New Failure Modes, Old Standards: Why Artificial Intelligence (AI) Demands Different Validation Approaches in Evidence Synthesis

Priccila Zuchinali[1], Kassandra Schaible[2], Allie Cichewicz[3]

[1]Thermo Fisher Scientific, Ottawa, ON, Canada; [2]Thermo Fisher Scientific, Pittsburgh, PA, USA; [3]Independent Consultant, Boston, MA, USA

MSR84

Background

- The rise of artificial intelligence (AI) has brought notable efficiency gains to systematic literature review (SLR) workflows, streamlining the entire process from searches through narrative synthesis. While AI tools expedite tasks, accuracy can be inconsistent; this introduces new challenges for ensuring high-quality outputs from AI-assisted workflows.
- Current approaches to validating AI performance rely almost exclusively on benchmarking against humans to determine whether AI can "match" human performance.

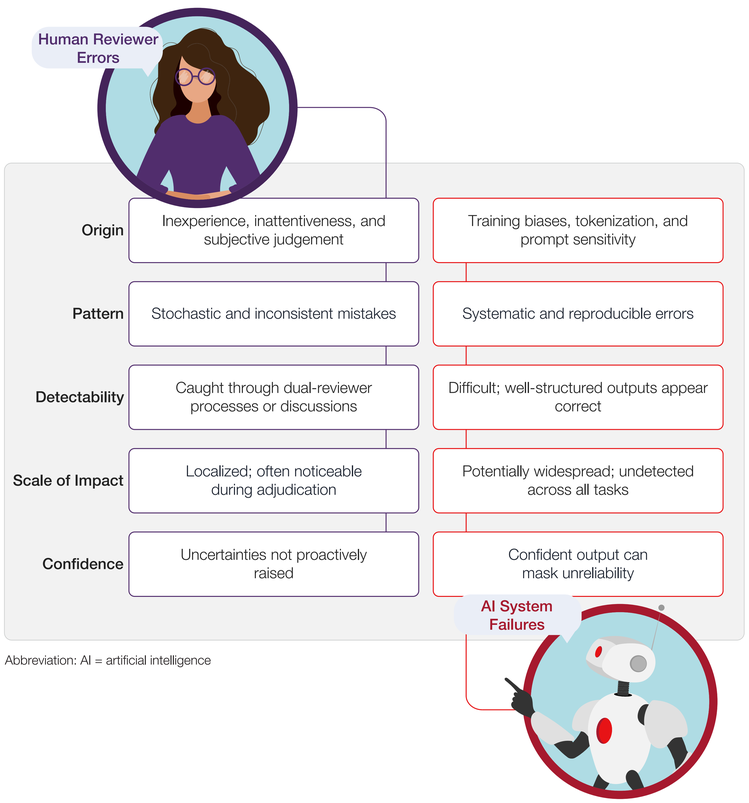
The Core Problem

- While statistical performance metrics are informative, they rely on assumptions that 1) the human benchmark is an absolute gold standard and is inherently flawless and 2) there is an appropriate benchmark for every review. Instead, how do you validate AI-assisted reviews in real time as you would check the work performed by a human reviewer? AI and human errors are not comparable in nature, distribution, or consequence; errors originate from fundamentally different sources, follow different patterns, and require different detection strategies, thus, making traditional validation methods insufficient. Applying human-centric validation frameworks to AI-assisted workflows creates blind spots that can compromise the integrity of the review.
- Therefore, this work aims to: 1) characterize the fundamental differences between human and AI errors in evidence synthesis; 2) identify task-specific AI failure modes across the SLR process; and 3) provide AI-specific validation recommendations with practical safeguards for each stage of the review process.

The Nature of Common SLR Errors

- Understanding what types of errors arise in the SLR process is essential to designing appropriate validation methods (**Figure 1**). Human errors are inconsistent, easier to catch, and usually limited in scope, while AI errors are systematic, harder to detect, and can scale across review tasks. We discuss a stage-specific, failure-mode-driven validation framework for AI-assisted SLRs.

Figure 1. Characteristics of Human vs. AI Errors



Abbreviation: AI = artificial intelligence

- AI can exhibit distinct failure modes at each stage of the review, with the underlying mechanisms varying between purpose-built evidence synthesis platforms (e.g., Nested Knowledge, DistillerSR) and generic large language models (LLMs; e.g., ChatGPT, Claude).
- Humans fail through fatigue, interpretation variance, and cross-reviewer inconsistency; purpose-built platforms fail through omission, misalignment, and structural constraint; and generic LLMs fail through fabrication, substitution, and miscalibrated confidence (**Figure 2**).

Figure 2. Main Errors in the SLR Process by Type of Reviewer

	1 Searches	2 Title/Abstract Screening	3 Full-text Screening	4 Data Extraction	5 RoB Assessment	6 Narrative Synthesis
Humans	Incompleteness from topic unfamiliarity; inconsistent syntax across sources; missed index terms	Inconsistent decisions from fatigue/ ambiguity; reviewer disagreement on edge cases	Overlooked PICOS elements or nuance; selective attention under time pressure	Transcription errors; misinterpreting data; inconsistent data entry; calculation errors	Unfamiliarity with tool mechanics; evaluator variability; inconsistent severity ratings	Personal interpretation bias; difficulty synthesizing conflicting results; variable writing styles
purpose-built platforms	Strategies biased toward input phrasing; limited sensitivity for multi-concept topics; iterative refinement drifts from original question	ML models underperform on underrepresented study types; integrated LLMs prompt-sensitive with inconsistent responses across batches	PDF parsing fails on complex/ scanned layouts; limited supplement extraction; PICOS in non-standard locations undetected	Omission/ misalignment of values; substitutes proximate but incorrect values when target data absent	Missed bias indicators in supplements/ non-standard sections; substitutes proximate justifications when evidence sparse	Oversimplification of findings; misweighting impactful studies; smoothing over heterogeneity and nuance
Generic LLMs	Invalid platform-specific search strings; hallucinated index terms; ignores syntax/field tags	Stateless batch processing with no memory of prior decisions; prompt-sensitive with shifting inclusion patterns	Truncation beyond context windows; misattribution across multi-arm papers; PICOS lost to truncation/ concatenation	High fabrication risk with confident phrasing; substitutes incorrect but real values when target data absent	Defaults to unclear/ moderate ratings; cannot distinguish domain-specific from surface bias	May fabricate findings/stats/ citations; draws on training data vs. reviewed evidence; confident language masks uncertainty

