ScitoVation
INNOVATIVE CELL BASED SCIENCE

ScitoVation Insight | Issue 1

Large Language Models in Information Synthesis: ScitoVation's Human in the Loop Approach

A growing number of artificial intelligence (AI) tools may search, screen, extract, summarize, and synthesize scientific information faster than traditional review methods. In life sciences, the volume of evidence is increasing faster than organizations can efficiently assess and act on it, so the appeal of these tools is clear. Speed, however, is not the same as scientific completeness. A system can generate an answer or report quickly while still missing relevant evidence, obscuring uncertainty, or extending beyond the context in which its performance has been demonstrated.

Adoption in regulated scientific environments remains cautious, and recent research helps explain why. The performance of Large Language Models (LLMs) varies across tasks, models, prompts, evidence bases, and workflows, creating challenges in circumstances where findings must be reproducible, traceable, and scientifically defensible (1–7). In that setting, the quality of an AI-generated output cannot be judged by its fluency or apparent completeness alone; the underlying evidence, assumptions, and limits of the workflow must remain inspectable. These limitations highlight the need for a Human-in-the-Loop (HITL) approach where AI accelerates defined tasks while scientists retain oversight of uncertainty, interpretation, and consequential decisions.

In this article, we examine what recent evidence tells us about where LLMs work well, where limitations remain, and how ScitoVation is addressing those challenges.

What recent evidence tells us about what works, and what doesn't

1. LLMs perform best on bounded, verifiable tasks. Individual studies have reported high sensitivity in title/abstract and full-text screening, while lower specificity and greater error have been reported in some full-text screening applications and in tasks such as literature searching and risk-of-bias assessment. (1–5) (Figure 1). These results suggest that LLMs can reduce the burden of repetitive information processing, particularly when the task has defined criteria and outputs can be checked against source material.

2. Performance does not automatically transfer from one use case to another. Some models have been shown to change screening decisions across repeated runs or lose accuracy when applied in a different review context (1–4). Fine-tuning and carefully designing prompts improved performance in some settings, but success remained tied to the evidence base and workflow tested (1–4). Performance must therefore be validated for the specific task, evidence base, scientific domain, and decision context in which the technology will be used; success in one workflow should not be assumed to demonstrate reliability in another.

3. Trust depends on the workflow, not just the model. One study processed 734 full-text PDFs using predefined fields, source-linked evidence, and flags for conflicts or unsupported information (6). That shifts the focus from model accuracy alone to whether outputs are controlled, traceable, and easy for a scientist to verify. A polished final report is therefore not, by itself, evidence of a defensible process. Defensibility depends on whether a reviewer can reconstruct where the evidence came from, how it was handled, and how the conclusion was reached. Even if no errors are found in a reviewed sample, that does not prove the true error rate is zero. For example, with no errors observed in 60 reviewed items, the upper one-sided 95% confidence bound on the true error rate was still 4.9% (5).

4. Scientific interpretation still requires human judgment. SciConBench found that even advanced models and deep-research agents produced incomplete and sometimes contradictory conclusions compared with expert-written reviews (7). Identifying or extracting what a study reports is not the same as deciding what an entire body of evidence means.

