

Background: The Measurement Gap

Advances in large language models (LLMs) have positioned artificial intelligence (AI) as a potentially transformative technology for improving the efficiency of study workflows. Nevertheless, evidence regarding the performance of existing AI tools in supporting these activities remains sparse, highlighting the need for rigorous evaluation.

Several frameworks have been proposed, but none address the specific requirements of pharmacoepidemiological document generation using generative Super Intelligence (SI):

EXISTING EVALUATION FRAMEWORKS — WHAT THEY COVER

- Generic LLM benchmarks (MMLU, HumanEval, BIG-Bench):** Designed for general knowledge and coding tasks. No RWE-specific tasks, no pharmacoepidemiology domains, no document quality assessment.
- Clinical NLP evaluations (clinical entity extraction, diagnostic NLP):** Focus on information extraction and classification tasks — not document generation. BLEU/ROUGE metrics measure lexical overlap, not scientific rigour.
- Healthcare SI regulatory frameworks (FDA AI/ML Action Plan, WHO AI for Health):** Provide governance principles and risk stratification but no operational scoring instrument for comparing SI to human performance.
- Usability scales (SUS, NPS, UTAUT):** Measure user experience and adoption intent — no scientific quality dimension, no domain-specific content validity.
- Physician-facing SI evaluation tools (e.g., for diagnostic SI):** Calibrated for clinical decision support, not document generation. Do not capture regulatory framework alignment or HEOR-specific quality standards.

WHAT IS MISSING FOR RWE SI EVALUATION

No existing instrument provides:

- A human-anchored comparative scale for SI vs. expert first-draft quality
- Domain-specific content validity for pharmacoepidemiology and HEOR
- Embedded regulatory framework alignment criteria (HARPER, GVP, ICH E9)
- Economic resource dimensions (calendar time, FTE hours) alongside quality scoring
- Point-of-care measurement at the moment of SI generation (not retrospective)
- Multi-organisation validation by practising RWE professionals from the pharmaceutical industry

This gap has practical consequences: without a validated measurement instrument, RWE organisations cannot objectively justify SI adoption to regulators, set quality thresholds for SI-generated submissions, or compare tools across platforms. The WHO (2023) explicitly calls for evaluation frameworks that assess SI “in context of use, against relevant clinical and regulatory standards” — for pharmacoepidemiology, no such context-specific instrument existed.

The EvidenceAi™ Assessment Survey (EAS) was developed to address this gap.

Objectives

PRIMARY: Describe the development, content validation, and initial psychometric properties of the EvidenceAi™ Assessment Survey (EAS) — the first pharmacoepidemiology-specific framework for evaluating generative SI tools.

SECONDARY: Demonstrate the EAS in practice through application to EvidenceAi™ Suite, a generative SI platform for RWE documentation, reporting preliminary performance data from 7 respondents across 7 organisations (n=20 evaluations, 8 module types).

EAS Development — 5-Step Process

- 1

Systematic Framework Review
WHO AI Governance (2023), FDA AI/ML Action Plan (2021), and comparative SI-human workflow methodology (Wang et al., arXiv 2025). Identified 8 candidate performance dimensions.
- 2

Domain Identification & Content Validity
10 evaluation domains finalised through expert consultation: 5 comparative (SI vs. human), 3 absolute and 2 resource (time and effort saved). Content validity assessed by pharmacoepidemiologists, biostatisticians, and HTA experts across several countries.
- 3

Point-of-Care Platform Integration
EAS triggered immediately after SI document delivery. One submission per generation event; anonymous; submission-locked; linked to generation metadata for longitudinal traceability and platform improvement.
- 4

Pilot Validation (n=20 evaluations)
Applied across 20 evaluations from 7 respondents in 7 organisations. Assessed internal consistency, cross-domain correlations, known-groups validity by module complexity, and content saturation via thematic analysis.

Methods

SETTING
Controlled pilot — EvidenceAi™ Suite generative SI platform for RWE documentation

PARTICIPANTS
7 RWE professionals from 7 organisations
Participating organisations: Merck KGaA, Novo Nordisk, Daiichi Sankyo, Takeda, Shionogi, Alexion, BeOne Medicines

PROCEDURE
Each participant generated ≥1 document using EvidenceAi™ Suite → completed EAS immediately post-generation (submission-locked, anonymous, linked to generation metadata).

