

Prospective validation of NeutrinoReview for LLM-assisted systematic review screening: an indoor air quality case study

Elias Sandner

`elias.sandner@cern.ch`

European Organization for Nuclear Research

Luca Fontana

World Health Organization

Michael T. Solomon

Università degli studi di Cassino e del Lazio Meridionale

Sumeya B. Abdella

Università degli studi di Cassino e del Lazio Meridionale

Elisa Caracci

Università degli studi di Cassino e del Lazio Meridionale

Luca Stabile

Università degli studi di Cassino e del Lazio Meridionale

Giorgio Buonanno

Università degli studi di Cassino e del Lazio Meridionale

Alice Simniceanu

World Health Organization

Igor Jakovljevic

European Organization for Nuclear Research

Andre Henriques

European Organization for Nuclear Research

Andreas Wagner

European Organization for Nuclear Research

Christian Gütl

Graz University of Technology

Research Article

Keywords: systematic review, screening, large language models, artificial intelligence, validation study

Posted Date: July 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10080817/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. The authors declare no financial competing interests. ES, LF, AS, IJ, AH, AW, and CG are involved in the development of NeutrinoReview and related LLM-assisted screening methods evaluated in this study. NeutrinoReview is an open-source research prototype.

Prospective validation of NeutrinoReview for
LLM-assisted systematic review screening: an
indoor air quality case study

Elias Sandner^{1,3*}, Luca Fontana², Michael T. Solomon⁴,
Sumeya B. Abdella⁴, Elisa Caracci⁴, Luca Stabile⁴,
Giorgio Buonanno^{4,5}, Alice Simniceanu², Igor Jakovljevic¹,
Andre Henriques¹, Andreas Wagner¹, Christian Gütl³

¹*CERN, Geneva, Switzerland.

²World Health Organization, Geneva, Switzerland.

³Graz University of Technology, Graz, Austria.

⁴Department of Civil and Mechanical Engineering, University of
Cassino and Southern Lazio, Cassino, FR, Italy.

⁵International Laboratory for Air Quality and Health, Queensland
University of Technology, Brisbane, QLD, Australia.

*Corresponding author(s). E-mail(s): elias.sandner@cern.ch;

Contributing authors: fontanal@who.int;

michaeltesfaye.solomon@unicas.it;

sumeyabehar.abdella@studentmail.unicas.it; elisa.caracci@unicas.it;

l.stabile@unicas.it; buonanno@unicas.it; simniceanua@who.int;

igor.jakovljevic@cern.ch; andre.henriques@cern.ch;

andreas.wagner@cern.ch; c.guetl@tugraz.at;

Abstract

Background: Systematic-review screening is commonly performed by at least two human reviewers to reduce missed-study risk, but this substantially increases workload. Large language model (LLM)-based pre-filtration may reduce screening burden, but prospective evidence from live review workflows is needed before

such tools can be considered for practical use. **Methods:** Six NeutrinoReview screening configurations were prospectively evaluated in parallel with an ongoing systematic review of measured indoor air pollutant concentrations in Italy. These configurations comprised three operating points of the 5-tier screening algorithm and three CAL-X likelihood-threshold settings, all implemented with the same self-hosted LLaMA 3.1-8B instruction model. NeutrinoReview outputs and the prospective evaluation protocol were locked before the manual consensus ground truth was available to the NeutrinoReview and analysis team. Performance was evaluated against the final manual consensus status, and observed human-screening decisions were used to construct single-human and double-human comparator workflows. **Results:** The performance-analysis dataset contained 1,344 records, including 112 finally included records. All six NeutrinoReview configurations retained all 112 finally included records, corresponding to 100% sensitivity against the manual consensus ground truth. Workload reduction differed by configuration, ranging from 37.1% with **CAL-99** to 74.0% with **5-tier_Full_Automation**. In the human-screening comparator workflows, **Single-human title/abstract** workflows missed five to seven finally included records, while **Double-human title/abstract** workflows missed zero to three. **Conclusions:** In this prospective use case, NeutrinoReview reduced title/abstract screening workload without excluding records that were finally included by the manual consensus process. The locked output design strengthens the credibility of the finding, but the result should not be interpreted as general validation across reviews, models, or settings. Further prospective evaluations and field tests are needed across additional reviews. Until such evidence is available, NeutrinoReview is most defensible as an adjunct to human screening using the conservative **CAL-99** configuration.

Keywords: systematic review, screening, large language models, artificial intelligence, validation study

1 Background

By synthesising findings across ideally all relevant literature, systematic reviews integrate evidence from individual studies. Traditional evidence hierarchies therefore commonly place systematic reviews at the top of the evidence pyramid while contemporary frameworks emphasise that the certainty of evidence provided by a systematic review depends on the quality, consistency, directness, and precision of the included studies. Systematic reviews may therefore be better understood not as inherently high-certainty evidence, but as a lens through which the underlying evidence is appraised and interpreted [1].

The reliability of a systematic review therefore depends on whether the review process successfully identifies the relevant body of literature. This typically involves retrieving a large set of potentially relevant candidate studies, which are then subjected to a thorough screening process comprising title and abstract screening followed by full-text screening. To reduce human error and selection bias, systematic-review guidance recommends independent study selection by at least two reviewers, with duplicate title-and-abstract screening considered desirable and duplicate full-text assessment essential [2]. This recommendation is supported by empirical evidence showing that single-reviewer screening can miss relevant studies and that duplicate screening reduces, but does not eliminate, missed-study risk [3–5].