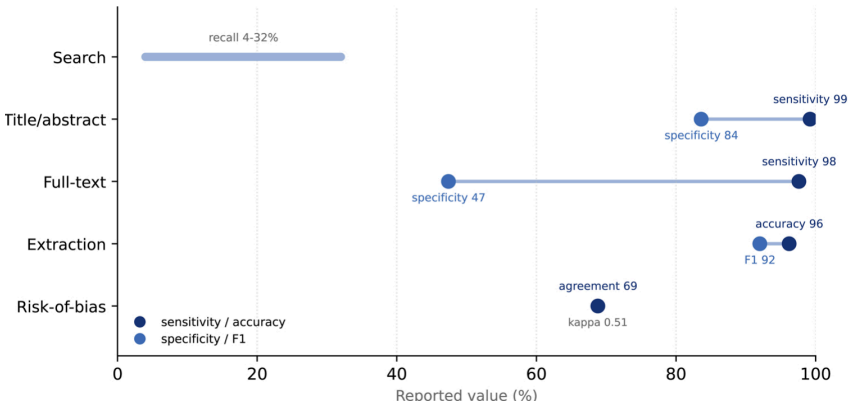

Figure 1. What the evidence shows across the information-synthesis workflow. High sensitivity has been reported for title/abstract and full-text screening, although specificity and other performance measures vary across studies. Performance has generally been less consistent for literature searching and risk-of-bias assessment. Reported results differ across studies, models, datasets, prompts, and evaluation criteria; values shown should therefore not be interpreted as pooled or generalizable estimates of AI performance. From Descamps et al., 2026 (5).

The current literature points to task-specific validation, source-level traceability, explicit handling of uncertainty and conflicts, and human scientific oversight as essential for effective AI-assisted information synthesis. These controls should be viewed as part of the scientific workflow itself rather than as optional software features added after an AI system has produced its answer.

How ScitoVation's information synthesis tool addresses these limitations

These findings have shaped ScitoVation's information-synthesis approach (Figure 2). Rather than treating an LLM as an autonomous reviewer or a report-generation engine, our tool is designed around the limitations identified in the literature, specifically variability, uncertainty, the need for provenance, and the continued importance of scientific judgment.

Our workflow uses AI to accelerate defined activities such as literature intake, screening, full-text review, structured extraction, and evidence organization, while scientists retain responsibility for scope, verification, interpretation, and final conclusions. The goal is not to maximize the number of automated steps, but to shorten the path to a defensible scientific decision while keeping the evidence inspectable at every consequential stage.

Auditability is necessary, but it is not sufficient. A workflow can be fully traceable and still produce a weak scientific conclusion if the underlying evidence is not interpreted in the appropriate toxicological and exposure context. ScitoVation therefore combines traceable AI-assisted workflows with domain-specific scientific review, so that clear sourcing and oversight strengthen, rather than replace, toxicological interpretation.

Where the scientific question requires it, synthesized evidence can also be evaluated alongside transcriptomic, biological-similarity, PBPK, and qIVIVE analyses rather than remaining a standalone literature report.

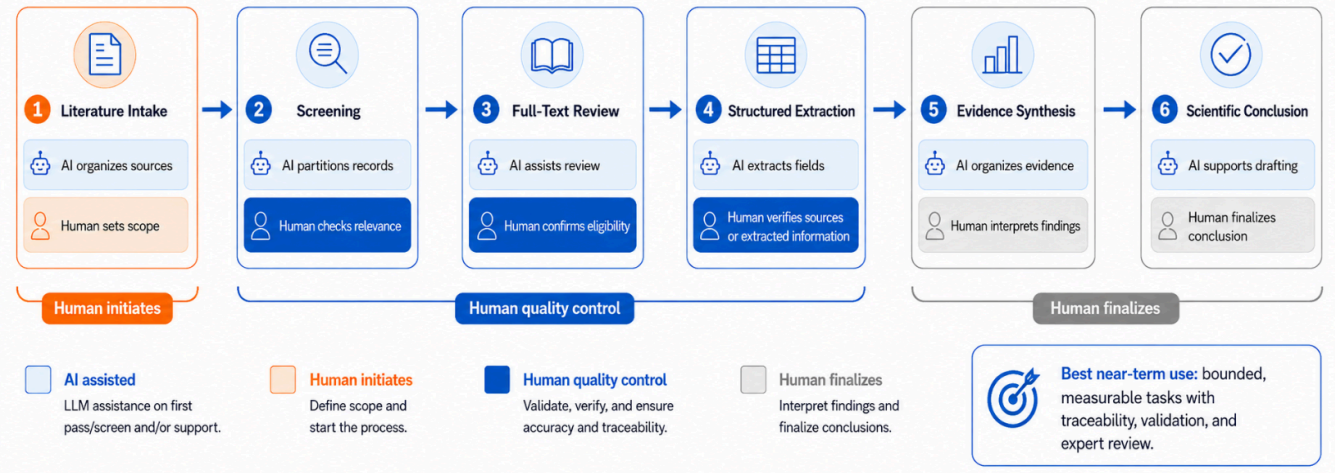

Figure 2. ScitoVation's HITL model for LLM-assisted information synthesis. AI supports defined stages of the workflow, including literature intake, screening, full-text review, structured extraction, and evidence synthesis. Humans define scope, perform quality control, interpret findings, and retain responsibility for final scientific conclusions.

