outputs is actually predictive of correctness, and whether users are appropriately informed of uncertainty to prevent over-reliance. Uncertainty is currently expressed through natural language text and high, medium, and low confidence indicators.

SEE returns high, medium, and low confidence indicators for each extracted field. The user interface visually distinguishes low-

confidence extractions to prevent over-reliance, flagging them for closer human review rather than silent acceptance. Explicit uncertainty handling is built in: the system uses natural text, provides ‘undetermined’ and ‘not answerable’ options with explanations from the AI, and testing is underway for a SEE numerical Confidence Score, starting with Abstracts and then PDFs. A calibration study is also in progress to quantify the relationship between model confidence scores and actual correctness, and to define action thresholds for each confidence level.

Domain 10: Security and Privacy

The framework requires that AI systems used in HEOR research comply with applicable data protection regulations, that customer data is handled securely, and that inputs are not used to train third-party models.

This last point is non-negotiable for organizations handling proprietary research data or unpublished clinical evidence. DistillerSR is GDPR and SOC 2 compliant. Customer data is encrypted in transit and at rest. Zero customer data is used for model training, re-training, fine-tuning, or correcting – a commitment that is contractual. This means organizations can use SEE on sensitive unpublished data such as clinical evaluation reports, health economic models and regulatory submissions in draft, without risk of IP exposure.

One Architecture, Several Frameworks

ELEVATE-GenAI is not the only standard organizations must satisfy. The NIST AI Risk Management Framework governs how we design, test, and validate AI capability. Cochrane’s RAISE guidance addresses AI use in evidence synthesis. The EU AI Act imposes obligations on

high-risk systems. These frameworks use different vocabulary, but they converge on the same underlying requirements: traceability, human oversight, documented validation and data governance.

What This Means in Practice for HTA Submissions

For teams preparing HEOR evidence for submission to NICE, CDA, G-BA, or other HTA bodies that reference the ELEVATE-GenAI framework, the practical implications are straightforward.

Using DistillerSR’s SEE for data extraction generates the documentation artifacts that ELEVATE-GenAI requires. The extraction audit trail provides model version traceability. The source citations satisfy the factuality verification domain. The accuracy metrics from our pre-release testing provide the quantitative performance evidence the framework asks for. The security architecture satisfies Domain 10. The confidence indicators satisfy Domain 9.

Teams do not need to build this infrastructure from scratch or retroactively document AI use after the fact. It is captured at the point of extraction, in a format that is structured for auditability and reporting.

What remains the team’s responsibility, and what the framework is clear about, cannot be delegated to any tool, is the human expert judgment applied to the AI’s suggestions. DistillerSR’s Expert-in-the-Loop design ensures that every AI suggestion requires a human decision: accept, adjust, or reject. That decision is logged. That log is the evidence of human oversight that regulators and HTA bodies require.

ELEVATE-GenAI establishes what good AI use looks like. DistillerSR was built to that standard.

DistillerSR’s Four Guiding Principles, Aligned to the Framework

DistillerSR’s AI development is guided by four principles that map directly to the ELEVATE-GenAI framework’s core concerns:

1. Defensibility Over Convenience

Every SEE output is source-cited, every decision is logged, and every model version is traceable. This is not the fastest way to build an AI feature, but it is the only way to build one that can survive regulatory scrutiny.

2. Human Expertise Embedded in the Workflow

SEE suggests, but experts decide. The framework’s requirement for structured human evaluation is not a constraint we work around; it is the design principle we built from. The design also accommodates a ‘human over the loop’ automation option that can be adopted after achieving sufficient trust in the system. Confidence indicators, visual uncertainty flags, and the one-click accept/adjust/reject interface are all expressions of this principle.

3. Audit-ready by Design

The audit trail is generated at the moment of extraction, capturing model version, prompt version, document ID, timestamp, and the reviewer’s decision. This is what ELEVATE-GenAI Domain 5 requires for reproducibility and what Domains 1 and 4 require for traceability.