This duplicate-screening requirement improves robustness but also increases the workload of systematic reviews. In an analysis of 195 completed PROSPERO-registered systematic reviews, Borah et al. found a mean duration of 67.3 weeks from registered project start to publication and a mean team size of five authors or team members [6]. Focusing specifically on study selection, Shemilt et al. estimated that conventional double screening in a case study involving 12,477 abstracts required 1,089 hours and cost GBP 75,139 [7].

Motivated by this workload, the broader project investigates how information technology can streamline systematic-review workflows, with a particular focus on large language model (LLM)-based literature screening. In previous work, two LLM-based screening approaches were introduced: a 5-tier prompting approach, which assigns candidate records to ordered relevance categories that can be mapped to binary screening decisions at different automation levels, and CAL-X, which augments binary LLM screening decisions with next-token likelihood scores to retain low-confidence exclusions for human review [8, 9]. Both approaches were integrated into NeutrinoReview, an open-source proof-of-concept review management tool for LLM-assisted screening [10, 11].

139 For screening automation in systematic reviews, the primary requirement is to
140
141 minimise the risk of excluding relevant studies. Thomas et al. report that, in consul-
142 tation with the Cochrane Information Retrieval Methods Group, the Cochrane RCT
143
144 classifier was required to recall at least 99% of randomized controlled trials before
145
146 adoption in Cochrane workflows [12]. Consistent with this safety-oriented framing, the
147
148 5-tier and CAL-X algorithms were developed as high-sensitivity pre-filtration systems
149 rather than as replacements for human screening. Workload reduction is therefore a
150
151 secondary objective and should not be achieved at the cost of increased missed-study
152
153 risk.

154 Although the sensitivity target is explicit, there is no universally established proce-
155
156 dure specifying how such sensitivity should be validated before real-world deployment.
157 Both the 5-tier and CAL-X algorithms were previously evaluated using retrospective
158
159 datasets from 17 real-world systematic reviews across various domains, comprising
160
161 more than 80,000 candidate studies [9]. This provides broader retrospective evidence
162
163 than is available for many screening-automation approaches [13], but it remains
164
165 insufficient to justify replacing human screeners in practical review settings.

166 Retrospective evaluations have inherent limitations because the ground truth is
167
168 already known when the evaluation is designed and conducted. This can reduce trust
169
170 in performance estimates, particularly when prompts, thresholds, eligibility represen-
171
172 tations, or analysis choices could have been adjusted after inspecting the reference
173
174 labels. Public benchmark datasets, such as the SYNERGY dataset for study-selection
175
176 research, are valuable for reproducible method development, but they may also raise
177
178 concerns about data contamination when evaluating large language models [14, 15].
179 Because LLM training corpora are usually not fully disclosed, prospective validation
180
181 on review data that were not publicly available before the model training cutoff pro-
182
183 vides a stronger safeguard against the possibility that performance reflects memorised
184
185 labels rather than genuine screening behaviour.

Recent studies have begun to evaluate LLMs within evidence-synthesis workflows, but prospective evidence for screening automation remains limited [13]. Spillias et al. evaluated ChatGPT in a human and AI collaboration for identifying literature in a scoping review and reported that repeated AI screening strategies could achieve very low omission rates. After collaborative reassessment, the best-performing AI screen had a false-negative rate of 0.5%. However, the approach relied on prompt refinement, repeated model queries, and human reconciliation rather than a locked-output validation of autonomous exclusions [16].

Park et al. prospectively evaluated a ChatGPT-assisted systematic-review workflow alongside expert reviewers and found strong performance for some downstream tasks, including full-text assessment and data extraction, but title-and-abstract screening sensitivity was only 69.8%, with 13 eligible records missed by the model [17].

These studies therefore support the feasibility of LLM-assisted review workflows, while also showing that current evidence remains insufficient for unsupervised replacement of human screeners.

This research gap motivates the present prospective validation study, which evaluates six NeutrinoReview screening configurations in parallel with a systematic review of measured indoor air pollutant concentrations in Italy [18]. The objective is to determine whether LLM-based screening using only title/abstract information can serve as a high-sensitivity pre-filtration step that reduces human screening workload without excluding records that are finally included by the manual consensus process. To maximise transparency and reduce opportunities for post hoc optimisation, the NeutrinoReview outputs were generated and locked before the manual consensus ground truth was available. The study addresses the following research questions:

1. What sensitivity and workload reduction are achieved by each NeutrinoReview configuration for title-and-abstract-based pre-screening in a live systematic-review workflow?

231 2. How does the observed performance of NeutrinoReview compare with single-human
232 and double-human screening workflows?
233

234 The remainder of the paper is structured as follows: [Methods](#) describes the
235 prospective validation design, source review data, NeutrinoReview configurations,
236 ground-truth definition, and analysis procedures. [Results](#) reports NeutrinoReview
237 sensitivity, workload reduction, illustrative cost estimates, and human-screening com-
238 parator workflows. [Discussion](#) interprets the safety-workload trade-off, compares the
239 findings with prior work, and discusses limitations and implementation consider-
240 ations. Finally, [Conclusions](#) summarises the practical implications for prospective
241 validation and cautious use of LLM-based pre-filtration and [Declarations](#) provides
242 ethics, data availability, competing interests, funding, author contributions, and
243 acknowledgements.
244
245
246
247
248
249
250