Abbreviations: LLM = large language model; PICOS = Population, Intervention, Outcomes, Study Design; RoB = risk of bias

Current Validation Standards

- Current validation standards were designed around properties of human error and are insufficient as the sole strategy when AI is involved:
  - **Search peer review** assumes a qualified reviewer can identify strategy errors. AI-generated searches may appear sound while containing platform-specific errors invisible to generalist reviewers.
  - **Pilot calibrations** align reviewer understanding through discussion. AI behavior is determined by training data, architecture, and prompting, not consensus.
  - **Dual independent review** assumes errors are stochastic. A second AI pass may replicate the same systematic mistake.
  - **Inter-rater agreement (e.g., Cohen's kappa)** measures concordance but cannot distinguish accurate agreement from shared systematic bias.
  - **Consensus resolution** relies on deliberation between reviewers. AI cannot deliberate; regenerated outputs may or may not differ.
  - **Risk-of-bias (RoB) domain adjudication** assumes evaluators apply domain-specific expertise. AI rates domains uniformly without distinguishing which require specialized judgment.
- These practices remain essential for human review components but leave AI-specific failure modes unaddressed, particularly for generic LLMs with no built-in safeguards.

AI-Specific Validation Approaches

- Effective validation of AI-assisted evidence synthesis requires safeguards designed around the specific AI failure modes while considering tool-specific capabilities and limitations. **Table 1** outlines recommended safeguards for validating AI-driven SLRs. Validation strategies should be calibrated according to the AI tool: purpose-built tools require checks on calibration accuracy, training representativeness, and threshold settings, while generic LLMs require intensive output verification, hallucination detection, and prompt sensitivity testing.

Table 1. AI-specific Validation Safeguards for SLRs

Safeguard	Description	Rationale
Control Prompt Design	Develop standardized, version-controlled prompts per task. Test on validation sets before deployment. Document iterations and effects on quality.	Prompt phrasing is a documented source of LLM inconsistency. Treating prompts as validated instruments mitigates a known failure mode and improves reproducibility.
Validate at Checkpoints	Validate AI outputs at natural workflow breakpoints (e.g., after each screening batch, extraction session, or synthesis pass) rather than only at task completion; check for drift in inclusion patterns, extraction accuracy, or synthesis claims before errors propagate	AI failures manifest as systematic drift not isolated errors: inconsistent criteria across batches, accumulating fabrications, or persistent template-driven outputs. End-stage review catches these too late to correct efficiently.
Target Task-Specific Failures	Design checks targeting known AI failure modes per stage. Searches: invalid/missing terms or syntax. Screening: inclusion drift/ prompt sensitivity. Extraction/RoB: verify values against source for fabrication, substitution, and prompt sensitivity. Synthesis: trace claims to supporting studies, checking for training-data contamination or unsupported assertions	Invalid syntax and hallucinated terms dominate searching; prompt sensitivity and inclusion drift dominate screening; fabrication and prompt sensitivity dominate extraction; training-data contamination and heterogeneity masking dominate synthesis. Generic validation misses stage-specific risk profiles. Even high-performing models produce plausible but incorrect details only caught by returning to source.
Audit Decision Patterns	Track and analyze patterns in AI decisions over time; monitor for clustering of similar judgments, unexpected shifts in inclusion rates, or disproportionate reliance on certain keywords	AI tools surfaces systematic biases that are invisible to spot-checks, particularly confidence-calibration failures producing consistent outputs across different inputs. They distinguish genuine evidence patterns from model artifacts.
Escalate to Humans	Define criteria for escalating AI outputs to human review: disagreement between multiple runs, low-confidence outputs, novel study designs outside training data, or borderline eligibility decisions	AI tools perform unevenly across contexts. Pre-defined escalation criteria prevent over-reliance in edge cases where failure risk is highest.

Abbreviations: AI = artificial intelligence; LLM = large language model; RoB = risk of bias

Conclusions

- AI tools fail in ways that often look superficially similar to human error but arise from distinct underlying mechanisms.
- Failure mechanisms also vary meaningfully between purpose-built platforms and generic LLMs.
- Much like economic models in health economics and outcomes research require calibration, validation, and skilled analysts, AI-assisted SLRs require technically proficient reviewers and layered validation protocols tailored to the specific tools in use.
- No single safeguard suffices; validation must be based on a thorough understanding that AI makes human-like errors, but on a thorough understanding of how AI actually fails.

Disclosures

Funding provided by Thermo Fisher Scientific. PZ and KS authors are employees of PPD™ Evidera™ Health Economics & Market Access, Thermo Fisher Scientific at the time this study was completed. AC is an independent consultant.

Acknowledgments

Editorial and graphic design support were provided by Caroline Cole and Kawthar Nakayima of Thermo Fisher Scientific.

© 2026 Thermo Fisher Scientific Inc. All rights reserved. All trademarks are the property of Thermo Fisher Scientific and its subsidiaries unless otherwise specified. This information is not intended to encourage use of these products in any manner that might infringe the intellectual property rights of others.