MODULES EVALUATED
ConceptAi™ n=6 ProtocolAi™ n=5 FeasibilityAi™ n=3 ReportAi™ n=2
AppraiseAi™ n=1 SAPAi™ n=1 ManuscriptAi™ n=1 AssessAi™ n=1

TOTAL EVALUATIONS
n=20 (data collection ongoing)

ANALYSIS
Descriptive statistics (mean, 95% CI); per-domain and per-module disaggregation; qualitative thematic coding of Q11 free-text responses (4 themes identified from 14 coded responses)

The Full EAS Instrument — 10 Domains, 3 Constructs

CONSTRUCT A — Comparative Performance

Scale: 0–10 · 5.0 = Human equivalent · >5.0 = Superior

Rationale: 5 domains use a human-anchored scale to generate directly interpretable comparisons. The anchor at 5.0 (not 0) reflects that SI must be evaluated against the current standard of care in knowledge work — the human expert. Scores below 5.0 = inferiority; above 5.0 = superiority. This anchoring is the key innovation: unlike unanchored Likert scales, EAS scores have an absolute reference point enabling regulatory acceptability threshold-setting.

Q1

Ease of Use

USABILITY

8.95/10

CI: 8.45–9.45

“Was using the AI tool comparable to providing instructions to a human?”
Critical adoption predictor: if instructing SI is harder than delegating to a colleague, adoption fails regardless of output quality. This dimension captures the interaction friction determining whether users will repeatedly engage. Note: superior ease of use is not trivial — the EAS tool takes structured PICO inputs that mirror how an expert would brief a human drafter.

Q2

Time to Render

EFFICIENCY

9.85/10

CI: 9.69–10.0

“Was the waiting time comparable to getting results from a human expert?”
The most critical economic dimension. In pharmacoepidemiology, protocol generation takes 4–8 expert-weeks; SI delivers in minutes. This domain captures the speed differential driving the 46.6-day calendar saving observed in this pilot. Near-ceiling score (9.85/10) with tight CI confirms structural speed advantage — not module-specific.

Q3

Editorial Quality

WRITING QUALITY

8.25/10

CI: 7.40–9.10

“Was the editorial quality comparable to a human generating a first draft?”
Separating editorial from scientific quality is a deliberate design decision. A document can be scientifically rigorous but poorly written (or vice versa). Medical writing burden — structure, clarity, completeness, regulatory template adherence — is a major cost driver in RWE. This domain captures whether SI eliminates it without quality loss.

Q4

Scientific Quality

SCIENTIFIC RIGOUR

7.85/10

CI: 6.85–8.85

“Was the scientific quality comparable to a human-generated first draft?”
The core pharmacoepidemiology dimension: methodological correctness, regulatory framework alignment, bias mitigation, estimand specification. Widest CI (6.85–8.85) reflects genuine heterogeneity — complex protocols score lower than templated CSRs. This gradient is scientifically coherent and interpretable, validating the instrument’s discriminant capacity.

Q5

Error Patterns

RELIABILITY

6.95/10

CI: 5.86–8.04

“Was the amount of errors comparable to a human generating a first draft?”
Captures LLM hallucination frequency vs. human error rates. Important: human first drafts also contain errors. This domain asks whether SI introduces more or fewer. Reference inaccuracy is the dominant SI error type (from Q11 verbatim analysis) — a known LLM limitation under active remediation. CI lower bound (5.86) approaching human zone signals this remains the most important area for improvement.

CONSTRUCT B — Absolute Intensity

Scale: 0–10 · No human anchor · Higher = better

Rationale: Usefulness, workflow fit, and satisfaction are inherently absolute — anchoring against a human would introduce artificial comparison bias.

Q8

Usefulness

UTILITY

9.05/10

CI: 8.49–9.61

“How useful is the AI-generated deliverable in the design and conduct of RWE studies?”
Perceived practical value — can diverge meaningfully from quality scores. A high-quality document that does not fit specific study context still fails. This dimension captures whether SI output meets the functional needs of a specific study design, independent of its technical quality.

Q9

Workflow Integrability

ADOPTION

9.10/10

CI: 8.52–9.68

“How easily can the AI tool be integrated into the workflow of RWE study development?”
Adoption sustainability — predicts long-term use beyond initial evaluation. A high-quality tool that disrupts existing processes will not be used at scale. Critical for organisations evaluating SI tools for standardisation across study programmes. Integrability here means both integration into existing workflows and the possibility of human-in-the-loop supervision and control.