251 252 253 **2 Methods** 254

255 A prospective validation study of NeutrinoReview, a proof-of-concept large language
256 model (LLM)-assisted screening application for systematic reviews, was conducted in
257 parallel with a systematic review of measured indoor air pollutant concentrations in
258 Italy [18]. The objective was to evaluate whether NeutrinoReview could act as a high-
259 sensitivity pre-filtration step: records classified as included by NeutrinoReview would
260 remain available for human assessment, whereas records classified as excluded would
261 be removed from the downstream title/abstract screening workload in the evaluated
262 workflow.
263
264
265
266
267

268 Figure 1 summarizes the prospective validation workflow rather than a deployed
269 screening workflow. After the source-review searches had been completed and dedu-
270 plicated candidate records were available, NeutrinoReview screening was performed
271 as a separate, non-interventional workflow from the manual screening process. The
272 prospective evaluation protocol and raw LLM screening outputs were uploaded to
273
274
275
276

Zenodo under embargo before the manual screening results were available to the NeutrinoReview and analysis team [19]. Final analyses used the locked NeutrinoReview outputs only after the manual consensus ground truth had become available.

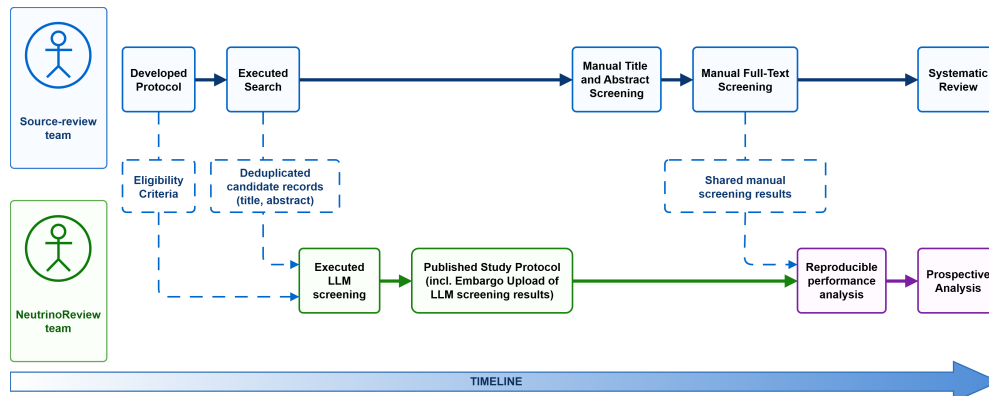


Fig. 1 Prospective validation workflow. Manual systematic-review screening and NeutrinoReview screening were conducted as separate workflows. NeutrinoReview received deduplicated candidate-record titles and abstracts together with the source-review eligibility criteria. The prospective evaluation protocol and raw LLM screening outputs were uploaded to Zenodo under embargo before the final manual screening results were available to the NeutrinoReview and analysis team. The final reproducible performance analysis compared the locked LLM outputs with the manual consensus ground truth and observed human screening decisions.

The primary analyses continued to evaluate each NeutrinoReview configuration individually against the manual consensus ground truth. A protocol-planned analysis of combined multi-algorithm screening scenarios was not pursued because all evaluated NeutrinoReview configurations retained all finally included records, leaving no missed-record pattern for such scenarios to address. Instead, reviewer-level Rayyan decisions from the three human screeners were used to construct single-human and double-human comparator workflows, which were benchmarked against the three-screener consensus ground truth and compared with the LLM screening scenarios.

2.1 Source review, candidate records, and ground truth

The source review was a systematic review of indoor air quality in Italy, covering measured concentrations of chemical and physical indoor air pollutants across residential,

public, occupational, and other indoor environments [18]. Searches were conducted in PubMed, Scopus, and Dimensions using combinations of controlled vocabulary and free-text terms for indoor air pollution, measured pollutant concentrations, and Italy or Italian regions. The source review protocol specified inclusion of English- or Italian-language studies from January 2000 onward that reported measured indoor pollutant concentrations in Italy, including peer-reviewed studies and methodologically transparent institutional reports [20]. Studies based only on modelled data, qualitative findings, outdoor-only measurements, biological contaminants, comfort parameters alone, or incomplete duplicate datasets were excluded.

After deduplication in Rayyan, the source review yielded 1,383 unique candidate records. This deduplicated candidate set, together with the eligibility criteria from the source review protocol, was provided for NeutrinoReview screening.

Manual screening for the source review was performed in Rayyan by three human reviewers. Records were first screened at title/abstract level. Records included after title/abstract consensus proceeded to full-text retrieval and full-text screening. Conflicts were resolved through discussion among all three reviewers. The manual screening workflow of the source review did not involve any LLM-based or other automated decision-making tool, and manual screeners were blinded to the NeutrinoReview results. Four Rayyan exports from this manual workflow were available for evaluation, comprising title/abstract individual decisions, title/abstract consensus decisions, full-text individual decisions, and full-text consensus decisions.

The ground truth was defined as the final manual consensus status. A record was labelled as included only if it was included after manual full-text consensus, following full-text assessment and resolution of any conflicts among the three human reviewers. Records excluded at title/abstract consensus or at full-text consensus were labelled as excluded. Records included at title/abstract screening but not retrieved for full-text assessment were reported separately and excluded from performance assessment.

The evaluation dataset reflects the original Rayyan screening exports rather than later data-extraction-stage changes in the source review. Four records included after Rayyan full-text screening were later excluded during data extraction, and one additional Rayyan record was marked as not retrieved compared with the PRISMA diagram of the source review. These discrepancies and the affected records are documented in Supplementary Note S1.

