Research Synthesis Methods (2026), **17**: 1297–1301
doi:10.1017/rsm.2026.10114

Research Synthesis Methods

LETTER TO THE EDITOR

Reframing workload saving in LLM-assisted screening: A commentary on Sciurti et al.

Honghao Lai[1], Jiayi Liu[1,2], Long Ge[3] and Janne Estill[4]

[1]Department of Health Policy and Management, Lanzhou University, China
[2]Evidence-Based Social Science Research Center, Lanzhou University, China
[3]Lanzhou University, China
[4]University of Geneva, Switzerland

Corresponding author: Long Ge; Email: gelong2009@163.com

Received: 5 May 2026 UTC; **Revised:** 24 June 2026 UTC; **Accepted:** 14 July 2026 UTC; published online 27 August 2026 UTC

Dear Editors

Sciurti and colleagues recently reported in this journal a careful evaluation of three compact large language models, GPT-4o mini, Llama 3.1 8B, and Gemma 2 9B, applied to title and abstract screening across three published systematic reviews on vaccine literacy, antibiotic-resistant *Acinetobacter baumannii*, and vitamin supplementation in COVID-19.[1] The three corpora differ in size, topic, and prevalence of eligible records, and the authors evaluated each model individually as well as a majority-voting ensemble at predefined rating thresholds. The headline finding, and the one most likely to shape practice, is the estimated workload reduction of 79% to 91% achieved by the ensemble at a 50-rating threshold across the three reviews, while sensitivity remained high. Compact models that run on local hardware at low cost, with savings of this magnitude, would be a meaningful addition to the systematic reviewer's toolkit.

This figure deserves close reading, however. It is a retrospective summary computed against ground-truth labels that already exist in the published reviews, and it describes the model's classification behaviour rather than the time a review team will save when deploying the workflow prospectively. The two are related, but they are not the same quantity. Whether the retrospective figure transfers to a real workflow depends on how the team chooses to handle the records the model excluded, and that choice is not visible in the metric itself.

The metric used by Sciurti et al., defined as the proportion of records correctly excluded by the model out of the total, follows the convention proposed by Syriani et al. and is now widely adopted in this literature.[2] As a property of the classifier, it is well defined and easy to compute. The risk is that, when reported as "workload saving", it invites a reading it cannot quite support. A practitioner encountering "79% to 91% workload saving" may reasonably infer that, by adopting the workflow, the team will only need to screen the remaining 9% to 21% of citations, on the grounds that the model has correctly removed the rest. The arithmetic of the metric encourages this reading, since saved and remaining work appear to add up to one. Yet the inference rests on a step the metric does not justify, namely the step of treating the model's excluded set as work that no longer needs to be done.

Systematic reviews are designed to identify, as completely as possible, all studies that meet the eligibility criteria, and methodological guidance treats missing an eligible study as a more serious error than retaining an irrelevant one. Sciurti et al. acknowledge this asymmetry explicitly when they argue that sensitivity should be prioritised and imprecision tolerated. The implication, however, runs further

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Society for Research Synthesis Methodology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/rsm.2026.10114 Published online by Cambridge University Press

than the design of the prompt. It also constrains how the team must treat the excluded set after the model has spoken. If completeness is to be preserved, exclusions cannot simply be discarded; they have to be either trusted or checked. To realise the estimated saving of 79% to 91% in practice, the review team must accept the residual risk of missed studies that comes attached to it. This risk is quantified by one minus the model's sensitivity in the corpus at hand. With Sciurti et al.'s ensemble configuration, sensitivity against the original title and abstract decisions ranged from 90.4% to 94.4%, which corresponds to expected false negative rates of roughly 5.6% to 9.6% within the eligible records of each review. Whether such a rate is acceptable is a judgement the metric cannot make for the team. If the team accepts it, the saving figure transfers to practice. If the team does not, the figure does not transfer.

