

AI-Assisted Literature Reviews in Higher Education: What the Science of Informatics Tells Us (and What It Doesn't)

John A. Nunnery 

Educational Leadership & Workforce Development, Old Dominion University, United States.

How to cite: Nunnery, J. A. (2026). AI-Assisted Literature Reviews in Higher Education: What the Science of Informatics Tells Us (and What It Doesn't). In: 12th International Conference on Higher Education Advances (HEAd'26). Valencia, 15-18 June 2026. <https://doi.org/10.4995/HEAd26.2026.21583>

Abstract

Generative AI tools are increasingly used by higher-education faculty and graduate students to accelerate literature reviews, particularly during abstract screening. Most evidence on AI-assisted screening, however, comes from informatics and systematic review methodology, where tools are evaluated under idealized, retrospective conditions that differ from real faculty workflows. This paper conducts a translational synthesis of that evidence, reframing abstract screening as a governance problem under epistemic constraint: abstracts are incomplete representations that impose an upper bound on screening reliability for both humans and AI. Drawing on four evidence domains (human screening error, validation regimes, stopping rules, and reproducibility), we show why reported performance gains cannot be taken at face value and why simulation-based accuracy does not ensure defensible practice. We derive five practice-oriented guidelines grounded in this framework: prioritizing low-inference measurement, separating filtering from exclusion, preserving uncertainty, documenting screening configurations, and calibrating AI use to review risk tolerance.

Keywords: *AI-assisted screening; Abstract screening; Human-in-the-loop machine learning; Research governance; Systematic review.*

1. Introduction

Across higher-education research, faculty and graduate students are increasingly experimenting with generative artificial intelligence (AI) tools to manage the growing scale of scholarly literature. Abstract screening has become a focal point of this experimentation because it is labor-intensive, repetitive, and often perceived as a bottleneck in review workflows (Gates et al., 2019; Tsou et al., 2020). In practice, AI tools are frequently framed as efficiency aids: mechanisms for accelerating screening, reducing workload, and shortening time to synthesis.

Most evidence on AI-assisted screening comes from informatics and systematic review methodology, where tools are evaluated under idealized, retrospective conditions using

historical inclusion decisions as fixed reference standards (Ferdinands et al., 2023; König et al., 2025). These evaluations differ from faculty-led workflows in important ways: they assume small teams, limited methodological support, stable eligibility criteria, and homogeneous domains.

In this paper, abstract screening is reframed as a governance problem under epistemic constraint. Reviewers do not evaluate studies directly; they evaluate incomplete representations (titles, abstracts, and limited metadata) that frequently omit or ambiguously report eligibility-relevant information, producing non-trivial error and disagreement even among trained reviewers (Wang et al., 2020; Waffenschmidt et al., 2019).

This framing has two implications. First, it imposes an upper bound on the reliability of abstract-only screening for both humans and AI: no screening system can recover information that is not present. Second, it shifts attention away from tool accuracy alone toward the design and governance of screening workflows. The critical question is not whether an AI system can reproduce historical labels, but whether the screening process remains defensible, auditable, and aligned with the review's purpose (van de Schoot et al., 2021; Lombaers et al., 2024).

Accordingly, this paper asks a practical question of direct relevance to higher-education faculty: what does the science of informatics tell us about when AI-assisted abstract screening is reliable, where its limits lie, and how these tools should be used responsibly in real research workflows? To address this question, we conduct a translational synthesis drawing on informatics research across four evidence domains: human screening error, validation regimes, stopping rules, and reproducibility. We use this synthesis to show why different forms of “validity” are not interchangeable and why efficiency gains must be interpreted through a governance lens. Building on this analysis, we offer concrete, practice-oriented guidance for faculty and graduate researchers, emphasizing measurement-first uses of AI, separation of filtering from exclusion, preservation of uncertainty, and documentation of screening configurations.

2. Screening as Governance Under Epistemic Constraint

These representational limits make early exclusion decisions disproportionately consequential: once a record is excluded at the abstract stage, it rarely re-enters the review.