2.2 NeutrinoReview screening

NeutrinoReview is a proof-of-concept platform that integrates two LLM-based screening algorithms described in previous work [8, 9]. The prototype is not production-ready and is used solely for research purposes. In this study, the evaluated implementation was NeutrinoReview commit 007f43d1 [11]. Screening was performed through the NeutrinoReview web application on November 1, 2025, using a self-hosted LLaMA 3.1-8B instruction model [21].

The deduplicated candidate records were entered into six separate NeutrinoReview review instances using title and abstract information together with the source-review eligibility criteria. No full-text information was provided to the LLM during NeutrinoReview screening. The six review instances represented the NeutrinoReview configurations evaluated in this study, covering both LLM-based screening algorithms and all integrated screening settings (Table 1).

Table 1 NeutrinoReview configurations evaluated in this study. Details of the 5-tier and CAL-X algorithms are provided in previous work [8, 9].

Configuration label	Algorithm	Pre-filtration rule or setting
5-tier_MaxSensitivity	5-tier	Excluded only category 5 records
5-tier_Balanced	5-tier	Excluded categories 4 and 5
5-tier_Full_Automation	5-tier	Retained only categories 1 and 2
CAL-95	CAL-X	Threshold 63, calibrated target 95% sensitivity
CAL-98	CAL-X	Threshold 80, calibrated target 98% sensitivity
CAL-99	CAL-X	Threshold 89, calibrated target 99% sensitivity

415 With the 5-tier algorithm, each candidate record is assigned by the LLM to one
416 of five relevance categories, ranging from highly relevant to not relevant, and these
417 categories are then converted into a binary decision according to a selected automa-
418 tion level [8]. Three 5-tier settings were evaluated: `5-tier_MaxSensitivity`, which
419 excluded only category 5 records; `5-tier_Balanced`, which excluded categories 4 and
420 5; and `5-tier_FullAutomation`, which retained only categories 1 and 2 for human
421 review.
422

426 With the CAL-X algorithm, the LLM is prompted for a binary include/exclude
427 response, and the next-token likelihood for the first response token is retrieved as an
428 uncertainty signal [9]. Records initially classified as excluded are retained for human
429 review when the likelihood of the exclusion response is below the selected thresh-
430 old. Three CAL-X settings were evaluated: `CAL-95` with threshold 63, `CAL-98` with
431 threshold 80, and `CAL-99` with threshold 89; these thresholds correspond to CAL-X
432 calibration targets of 95%, 98%, and 99% sensitivity, respectively [9]. NeutrinoReview
433 provides built-in options for configurations calibrated to 95% and 99%; the threshold
434 for `CAL-98` was therefore entered as a custom setting.
435

441 The proof-of-concept implementation of NeutrinoReview did not yet include a
442 feature to export results immediately after LLM screening. Therefore, a JavaScript
443 function was executed via the browser console to automatically mark all records as
444 included in the manual-screening interface, enabling export of the LLM screening
445 outputs after automated screening. The exported NeutrinoReview outputs were locked
446 and uploaded to Zenodo under embargo before the final manual consensus results were
447 available to the NeutrinoReview and analysis team [19].
448
449
450
451
452
453
454
455
456
457
458
459
460

2.3 Evaluation methods

Evaluation used the six locked NeutrinoReview exports together with the Rayyan-derived manual screening decisions described above. For each NeutrinoReview configuration, the binary NeutrinoReview decision was interpreted as a pre-filtration decision. Records labelled included by NeutrinoReview were counted as records requiring human review. Records labelled excluded by NeutrinoReview were counted as workload saved.

The two primary performance metrics were sensitivity (Equation 1), which quantified the proportion of relevant records retained for human assessment, and workload reduction (Equation 2), which quantified the proportion of records filtered out before human screening.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

$$\text{Workload reduction} = \frac{\text{TN} + \text{FN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (2)$$

TP : true positives – relevant records retained for human review,

FN : false negatives – relevant records excluded by pre-filtration,

TN : true negatives – irrelevant records excluded by pre-filtration,

FP : false positives – irrelevant records retained for human review.

Additional metrics included specificity, positive predictive value, negative predictive value, accuracy, the number of records requiring human review, the number of records saved from human screening, and the number of finally included records missed by each configuration.

For comparison with NeutrinoReview, human-screening comparator workflows were constructed from the reviewer-level Rayyan decisions. **Single-human title/abstract** scenarios used each reviewer’s title/abstract decision alone. **Single-human title/abstract plus full-text** scenarios required the same

reviewer to include the record at both title/abstract and full-text stages. Double-human title/abstract scenarios retained a record if either reviewer in a pair included it at title/abstract screening. Double-human title/abstract plus full-text scenarios used symmetric pair logic: at least one reviewer in the pair had to include the record at title/abstract screening and at least one reviewer in the same pair had to include it at full-text screening. The same performance metrics were calculated for all NeutrinoReview configurations and human-screening comparator workflows.

Cost and workload simulations were restricted to title/abstract screening labour. The baseline scenario was double independent human title/abstract screening of all records in the performance-analysis dataset. NeutrinoReview was modelled as a pre-filtration step before title/abstract screening: a NeutrinoReview-using screener screened only records retained by the selected configuration, whereas any non-NeutrinoReview screener screened all records.