The case in which the team does not accept the residual risk deserves separate attention, because it is here that the gap between the reported metric and real workload becomes visible. When a team requires complete recovery of any false negatives in the excluded set, the model offers no reliable means of locating them: confidence scores and other internal signals were already used to produce the exclusion decision, and tools such as short rationales or supporting excerpts may reduce the time to inspect individual records but cannot themselves identify which records are safe to skip. For a team that will not accept any risk of having missed an eligible study, this means that every excluded record must be re-examined—unless the workflow can partition the excluded set in a way that provides a verifiable guarantee of complete sensitivity within at least part of it. When no such partition is available, the entire excluded set remains in scope for verification.

Once verification covers the full excluded set, the arithmetic of the workload reduction metric breaks down. The work being verified is no longer limited to the small portion the model handled wrongly; it is the entire portion the model handled at all. In the configuration above, that means roughly 80% to 95% of each corpus has to be re-examined in addition to the records the model retained. The time saved at the LLM step is offset by the time spent at the verification step, and the net workload reduction shrinks well below the headline 79% to 91%. In the limiting case, where the team trusts no part of the excluded set without re-examination, the saving approaches zero. The same model, the same prompt, and the same data therefore support very different statements about workload, and the difference is driven entirely by the verification policy the team has adopted.

The link between classifier behaviour and real workload reduction can be made precise by introducing the idea of a verification-exempt subset. This is the part of the excluded set for which the workflow can credibly assert that no eligible study has been missed, either because the subset has been empirically shown to have 100% sensitivity, or because the design provides another form of guarantee. Records outside this subset still require human checking. The larger the verification-exempt subset, the larger the genuine saving, and a workflow that cannot identify any verification-exempt subset at all cannot, strictly speaking, claim a workload reduction beyond what the team is willing to take on as missed-study risk.

Building on this, we suggest a simple verification-aware measure. Let $W_{manual} = n \cdot t$ denote the baseline workload of fully manual screening, where n is the total number of records and t the time per record. The workload of the LLM-assisted workflow can be written as

$$W_{LLM} = T_{setup} + n_{keep} t_{keep} + q_{exc} n_{exc} t_{verify}$$

Here, T_{setup} is the one-time cost of deploying the LLM workflow, encompassing prompt development and testing, any calibration runs needed to select the operating threshold, and the time or compute cost of running the model on the full corpus; n_{keep} and n_{exc} are the numbers of records retained and excluded, respectively; q_{exc} is the proportion of excluded records that the team's pre-specified verification policy designates for human review—a value set at the protocol stage, before the model is run, as a commitment about verification effort rather than a claim about classifier performance; and t_{keep} and t_{verify} are the unit times for reviewing a retained record and verifying an excluded record, respectively. Note that t_{verify} may be shorter than t_keep when LLM-generated confidence scores,

https://doi.org/10.1017/rsm.2026.10114 Published online by Cambridge University Press

brief rationales, or supporting excerpts are available to guide each verification judgment. We therefore propose a new metric, the verification-adjusted workload reduction (VAWR), defined as:

$$\mathrm{VAWR} = 1 - \frac{\mathrm{W_{LLM}}}{\mathrm{W_{manual}}}$$

VAWR is a workload-accounting metric. It takes q_{exc} as a given input and returns the workload reduction conditional on that verification policy. It does not specify what q_{exc} should be, nor does it validate whether a chosen verification strategy provides sufficient statistical assurance against false negatives. Those are separate questions that require a sampling design, a sample size justification, and a principled risk threshold.

Three features of this expression are worth noting. First, q_{exc} is where verification policy enters the calculation, and it can take any value between 0 and 1, allowing a single formula to describe a spectrum of strategies from full trust to full re-screening. Second, when q_{exc} approaches 0, T_{setup} is negligible, and $t_{keep} \approx t_{verify} \approx t$, VAWR approaches n_{exc}/n—the proportion of all records excluded by the model, which is the quantity a prospective team can directly observe. This differs from the conventional workload-saving metric (TN/N), which counts only correctly excluded records and requires post-hoc access to ground-truth labels. The two quantities coincide only when sensitivity is perfect (FN = 0); when false negatives are present, n_{exc}/n slightly exceeds TN/N because it credits the workflow for exclusions that were, in retrospect, erroneous. This is precisely the condition under which the trust assumption embedded in $q_{exc} = 0$ is consequential: the prospective team cannot distinguish true negatives from false negatives within the excluded set, so the workload they forgo includes both. Third, when q_{exc} approaches 1, even a classifier with high specificity yields little practical reduction, because the excluded records must be looked at anyway. The metric thereby makes explicit the conditions under which the standard workload saving figure can be trusted as an estimate of real-world time savings.