4. Your Evidence is Never Used to Train Third-Party Models

This maps directly to ELEVATE-GenAI Domain 10. organizations doing proprietary HEOR research cannot afford to have their unpublished data ingested by general purpose LLM training pipelines. We have structured our data architecture to ensure this cannot happen.

DistillerSR Regulatory-Grade AI

Two decades of building AI for regulated research

DistillerSR has been applying AI to evidence synthesis since long before it became fashionable. Our first peer-reviewed AI paper was published in 2009. The first generation of DistillerSR AI shipped in 2017, followed by AI Classifiers in 2019 and AI Re-rank in 2020 – deterministic machine learning applied to screening and prioritization years before large language models entered the field. Smart Evidence Extraction followed in 2024, alongside our adoption of the NIST AI Risk Management Framework. Tabular data extraction, MCP connectivity, and agentic, fully automated workflows arrived in 2025.

That work is documented in public. Research by DistillerSR employees on AI in evidence synthesis has been cited in more than 650 peer-reviewed publications. The platform is relied on by over 250 organizations, including 80% of the top pharmaceutical and medical device companies, and currently holds more than 24 million references.

We also help write the standards we are held to. DistillerSR leads the AIDESL consortium under Eureka-ITEA, participates in NICE's HTA Lab on governance for agentic AI, and is one of five regulatory technology providers supporting the EU's DIGICAP project on digitizing conformity assessment for medical devices and IVDs.

The question of whether GenAI can be used in HTA submissions has been answered. The frameworks are in place. The answer is yes, provided the AI system and the process followed meets the documentation, accuracy, reproducibility, HITL, and security standards that ELEVATE-GenAI establishes and that HTA bodies are beginning to enforce.

The question now is whether your current AI tooling can support that documentation. General-purpose LLMs cannot. Purpose-built, regulatory-grade AI platforms can.

DistillerSR's alignment with ELEVATE-GenAI across all 10 domains is the result of building SEE from the ground up for the requirements of regulated evidence synthesis in mind; defensibility, human oversight, audit readiness, and data security built into the architecture. ☞

References:

1. Fleurence, Rachael L., Dalia Dawoud, Liang Bian, Mitchell K. Higashi, Xiaoyan Wang, Hua Xu, Jagpreet Chhatwal, Turgay Ayer, and ISPOR Working Group on Generative AI. "ELEVATE-GenAI: Reporting Guidelines for the Use of Large Language Models in Health Economics and Outcomes Research: An ISPOR Working Group Report." *Value in Health*, vol. 28, no. 11, Nov. 2025, pp. 1611-1625. <https://doi.org/10.1016/j.jval.2025.06.018>[https://www.valueinhealthjournal.com/article/S1098-3015\(25\)02455-6/pdf](https://www.valueinhealthjournal.com/article/S1098-3015(25)02455-6/pdf)
2. Linardon, Jake, Hannah K. Jarman, Zoe McClure, Cleo Anderson, Claudia Liu, and Mariel Messer. "Influence of Topic Familiarity and Prompt Specificity on Citation Fabrication in Mental Health Research Using Large Language Models: Experimental Study." *JMIR Mental Health*, Nov. 12, 2025. <https://doi.org/10.2196/80371>
3. National Institute for Health and Care Excellence. "Use of AI in Evidence Generation: NICE Position Statement." NICE, Aug. 2024. <https://www.nice.org.uk/corporate/ecd11/resources/use-of-ai-in-evidence-generation-nice-position-statement-pdf-40464268944325>
4. Canadian Drug Agency. "CDA-AMC Position Statement on the Use of AI in the Generation and Reporting of Evidence." CDA-AMC. https://www.cda-amc.ca/sites/default/files/MG%20Methods/Position_Statement_AI_Renumbered.pdf
5. U.S. Food and Drug Administration. "Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations." FDA, Jan. 2025. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device>
6. DistillerSR ELEVATE-GenAI Compliance Analysis: internal assessment document, April 2026. (Available on request for enterprise customers.)