The screening time per record was estimated from the observed Rayyan title/abstract screening times of the three human reviewers. An illustrative labour cost of CHF 100 per hour was applied. The simulation excluded full-text screening, conflict resolution, coordination, NeutrinoReview runtime, software operating costs, and infrastructure costs. The focused cost scenarios compared fully manual double title/abstract screening with one or two screeners using either CAL-99 or 5-tier_Full_Automation as pre-filtration. These configurations were selected as the most conservative and least conservative pre-filtration options, respectively.

3 Results

The Rayyan exports contained 1,383 deduplicated candidate records. After title/abstract consensus screening, 201 records were advanced to full-text retrieval and full-text screening. Final manual consensus classified 112 records as included and 50 records as excluded after full-text assessment. A further 39 records were labelled as

not retrieved and were excluded from performance denominators. The performance-analysis dataset therefore contained 1,344 records, including 112 finally included records and 1,232 finally excluded records.

3.1 NeutrinoReview decision counts and performance

All six NeutrinoReview configurations retained all 112 finally included records, corresponding to 100% sensitivity against the manual consensus ground truth. No finally included record was classified as excluded by any NeutrinoReview configuration. Across the six configurations, the number of records retained for human review ranged from 350 to 846 of 1,344 records (Table 2). The least conservative configuration, labelled `5-tier_Full_Automation`, retained 350 records (26.0%) and excluded 994 records (74.0%). The most conservative setting, `CAL-99`, retained 846 records (62.9%) and excluded 498 records (37.1%). Among the `CAL-X` configurations, retained records increased from 451 (33.6%) with `CAL-95` to 639 (47.5%) with `CAL-98` and 846 (62.9%) with `CAL-99`. Workload reduction differed substantially between configurations, ranging from 37.1% for `CAL-99` to 74.0% for `5-tier_Full_Automation`.

Table 2 NeutrinoReview performance against the manual consensus ground truth. The Excluded column reports records filtered before human screening and therefore corresponds to workload reduction.

Configuration	Retained	Excluded	Sensitivity	Specificity
<code>5-tier_Full_Automation</code>	350 (26.0%)	994 (74.0%)	100.0%	80.7%
<code>5-tier_Balanced</code>	613 (45.6%)	731 (54.4%)	100.0%	59.3%
<code>5-tier_MaxSensitivity</code>	645 (48.0%)	699 (52.0%)	100.0%	56.7%
<code>CAL-95</code>	451 (33.6%)	893 (66.4%)	100.0%	72.5%
<code>CAL-98</code>	639 (47.5%)	705 (52.5%)	100.0%	57.2%
<code>CAL-99</code>	846 (62.9%)	498 (37.1%)	100.0%	40.4%

3.2 Human-screening comparator workflows

Single-human comparator workflows missed finally included records in all scenarios (Figure 2). Single-human title/abstract workflows achieved sensitivities between 93.8% and 95.5%, missing five to seven finally included records. Single-human title/abstract plus full-text workflows achieved sensitivities between 85.7% and 94.6%, with six to 16 finally included records missed.

Double-human workflows reduced the number of missed finally included records. The three possible Double-human title/abstract reviewer pairs achieved sensitivities between 97.3% and 100.0%, missing zero to three finally included records. The Double-human title/abstract plus full-text workflows did not change the missed-inclusion counts, indicating that no additional finally included records were lost at the full-text-decision stage after retention by at least one reviewer in a pair at title/abstract screening.

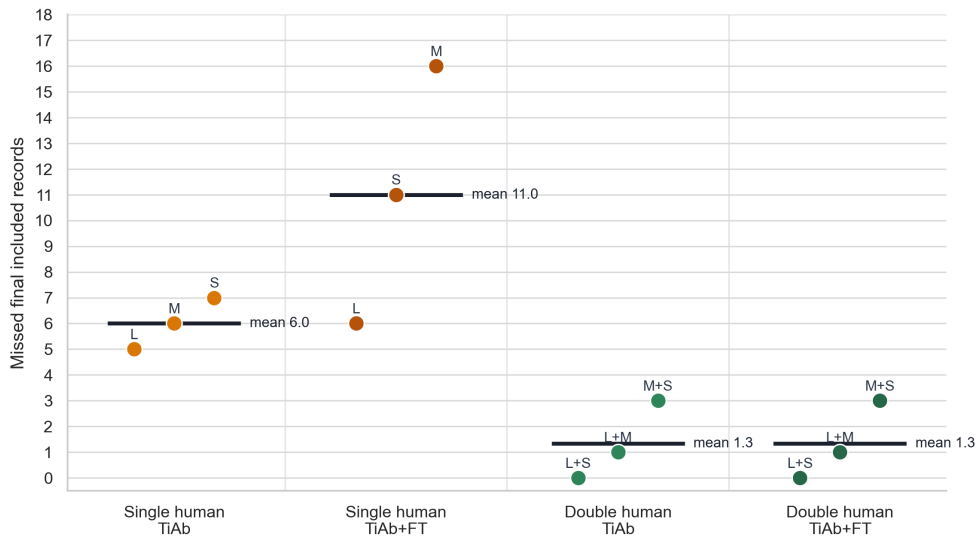


Fig. 2 Missed finally included records relative to the manual consensus ground truth in human-screening comparator workflows. Points represent individual reviewers for single-human workflows and reviewer pairs for double-human workflows; horizontal markers show the mean within each workflow type. L, M, and S denote the three human screeners.