The relevant evaluative question is not whether an AI system reproduces historical labels, but whether the screening process as a whole remains defensible, auditable, and aligned with the review's purpose (van de Schoot et al., 2021; Lombaers et al., 2024; O'Connor et al., 2019; O'Mara-Eves et al., 2015).

3. Abstracts as Samples, Not Proxies

A central premise of this paper is that abstracts should be treated as samples of underlying studies rather than as faithful proxies. Empirical research shows that abstracts frequently omit or ambiguously report key methodological features, leading to disagreement and error during abstract screening even among trained human reviewers (Wang et al., 2020; Waffenschmidt et al., 2019).

This has direct implications for screening reliability. If eligibility-relevant information is absent from the abstract, no classifier—human or automated—can recover it with certainty. The achievable reliability of abstract-only screening is therefore bounded by representational limits rather than reviewer competence alone (Wang et al., 2020; Dennstädt et al., 2024).

This motivates a distinction between measurement validity and decision validity. Measurement tasks concern what is explicitly stated in the abstract, while decision tasks concern whether a study should be included. Informatics studies implicitly rely on this distinction when demonstrating that screening “errors” often reflect abstract reporting limitations rather than reviewer negligence (Waffenschmidt et al., 2019).

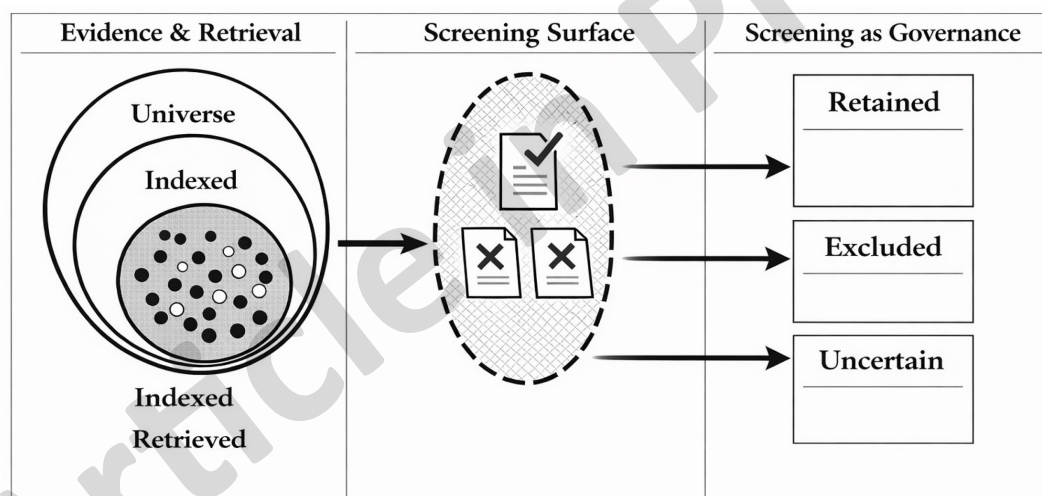


Figure 1. Abstract screening operates on a thin, incomplete sample of the evidence universe, requiring governance decisions under uncertainty as retrieved records are partitioned into inclusion, exclusion, and indeterminate states.

4. Validation Regimes and the Limits of Performance Claims

Much of the evidence base on AI-assisted screening evaluates systems under retrospective, simulation-based conditions, using historical inclusion decisions as fixed reference standards (Ferdinands et al., 2023; König et al., 2025). These simulations eliminate sources of real-world

variability such as reviewer disagreement, uncertainty, and fatigue, thereby inflating apparent certainty.

Informatics research distinguishes among simulation validity, pipeline validity, and governance validity—each addressing different risks and decision contexts (Ferdinands et al., 2023; König et al., 2025; van de Schoot et al., 2021). For higher-education researchers, the critical implication is that validation regimes are not interchangeable. Performance in retrospective simulations does not guarantee performance in live workflows, nor does it establish that a process is auditable or defensible under scrutiny (van de Schoot et al., 2021; Lombaers et al., 2024).