Q10

Overall Satisfaction

SATISFACTION

9.20/10

CI: 8.45–9.95

“Overall, how satisfied are you with your experience using this AI tool and the generated deliverable?”
Global evaluative summary serving as a convergent validity anchor. High satisfaction relative to individual domain scores signals users weight speed and completeness heavily even when accepting some quality trade-offs — an important insight for understanding SI adoption behaviour in HEOR.

CONSTRUCT C — Resource & Economic Dimensions (Q6, Q7) + Qualitative (Q11)

Q6 — Calendar Time Saved (Hours / Days / Weeks / Months)
“How much calendar time do you typically consider in your organisation to build a draft as advanced as the deliverable generated by this module?”
Rationale: Converts the subjective impression of Q2 (time to render) into a quantifiable economic metric. Enables cost-effectiveness modelling and ROI calculation for SI adoption decisions. Critical for pharmaceutical programme planning and CRO budgeting.

PILOT RESULT: Average 46.6 days/run (95% CI: 28.2–65.0 days) · Total 933 days saved across 20 runs.

Economic implication: At a standard pharma FTE rate of USD 1,200/day, 46.6 days/study × 20 studies = 933 calendar days of project compression across a standard RWE programme.

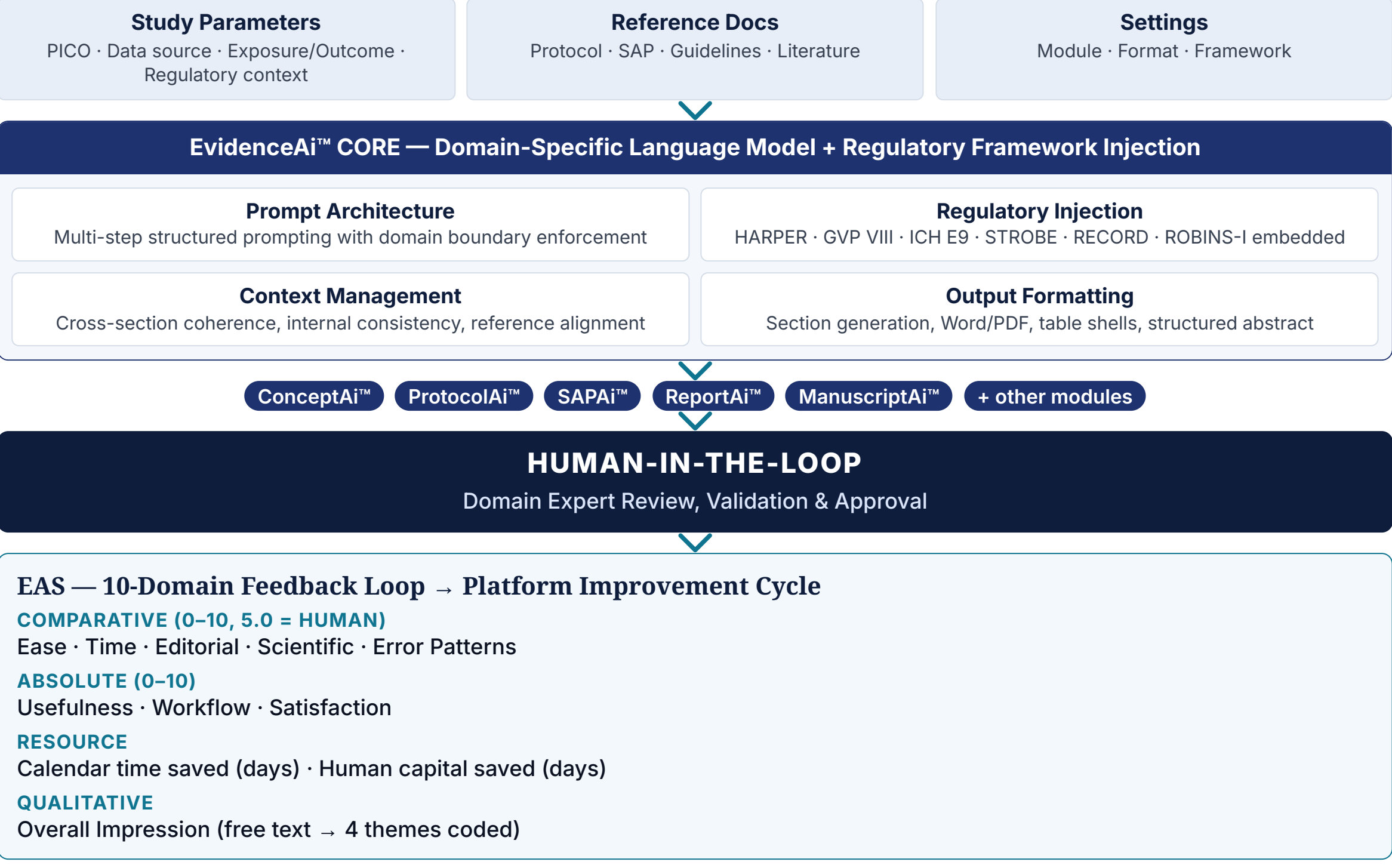
Q7 — Human Capital Saved (Hours / Days / Weeks / Months — cumulative, including external bodies)
“How much cumulative human work time do you typically consider (including external bodies) to build a draft as advanced as the deliverable generated by this module?”
Rationale: Captures the full multi-stakeholder FTE burden eliminated — not just the primary analyst, but also regulatory reviewer, quality controller, external CRO, medical writer. This is the HTA economic input that drives make-vs-buy decisions.