Applied to the three reviews in Sciurti et al., the metric reveals a wide range of plausible savings depending on which verification policy a team adopts. We focus on the majority-voting configuration at the 50-rating threshold, which the authors recommend as a reasonable default when an optimal threshold is not known in advance. For simplicity, we assume $T_{setup} \approx 0$ and $t_{keep} \approx t_{verify} \approx t$, so that differences across scenarios are driven entirely by q_{exc}. The numbers below are recomputed from values reported in their Table 2.

Consider first the conservative position, in which title and abstract decisions made by the original reviewers are taken as the standard of interest, and any sensitivity below 100% triggers full re-screening of the excluded set. Under this policy $q_{exc} = 1$ in the VL review (majority-voting sensitivity 90.4%), the AB review (sensitivity 91.3%), and the COVID-19 review (sensitivity 94.4%). VAWR in all three cases drops well below the reported workload savings of 86.1%, 78.8%, and 87.4%. Now consider a more permissive position, in which full-text inclusion is treated as the standard of interest. Sciurti et al. report 100% sensitivity for majority voting at the 50-rating threshold against this reference in all three reviews, so $q_{exc} \approx 0$ and VAWR returns to the neighbourhood of the reported figures. A middle position, such as a 10% random audit of the excluded set ($q_{exc} = 0.10$), reduces the verification burden and places VAWR between these two extremes. It is important, however, to distinguish the workload implications of such an audit from its evidentiary value. Setting $q_{exc} = 0.10$ determines the VAWR calculation, but the inferential claim that the remaining 90% of exclusions are safe to skip without further checking depends on whether the audit was statistically powered for that purpose. Finding zero false negatives in a 10% sample does not by itself establish an acceptable false-negative rate in the unverified portion unless the sample was sized against a prespecified confidence bound on the false-negative rate. A partial audit therefore functions as a practical verification policy with computable workload implications, but its inferential claims require a sampling design and sample size justification to be stated alongside the VAWR figure.

The same workflow, then, supports very different statements about workload reduction depending on a choice that has nothing to do with model behaviour. We do not read this as a flaw of Sciurti et al.'s

https://doi.org/10.1017/rsm.2026.10114 Published online by Cambridge University Press

analysis, since their numbers are correct given the metric they used. We read it as a sign that the metric itself needs a companion piece of information, namely the verification policy under which it is to be interpreted.

A second observation arises from the full-text comparison itself. In all three reviews, several configurations achieved 100% sensitivity against full-text inclusion, which means that, in retrospect, no truly eligible study was lost in the LLM exclusions for these corpora. The result is encouraging but retrospective. A reviewer running a new review prospectively cannot know in advance whether such full-text-level sensitivity will hold for their corpus, since it depends on properties of the literature that are only revealed at the full-text stage. Prudent practice will assume it does not until evidence accrues. VAWR is meant to capture this prospective uncertainty by making the verification policy, rather than a hoped-for retrospective property, the basis on which workload claims are made.