3.3 Workload and cost simulations

The fully manual baseline required two independent title/abstract screeners to screen all 1,344 records, corresponding to 2,688 human screening events. At an observed average screening time of 41.3 seconds per record, this represented 30.8 hours of title/abstract screening labour and an illustrative labour cost of CHF 3,083.

The focused cost simulations compared the fully manual baseline with `CAL-99` and `5-tier_Full_Automation` as pre-filtration strategies (Table 3). If one of two title/abstract screeners used `CAL-99` pre-filtration, 498 screening events were avoided, corresponding to 5.7 hours and CHF 571 saved. If both screeners used `CAL-99`, 996 screening events were avoided, corresponding to 11.4 hours and CHF 1,142 saved.

Using `5-tier_Full_Automation` as pre-filtration yielded larger workload reductions. If one screener used this configuration, 994 screening events were avoided, corresponding to 11.4 hours and CHF 1,140 saved. If both screeners used this configuration, 1,988 screening events were avoided, corresponding to 22.8 hours and CHF 2,280 saved.

Table 3 Title/abstract screening workload and illustrative labour-cost simulations.

Scenario	Events	Time (hours)	Cost (CHF)	Savings (CHF)
Fully manual double screening	2,688	30.8	3,083	0
One screener with <code>CAL-99</code>	2,190	25.1	2,512	571
Both screeners with <code>CAL-99</code>	1,692	19.4	1,941	1,142
One screener with <code>5-tier_Full_Automation</code>	1,694	19.4	1,943	1,140
Both screeners with <code>5-tier_Full_Automation</code>	700	8.0	803	2,280

4 Discussion

This study evaluated NeutrinoReview as a high-sensitivity pre-filtration step for title/abstract screening in a live systematic-review workflow. The findings directly

691 answer the two research questions. RQ1 asked what sensitivity and workload reduc-
692 tion were achieved by each NeutrinoReview configuration for title-and-abstract-based
693 pre-screening. In this dataset, all six configurations achieved 100.0% sensitivity against
694 the manual consensus ground truth, because none of the 112 finally included records
695 was excluded by NeutrinoReview. Workload reduction differed by configuration:
696 37.1% for CAL-99, 52.0% for 5-tier_MaxSensitivity, 52.5% for CAL-98, 54.4% for
697 5-tier_Balanced, 66.4% for CAL-95, and 74.0% for 5-tier_FullAutomation. Thus,
698 all evaluated configurations met the high-sensitivity objective in this source review,
699 while the achieved workload reduction depended strongly on the selected operating
700 point.

701 RQ2 asked how the observed performance of NeutrinoReview compared with
702 single-human and double-human screening workflows. In this evaluation, Neutri-
703 noReview showed higher sensitivity than the observed single-human title/abstract
704 workflows, which missed five to seven finally included records, and at least compara-
705 ble sensitivity to the observed double-human title/abstract workflows, which missed
706 zero to three finally included records. This comparison should be interpreted as a
707 workflow-specific benchmark rather than as evidence that NeutrinoReview is generally
708 superior to human screening, because all estimates come from one review, one model
709 implementation, and one set of eligibility criteria.

710 The prospective locked-output design is central to the credibility of this finding.
711 NeutrinoReview outputs were generated, locked, and uploaded before the manual con-
712 sensus ground truth was available to the NeutrinoReview and analysis team [19]. This
713 reduces concerns that prompts, thresholds, or analysis choices were adjusted after
714 inspecting the final labels and helps address data-contamination concerns because the
715 final screening labels came from a live review rather than from a public benchmark
716 that may have appeared in model training data.

The results support interpreting NeutrinoReview as a configurable pre-filtration approach, but not all operating points should be treated equally. For cautious practical use, CAL-99 remains the most defensible setting because it combines the present prospective finding with stronger prior retrospective evidence for high sensitivity [9]. Less conservative settings also missed no finally included records in this review, but prior retrospective evaluations indicate that they carry a higher missed-study risk [8, 9]. Their larger workload reductions are therefore relevant mainly where a higher risk of missed records is acceptable.

The human-screening comparator workflows provide a realistic reference for interpreting these results. The missed-record pattern observed here is consistent with prior empirical evidence showing that single-reviewer screening can miss relevant studies and that duplicate screening reduces, but does not eliminate, this risk [3–5]. The findings also extend the emerging prospective literature on LLM-assisted evidence synthesis. Spiliadis et al. reported very low omission rates after repeated ChatGPT screening, prompt refinement, and collaborative reassessment, whereas the present study used locked NeutrinoReview outputs generated before the final manual labels were available [16]. Park et al. reported substantially lower title/abstract screening sensitivity in a prospective ChatGPT-assisted workflow, with 13 eligible records missed by the model [17]. These comparisons should not be read as evidence that NeutrinoReview is generally superior, because the studies differed in topic, model, prompts, workflow design, and reference standard. They instead underline the need for prospective, workflow-specific validation.