Many validation studies use oracle-derived human labels rather than actual reviewer inputs, making them useful for characterizing algorithmic behavior under idealized conditions but not for estimating live-workflow performance. They do not capture reliability attenuation from real reviewer disagreement and uncertainty, so their performance estimates necessarily overstate live-workflow results to an indeterminate degree.

5. Measurement-First Logic for Responsible AI Use

A measurement-first approach prioritizes AI for low-inference extraction tasks (asking what is explicitly stated in an abstract) rather than holistic eligibility judgments (deciding whether a study qualifies overall). The literature on screening prioritization supports this conservative stance: low-inference tasks show more stable performance across domains and prevalence conditions than eligibility judgments, while stopping rules and auto-exclusion thresholds encode explicit trade-offs between workload reduction and miss-risk (Gates et al., 2019; Tsou et al., 2020; Callaghan & Müller-Hansen, 2020; Boetje & van de Schoot, 2024).

Reproducibility studies further emphasize that AI-assisted workflows must preserve decision traces, tool versions, and configuration details to remain interpretable and auditable (van de Schoot et al., 2021; Lombaers et al., 2024).

6. Practical Implications for Higher-Education Faculty

The informatics evidence reviewed above does not imply that faculty should avoid AI tools in literature reviews. Rather, it implies that how AI is used matters more than which tool is used. Each guideline below integrates the relevant governance constraint with its practical implication.

6.1. Prioritize low-inference measurement over eligibility judgment.

Because abstracts are samples rather than proxies and AI cannot recover absent information, assistance is most defensible for low-inference measurement tasks: asking whether specific

information is explicitly present (publication year, study type, population keywords, named methods or outcomes) rather than whether a study qualifies overall. Such tasks show more stable performance across domains and prevalence conditions (Gates et al., 2019; Tsou et al., 2020; Dennstädt et al., 2024).

6.2. Separate filtering from exclusion.

Because abstract-stage exclusions are irreversible and governance requires human authority over final decisions, AI tools should filter or prioritize records without substituting for human exclusion decisions. Records below any AI-assigned threshold should still be sampled or batch-reviewed before permanent exclusion, converting efficiency gains into workload reduction without introducing hidden miss-risk (Callaghan & Müller-Hansen, 2020; Boetje & van de Schoot, 2024).

6.3. Preserve uncertainty as a screening outcome.

Because representational limits are irreducible and screening uncertainty reflects missing information rather than reviewer error, teams should code a third outcome category (uncertain / needs full text) and route those records to full-text retrieval rather than forcing binary include/exclude decisions at the abstract stage (Wang et al., 2020; Waffenschmidt et al., 2019).

6.4. Document the screening configuration.

Because validation regimes are not interchangeable and AI configuration choices materially affect results, workflows must document the tool used, prompt or criteria text, decision thresholds, and any iteration changes. A methods appendix or project log recording these details keeps the screening process auditable and reproducible (van de Schoot et al., 2021; Lombaers et al., 2024).

6.5. Calibrate AI use to review purpose and risk tolerance.

Because miss-risk tolerance varies by review purpose, governance thresholds should reflect the stakes of the work. A doctoral seminar mapping themes may use AI-assisted prioritization to accelerate exploration; a grant-funded systematic review informing institutional policy should adopt conservative thresholds, human oversight, and audit sampling of excluded records (Boetje & van de Schoot, 2024; O'Connor et al., 2019).

Table 1 provides a summary of these recommendations.

Table 1. Summary recommendations for using AI in Abstract Screening for Systematic Reviews

Recommendation	Underlying Constraint	Governance Implication
6.1 Prioritize low-inference measurement	Abstracts are not proxies; AI cannot recover absent information	AI is most reliable on what is explicitly stated
6.2 Separate filtering from exclusion	Abstract-stage exclusions are irreversible; governance requires human authority over final decisions	AI filtering should not silently substitute for human exclusion decisions
6.3 Preserve uncertainty as a screening outcome	Representational limits are irreducible; screening uncertainty reflects missing information, not reviewer error	Uncertainty is informative and should be preserved, not forced into binary decisions
6.4 Document the screening configuration	Validation regimes are not interchangeable; AI configuration choices materially affect results	Workflows must be documented to remain auditable and reproducible
6.5 Calibrate AI use to review purpose and risk tolerance	Miss-risk tolerance varies by review purpose; governance thresholds should reflect stakes	AI use should be calibrated to the review's acceptable level of miss-risk