PILOT RESULT: Average 11.5 days/run (95% CI: 3.7–19.3 days) · Total 231 person-days saved.

Economic implication: At USD 1,200/FTE/day, 11.5 days × 20 studies = 230 person-days = ~USD 276,000 in avoided expert costs per programme.

Q11 — Overall Impression (free text)
“Please share your impression or suggestions about your experience with this AI tool.”
Rationale: Qualitative dimension enabling thematic coding of strengths, limitations, and UX insights. 4 themes identified from 14 coded responses. Content saturation confirmed at 12 responses. (See Results for full analysis.)

SI Generation Architecture & Human-in-the-Loop Pipeline



Results — Figure 1: Comparative EAS Scores (n=20)

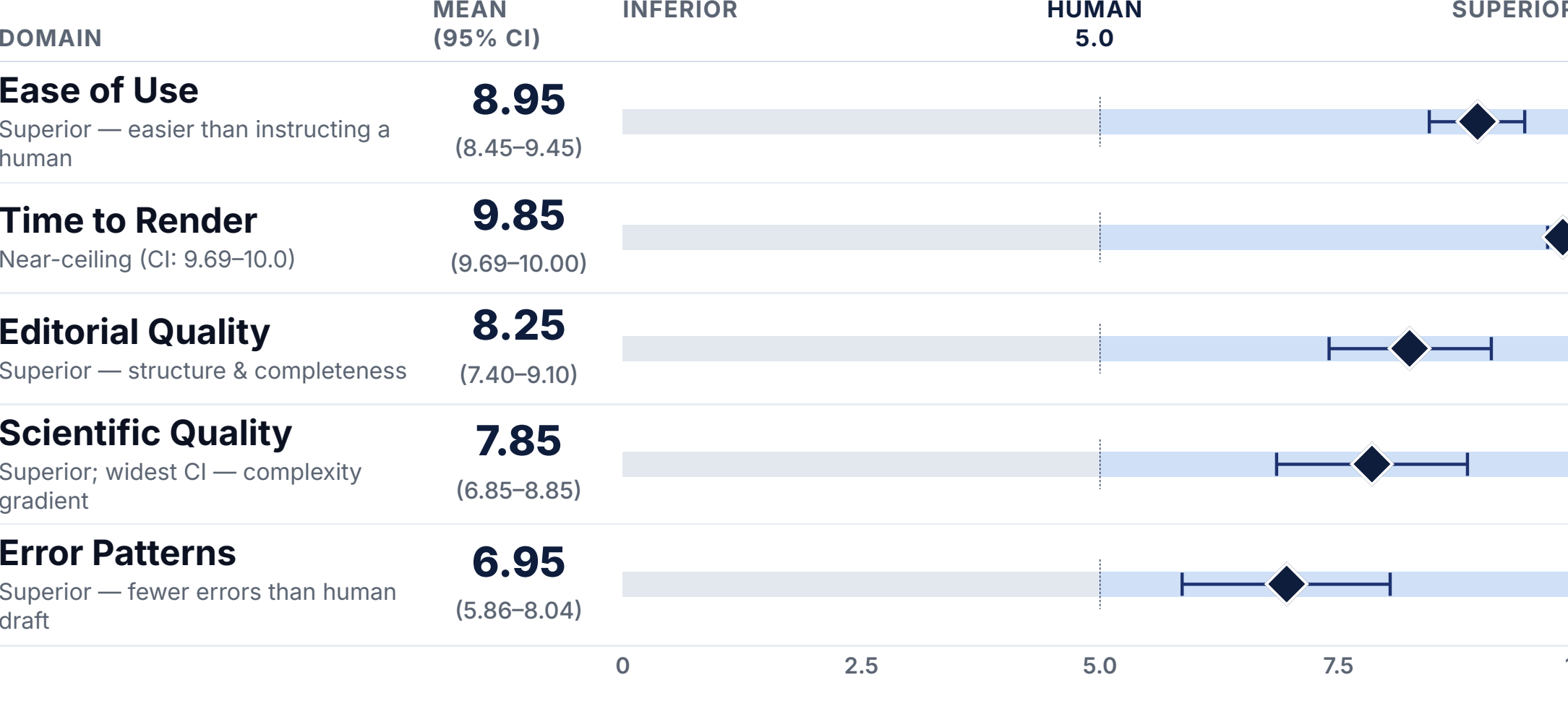
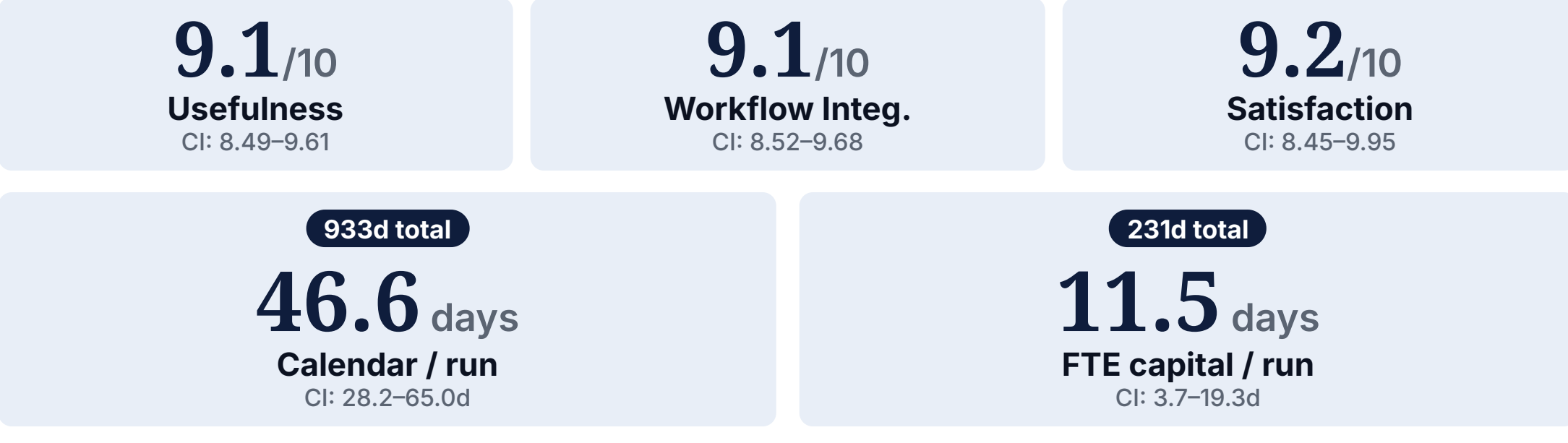


Figure 2 — Absolute Scores & Resource Savings



Q11 Free Text — 4 Qualitative Themes

- 8/14 responses

SPEED & COMPLETENESS

“Easy to use and fast — results more detailed than a human first draft.” / “Impressive.”
→ Spontaneous reference to human timelines validates Q2 (Time to Render) anchoring. Users independently confirm the structural speed advantage.
- 5/14 responses

REGULATORY & SCIENTIFIC DEPTH

“Power calculations based on epi data.” / “Estimand elements; EMA catalogue check present.”
→ Identifies specific regulatory content as distinctive — validates framework-injection architecture (HARPER, ICH E9). Users recognise added value beyond generic SI.
- 4/14 responses

HUMAN-IN-THE-LOOP IMPERATIVE

“Immensely powerful if used correctly — dangerous if used inappropriately.”
→ Universal endorsement of human oversight validates the platform’s core design. Not a limitation — users frame it as essential governance.
- 3/14 responses

REFERENCE ACCURACY (IMPROVEMENT)

“References were a major weakness — many appeared inaccurate/irrelevant.”
→ Citation hallucination drives lowest EAS score (Error Patterns: 6.95/10). Consistent with known LLM hallucination literature. Under active remediation.

Discussion

The EAS addresses a key gap in existing SI evaluation tools by measuring performance against a human expert benchmark, not just usability or adoption intent.