The main boundary of this study is that zero missed finally included records in one source review does not establish general safety across review topics, eligibility criteria, search strategies, models, prompts, or deployments. The evaluation used one review topic, one LLM implementation, and title/abstract inputs only. The manual consensus ground truth is the appropriate practical reference standard for this evaluation, but it

783 remains human-derived and does not prove the true relevance status of every record.
784 The workload and cost simulations were illustrative and excluded full-text screening,
785 conflict resolution, coordination, NeutrinoReview runtime, software operating costs,
786 and infrastructure costs.
787

789 In this specific review, NeutrinoReview could have reduced screening workload
790 without changing the set of finally included records. However, this can only be
791 known after prospective validation against the manual consensus process. Real-world
792 use should therefore begin with conservative, supervised, or parallel pre-filtration
793 workflows, especially CAL-99, rather than unsupervised replacement of established
794 screening procedures. Future work should repeat locked prospective validations across
795 multiple reviews, domains, and review teams, and field-test designs should examine
796 both records that might be missed by LLM pre-filtration and records that might be
797 recovered because LLM assistance disagrees with erroneous human exclusions.
798
799

806 5 Conclusions

807 This prospective evaluation found that all evaluated NeutrinoReview configurations
808 retained all finally included records while reducing title/abstract screening workload by
809 37.1% to 74.0%, depending on the selected setting. The human-screening comparator
810 workflows further showed that NeutrinoReview should be interpreted relative to real
811 screening workflows, which can also miss relevant records. The locked-output design
812 strengthens the credibility of this finding, but the result should be interpreted as
813 evidence from one live systematic-review workflow rather than as general validation
814 across reviews or settings.
815

816 For high-sensitivity systematic-review pre-filtration, conservative settings such as
817 CAL-99 remain the most defensible option. NeutrinoReview is therefore most promis-
818 ing as a supervised workflow component, for example as an additional screener,
819 while further prospective evaluations and field tests are needed before moving toward
820
821
822
823
824
825
826
827
828

replacement of human screeners. Such field tests should examine both records missed
by NeutrinoReview and records retained by NeutrinoReview that human screeners
would otherwise have missed.

List of abbreviations

<i>LLM</i>	Large language model	829
<i>CAL-X</i>	Next-token likelihood-calibrated LLM screening algorithm	830
<i>CAL-95/CAL-98/CAL-99</i>	<i>CAL-X</i> configurations targeting 95%, 98%, and 99% sensi- tivity, respectively	831
<i>PROSPERO</i>	International Prospective Register of Systematic Reviews	832
<i>GBP</i>	British pounds sterling	833
<i>RCT</i>	Randomized controlled trial	834
<i>AI</i>	Artificial intelligence	835
<i>TP</i>	True positive	836
<i>FN</i>	False negative	837
<i>TN</i>	True negative	838
<i>FP</i>	False positive	839
<i>CHF</i>	Swiss francs	840

Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. The prospective evaluation protocol and
embargoed locked-output record are available on Zenodo [19]. The supplementary
material for this manuscript, including the analysis notebook, Rayyan-derived inputs,
processed datasets, supporting script, and Supplementary Note S1 documenting recon-
ciliation of the Rayyan exports with the source-review PRISMA diagram, is available

875 on Zenodo [22]. The source systematic-review protocol is available as an OSF registra-
876 tion [20]. The NeutrinoReview codebase is available from the CERN GitLab repository
877 [11]. The hosted NeutrinoReview application is access-restricted and available on
878 request subject to CERN access constraints.
879

881 **Competing interests.** The authors declare no financial competing interests. ES, LF,
882 AS, IJ, AH, AW, and CG are involved in the development of NeutrinoReview and
883 related LLM-assisted screening methods evaluated in this study. NeutrinoReview is
884 an open-source research prototype.
885

888 **Funding.** This work was conducted in the context of ES’s PhD project, which is
889 funded by the joint CERN-WHO ARIA project. The funding arrangement did not
890 influence the study design, data analysis, interpretation of results, or decision to submit
891 the manuscript beyond the scientific contributions of the authors affiliated with CERN
892 and WHO.
893

896 **Authors’ contributions.** ES conceived and designed the study, executed the
897 prospective NeutrinoReview evaluation, curated and analysed the data, interpreted the
898 results, and drafted the manuscript. LF, MTS, and SBA conducted the manual screen-
899 ing in the source systematic review and provided the source-review screening data. LF
900 and AS provided methodological guidance on the prospective evaluation design. EC,
901 LS, and GB provided source-review and indoor-air-quality domain expertise, interpre-
902 tation input, and manuscript review. LF, AS, AH, IJ, AW, and CG provided scientific,
903 technical, and supervisory input during the project. All authors critically reviewed the
904 manuscript and approved the final version.
905

911 **Acknowledgements.** The authors thank all contributors to the open-source Neu-
912 trinoReview platform from CERN, WHO, and the CoDiS Lab at Graz University
913 of Technology. The authors also thank Kavita Kothari and participants of the 2025
914 International Collaboration for the Automation of Systematic Reviews meeting for
915

916
917
918
919
920

discussions on validation approaches for LLM-assisted systematic-review screening, which provided valuable insights for the present study.

Use of artificial intelligence tools. OpenAI Codex and ChatGPT were used as assistive tools during preparation of the Python analysis code and manuscript, including coding support, debugging, LaTeX editing, language refinement, and drafting support for author-written text. This use was limited to editorial and technical assistance and did not replace author-led scientific, analytical, or editorial decisions. All AI-assisted outputs were reviewed, revised, and verified by the authors, who take full responsibility for the final manuscript, code, data presentation, and conclusions. The use of LLMs within the evaluated NeutrinoReview screening workflow is described in the Methods section.

References

- [1] Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evidence-Based Medicine. 2016;21(4):125–127. <https://doi.org/10.1136/ebmed-2016-110401>.
- [2] Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Metzendorf MI, et al. Chapter 4: Searching for and selecting studies [web page]. Cochrane; 2025 [updated March 2025; cited 19 May 2026]. Available from: <https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-04>. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for Systematic Reviews of Interventions version 6.5.1.
- [3] Waffenschmidt S, Knelangen M, Sieben W, Böhn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC Medical Research Methodology. 2019;19(1):132. <https://doi.org/10.1186/s12874-019-0782-0>.