7. Conclusion

AI-assisted literature review is a methodological and governance problem shaped by representation limits, reviewer error, validation regimes, and audibility requirements (Wang

et al., 2020; Ferdinands et al., 2023; O'Connor et al., 2019). Responsible use requires shifting from a tool-centered question (Is this model accurate?) to a workflow-centered one (What decisions are being delegated, under what constraints, and with what audit trail?). The five guidelines above translate this governance framework into practice, enabling faculty to reduce workload without compromising the defensibility or trustworthiness of their reviews.

References

- Boetje, J., & van de Schoot, R. (2024). The SAFE procedure: A practical stopping heuristic for active learning–based screening in systematic reviews and meta-analyses. *Systematic Reviews*, 13(1), Article 14. <https://doi.org/10.1186/s13643-024-02406-7>
- Callaghan, M. W., & Müller-Hansen, F. (2020). Statistical stopping criteria for automated screening in systematic reviews. *Systematic Reviews*, 9(1), Article 316. <https://doi.org/10.1186/s13643-020-01521-4>
- Dennstädt, M., Moher, D., Vach, W., Schäfer, M., Seidler, A. L., Löffler, C., et al. (2024). Title and abstract screening using large language models: A methodological review. *Systematic Reviews*, 13(1), Article 158. <https://doi.org/10.1186/s13643-024-02575-4>
- Ferdinands, G., Thomas, J., Marshall, I. J., Noel-Storr, A., Elliott, J. H., & Wallace, B. C. (2023). Performance of active learning models for screening prioritization in systematic reviews: A simulation study. *Systematic Reviews*, 12(1), Article 14. <https://doi.org/10.1186/s13643-023-02130-7>
- Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. *Systematic Reviews*, 8(1), Article 278. <https://doi.org/10.1186/s13643-019-1222-2>
- König, L., Zitzmann, S., & Hecht, M. (2025). Strategizing AI utilization for psychological literature screening: Balancing efficiency, sensitivity, and transparency. *Research Synthesis Methods*, 16(1), 123–138. <https://doi.org/10.1002/jrsm.1672>
- Lombaers, P., de Bruin, J., & van de Schoot, R. (2024). Reproducibility and data storage for active learning–aided systematic reviews. *Applied Sciences*, 14(2), Article 655. <https://doi.org/10.3390/app14020655>
- O'Connor, A. M., Tsafnat, G., Thomas, J., Glasziou, P., Gilbert, S. B., & Hutton, B. (2019). A question of trust: Can we build an evidence base to gain trust in systematic review automation technologies? *Systematic Reviews*, 8(1), Article 143. <https://doi.org/10.1186/s13643-019-1062-0>
- O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: An evaluation of four tools. *Systematic Reviews*, 4(1), Article 5. <https://doi.org/10.1186/2046-4053-4-5>
- Tsou, A. Y., Treadwell, J. R., Erinoff, E., & Schoelles, K. M. (2020). Machine learning for screening prioritization in systematic reviews: Comparative performance of classification models. *Systematic Reviews*, 9(1), Article 1. <https://doi.org/10.1186/s13643-019-1253-8>
- van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdemans, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., &

- Tummers, L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2), 125–133. <https://doi.org/10.1038/s42256-020-00287-7>
- Wang, Z., Nayfeh, T., Tetzlaff, J., O’Blenis, P., Murad, M. H., & Waffenschmidt, S. (2020). Error rates of human reviewers during abstract screening in systematic reviews. *PLOS ONE*, 15(1), e0227741. <https://doi.org/10.1371/journal.pone.0227741>
- Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S., & Pieper, D. (2019). Single screening versus conventional double screening for study selection in systematic reviews: A methodological systematic review. *BMC Medical Research Methodology*, 19(1), Article 132. <https://doi.org/10.1186/s12874-019-0782-0>