Key features of ScitoVation's information synthesis tool:

1. Bounded, reviewable automation. AI is applied to tasks such as screening and structured extraction where criteria can be clearly defined and outputs checked against source material.

2. Context-specific validation. Performance is evaluated within the intended workflow and evidence base rather than assumed to transfer from another model, prompt, or use case.

3. Traceability and auditability. Source-linked outputs, structured fields, conflict flags, and documented review steps help scientists verify how information was identified and synthesized.

4. Human judgment at critical points. Scientists remain responsible for resolving uncertainty, reviewing critical findings, evaluating conflicting evidence, and interpreting the overall evidence base.

Together, these features are intended to accelerate routine information-processing tasks while preserving the traceability, verification, and scientific judgment needed for regulated scientific use.

Conclusion

Current evidence does not support replacing scientists with end-to-end LLM automation for information synthesis. It does, however, show that LLMs can accelerate defined stages, particularly screening and structured extraction, when performance is validated, outputs are traceable, and experts retain responsibility for interpretation and final conclusions. Neither speed, source volume, nor report length is a substitute for demonstrated scientific reliability.

Ultimately, the most effective information-synthesis systems may not be those that automate the most. They will be those that automate the right tasks while keeping the underlying evidence traceable, uncertainty visible, and human oversight built into the workflow.

At ScitoVation, these are the principles guiding our information-synthesis tool. We use AI to accelerate the work that machines can perform well, while keeping scientists firmly in control of the evidence and the conclusions drawn from it. The objective is not simply to produce information faster, but to make fragmented evidence usable for a robust and grounded toxicological or risk-assessment decision.

If you're exploring how LLMs could support your information-synthesis workflow, talk to a scientist. ScitoVation's HITL tool can help turn fragmented literature, regulatory documents, and scientific data into clear, structured, source-linked evidence that scientists can interrogate, interpret, and use in toxicology and risk-assessment decisions.

References

1. Yamoah K, Schroeder N, Dorley E, Rani N, Schutz C. Fine-Tuning A Large Language Model for Systematic Review Screening. arXiv. 2026 Mar 25. doi:10.48550/ARXIV.2603.24767

2. Hida GS, Ribeiro DM, Yahata E. Beyond Accuracy: LLM Variability in Evidence Screening for Software Engineering SLRs. arXiv. 2026 Apr 29. doi:10.48550/ARXIV.2604.27006

3. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT -4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Research Synthesis Methods. 2024 Jul;15(4):616–26. doi:10.1002/jrsm.1715

4. Fleurence RL, Qureshi R, Aggarwal R, Bian J, Cheyne S, Dawoud D, et al. The Use of Generative Artificial Intelligence in Systematic Literature Reviews: A Rapid Review of the Literature. Value in Health. 2026 Jul;S1098301526025325. doi:10.1016/j.jval.2026.06.012

5. Descamps J, Bouguennec N, Porcher R, Resche-Rigon M, Bouché PA. Artificial intelligence in systematic reviews and meta-analyses: Task-specific performance, residual error quantification, and human oversight. Orthopaedics & Traumatology: Surgery & Research. 2026 Jul;104793. doi:10.1016/j.otsr.2026.104793

6. Mortezaagha P, Shaw J, Sun B, Rahgozar A. From chaos to clarity: schema-constrained AI for auditable biomedical evidence extraction from full-text PDFs. BMC Med Res Methodol. 2026 Apr 14;26(1):119. doi:10.1186/s12874-026-02847-8

7. Jung H, Diniz PV, Roveda JRC, da Silva AF, Jung H, Tsai E, et al. Can AI Agents Synthesize Scientific Conclusions? arXiv. 2026 Jun 9. doi:10.48550/ARXIV.2606.11337