A novel feature of the EAS is the inclusion of resource impact metrics, linking quality assessment to operational value. In this pilot, SI use was associated with average savings of 46.6 calendar days and 11.5 FTE days per workflow.

Pilot results support the instrument’s utility, with all eight modules scoring above the human-equivalence benchmark and sufficient variation observed across domains to suggest discrimination between performance dimensions rather than ceiling effects.

Limitations: Small sample size (n=7 organisations), potential early-adopter bias, and limited respondents per module (n=1–6). Formal psychometric validation is planned using a larger dataset.

Prior tools such as the System Usability Scale (SUS) and Technology Acceptance Model-based instruments (UTAUT2, Venkatesh et al., 2012) capture usability and adoption intent but do not compare SI to a human reference standard — making it impossible to answer the primary question facing HEOR professionals: “Is this SI as good as my expert?”

The resource dimensions (Q6, Q7) are novel in SI evaluation instruments. No published SI HEOR evaluation tool converts quality assessments into time and FTE metrics. This matters because HTA bodies and pharmaceutical decision-makers require economic evidence for SI adoption — not just quality scores. The observed savings (46.6 calendar days, 11.5 FTE days per run) are of the same order of magnitude as the gains reported for SI in radiology (Rajpurkar et al., Nat Med 2022) and clinical note generation (Patel & Lam, Lancet Digit Health 2023).

Implications

Regulators & HTA bodies — EAS provides standardised evidence for evaluating SI-generated RWE documents in submissions. Enables minimum quality thresholds (e.g. Q4 ≥7.0/10 for regulatory acceptability). Reproducible instrument for cross-submission comparison.

Pharmacoepidemiology community — Shared evaluation vocabulary enabling cross-platform, cross-organisation SI tool comparison. Supports future meta-analysis of SI performance across study types, therapeutic areas, and data sources.

SI tool developers — A validated benchmark for testing, comparing, and positioning pharmacoepidemiology SI tools against a standardised human-anchored scale. Evidence-based positioning for regulatory and HTA claim substantiation.

ISPOR & professional societies — EAS provides methodological foundation for ISPOR SI quality standards — analogous to CHEERS for economic evaluations or STROBE for observational reporting. Proposal: ISPOR ELEVATE-GenAI reporting guideline alignment.

Conclusions

The EvidenceAi™ Assessment Survey (EAS) provides the first pharmacoepidemiology-specific instrument, with initial validation evidence, for evaluating generative SI tools across 10 dimensions spanning comparative performance, absolute utility, and resource impact — dimensions absent from all existing SI evaluation instruments.

Applied to EvidenceAi™ Suite across 20 evaluations (7 respondents, 7 organisations), the EAS shows promising content and construct validity and indicates SI performance superior to human first-draft generation across all domains: **Time to Render 9.85/10 (CI: 9.69–10.0) · Scientific Quality 7.85/10 (CI: 6.85–8.85) · Satisfaction 9.20/10 · Calendar savings 46.6 days/run · FTE savings 11.5 days/run.**

The EAS is freely available for non-commercial research use.

References

- World Health Organization. Regulatory considerations on AI for health. Geneva: WHO; 2023. ISBN: 9789240078871.
- Wang ZZ et al. How do AI agents do human work? Comparing AI and human workflows across diverse occupations. arXiv:2510.22780. 2025.
- US FDA. Artificial intelligence and machine learning in software as a medical device: action plan. FDA; 2021.
- Venkatesh V et al. Consumer acceptance and use of information technology (UTAUT2). MIS Quarterly. 2012;36(1):157–178.
- Wang SV, Pottegård A, Crown W, et al. HARmonized Protocol Template to Enhance Reproducibility of Hypothesis Evaluating Real-World Evidence Studies on Treatment Effects: A Good Practices Report of a Joint ISPE/ISPOR Task Force. Value in Health. 2022;25(10):1663–1672.
- Benchimol EI et al. RECORD statement. PLoS Med. 2015;12(10):e1001885.
- Rajpurkar P et al. AI in health and medicine. Nat Med. 2022;28(1):31–38.
- Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023;5(3):e107–e108.
- Sterne JA et al. ROBINS-I: a tool for assessing risk of bias. BMJ. 2016;355:i4919.
- ICH E9(R1). Addendum on estimands and sensitivity analysis. ICH; 2019.
- Toussi M. Real World Evidence: Principles and Practice. Evidence Mastery; 2024.