Several practical implications follow. For review teams, q_{exc} should be specified at the protocol stage as part of the verification policy, alongside the screening method and conflict-resolution rules. Several considerations can inform this pre-specification. The primary driver is the acceptable false-negative rate for the review in question: a Cochrane review aiming for comprehensive evidence synthesis may warrant lower tolerance for missed studies—and therefore a higher q_{exc}—than a rapid review supporting a time-sensitive policy decision. Where published validation data exist for a similar LLM configuration and corpus type, the observed sensitivity in those studies provides an empirical anchor for estimating the residual false-negative rate at a given q_{exc}. Team size and reviewer experience enter as practical constraints rather than as risk parameters: a larger or more experienced team may be able to sustain a higher verification burden at lower per-record cost, affecting the feasibility of different q_{exc} choices without changing the acceptable risk threshold itself. The time parameters t_{keep} and t_{verify} can be estimated from pilot data or from published benchmarks, which typically report title-and-abstract screening at 5–40 seconds per record; t_{verify} may be shorter when LLM-generated outputs are available to guide the check. Stating in advance whether the team will accept LLM exclusions, audit a sample of them, or re-screen the full set turns workload saving from a property reported after the fact into a quantity that can be planned for, and it allows reviewers, peer reviewers, and editors to assess whether the chosen policy is appropriate to the question at hand. It is also worth distinguishing between two uses of q_{exc}: as a prospective policy commitment, in which the team prespecifies the verification fraction before running the model; and as an empirically informed quantity, in which the verification fraction is calibrated by the results of a validation study. In de novo reviews, only the first use is available in advance. The second requires a powered audit and a stated confidence bound on the false-negative rate in the unverified portion; its inferential implications are a separate question from the VAWR calculation and should be reported alongside it rather than folded into it.

For developers of LLM-assisted screening systems, lowering q_{exc} may be more valuable than further pushing classifier sensitivity.[3] Calibrated confidence outputs that flag low-confidence exclusions for human review, abstention mechanisms that route ambiguous records to an "uncertain" bucket rather than to the excluded set, and risk stratification of the excluded set are all design choices that shrink the part of the workflow needing verification without compromising completeness. Each of these moves a portion of the excluded set into a verification-exempt subset that can be defended on grounds other than blanket trust. Reducing T_{setup} is a complementary design goal. Workflows that require extensive prompt engineering, threshold calibration runs, or large pilot sets before deployment raise the setup cost term in VAWR, which erodes the net workload reduction particularly for smaller reviews where the corpus may not be large enough to amortise the initial investment. Streamlined deployment pipelines and validated default configurations that transfer across corpus types would widen the class of reviews for which LLM-assisted screening offers a genuine net saving.

More broadly, we suggest that workload saving be reported together with the pre-specified verification policy, the values of n_{keep} and n_{exc}, an estimate or bound on q_{exc} under that policy, and—where q_{exc} is derived from a sample audit rather than a blanket policy commitment—the statistical basis for that estimate: the sampling design, the achieved sample size, and the confidence interval on the false-negative rate in the unverified portion of the excluded set. None of this requires new infrastructure, and

https://doi.org/10.1017/rsm.2026.10114 Published online by Cambridge University Press

it would sit comfortably alongside the reporting items already promoted by recent guidance on AI in evidence synthesis,[4] where it could plausibly feature in future updates as the verification dimension of workload reporting.

Competing interest statement. The authors declare that no competing interests exist.

Data availability statement. No new data were generated or analysed in support of this research.

Funding statement. The authors declare that no specific funding has been received for this article.

Ethics statement. Ethical approval was not required for this manuscript because it is an opinion article and does not involve human participants, animal subjects, identifiable personal data, or any experimental procedures.

References

[1] Sciurti A, Migliara G, Siena LM, et al. Compact large language models for title and abstract screening in systematic reviews: An assessment of feasibility, accuracy, and workload reduction. *Res Synth Methods*. 2026;17(2): 332–347. https://doi.org/10.1017/rsm.2025.10044.

[2] Syriani E, David I, Kumar G. Screening articles for systematic reviews with ChatGPT. *J Comput Lang*. 2024; 101287–101287. https://doi.org/10.1016/j.cola.2024.101287.

[3] Marshall IJ, Wallace BC. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;8(1): 163. https://doi.org/10.1186/s13643-019-1074-9.

[4] Farotimi O, Dunn A, Van Lissa CJ, Polanin JR, Mavridis D, Pigott TD. Guidance for manuscript submissions testing the use of generative AI for systematic review and meta-analysis. *Res Synth Methods*. 2026;17(2): 237–239. https://doi.org/10.1017/rsm.2025.10058.

Cite this article: Lai H, Liu J, Ge L, Estill J. Reframing workload saving in LLM-assisted screening: A commentary on Sciurti et al.. *Research Synthesis Methods*. 2026;17: 1297–1301. https://doi.org/10.1017/rsm.2026.10114

https://doi.org/10.1017/rsm.2026.10114 Published online by Cambridge University Press