967 [4] Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N,
968 et al. Single-reviewer abstract screening missed 13 percent of relevant studies:
969 a crowd-based, randomized controlled trial. *Journal of Clinical Epidemiology*.
970 2020;121:20–28. <https://doi.org/10.1016/j.jclinepi.2020.01.005>.
971
972
973
974 [5] Stoll CRT, Izadi S, Fowler S, Green P, Suls J, Colditz GA. The value of a second
975 reviewer for study selection in systematic reviews. *Research Synthesis Methods*.
976 2019;10(4):539–545. <https://doi.org/10.1002/jrsm.1369>.
977
978
979
980 [6] Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers
981 needed to conduct systematic reviews of medical interventions using data from
982 the PROSPERO registry. *BMJ Open*. 2017;7(2):e012545. <https://doi.org/10.1136/bmjopen-2016-012545>.
983
984
985
986
987 [7] Shemilt I, Khan N, Park S, Thomas J. Use of cost-effectiveness analysis to
988 compare the efficiency of study identification methods in systematic reviews.
989 *Systematic Reviews*. 2016;5(1):140. <https://doi.org/10.1186/s13643-016-0315-4>.
990
991
992
993 [8] Sandner E, Hu B, Simniceanu A, Fontana L, Jakovljevic I, Henriques A, et al.
994 Screening Automation for Systematic Reviews: A 5-Tier Prompting Approach
995 Meeting Cochrane’s Sensitivity Requirement. In: 2024 2nd International Con-
996 ference on Foundation and Large Language Models (FLLM). IEEE; 2024. p.
997 150–159. Available from: <https://doi.org/10.1109/FLLM63129.2024.10852425>.
1000
1001
1002 [9] Sandner E, Negovetic M, Kothari K, Taj I, Fontana L, Henriques A, et al. Cal-X:
1003 Enhancing Systematic Review Screening with LLMs and Next-Token Likelihood
1004 Calibration. In: 2025 3rd International Conference on Foundation and Large
1005 Language Models (FLLM). IEEE; 2025. p. 1071–1078. Available from: <https://doi.org/10.1109/FLLM67465.2025.11391146>.
1006
1007
1008
1009
1010
1011
1012

- [10] Sandner E, Jakovljevic I, Simniceanu A, Fontana L, Henriques A, Wagner A, et al. NeutrinoReview: Concept Proposal for an Open Source Review Management Tool. In: 6th International Open Search Symposium ossym2024; 2024. p. 53.
- [11] CERN CAiMIRA WP4. NeutrinoReview: Prototype of a Semi-Automated Systematic Review Management Tool [GitLab repository]. CERN GitLab; 2025 [cited 3 June 2026]. Available from: <https://gitlab.cern.ch/caimira/caimira-wp4/neutrinoreview>. Commit 007f43d13444ab9b8667023152a3f7d86b826eb4.
- [12] Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *Journal of Clinical Epidemiology*. 2021;133:140–151. <https://doi.org/10.1016/j.jclinepi.2020.11.003>.
- [13] Sandner E, Fontana L, Kothari K, Henriques A, Jakovljevic I, Simniceanu A, et al. Evaluating large language models for literature screening: A systematic review of sensitivity and workload reduction. In: Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025). SCITEPRESS - Science and Technology Publications; 2025. p. 508–517. Available from: <https://doi.org/10.5220/0013562900003967>.
- [14] De Bruin J, Ma Y, Ferdinands G, Teijema J, Van de Schoot R. SYNERGY - Open machine learning dataset on study selection in systematic reviews. *DataverseNL*. 2023;Version 1. <https://doi.org/10.34894/HE6NAQ>.
- [15] Magar I, Schwartz R. Data Contamination: From Memorization to Exploitation. In: Proceedings of the 60th Annual Meeting of the Association for Computational

1059 Linguistics (Volume 2: Short Papers). Dublin, Ireland: Association for Compu-
1060 tational Linguistics; 2022. p. 157–165. Available from: [https://aclanthology.org/](https://aclanthology.org/2022.acl-short.18/)
1061 [2022.acl-short.18/](https://aclanthology.org/2022.acl-short.18/).
1062
1063
1064
1065 [16] Spillias S, Tuohy P, Andreotta M, Annand-Jones R, Boschetti F, Cvitanovic C,
1066 et al. Human-AI collaboration to identify literature for evidence synthesis. *Cell*
1067 *Reports Sustainability*. 2024;<https://doi.org/10.1016/j.crsus.2024.100132>.
1068
1069
1070
1071 [17] Park Y, Zhang HS, Bai SW. Large language models in systematic review and
1072 meta-analysis of surgical treatments for vaginal vault prolapse. *npj Digital*
1073 *Medicine*. 2026;9:262. <https://doi.org/10.1038/s41746-026-02431-w>.
1074
1075
1076 [18] Fontana L, Simniceanu A, Di Marco M, Silenzi A, Solomon MT, Abdella SB, et al.
1077 Indoor air quality and health-relevant exposure in Italy: a systematic review of
1078 measured pollutant concentrations, guideline exceedances, and radon-attributable
1079 risk. *Building and Environment*. 2026;Unpublished manuscript, submitted for
1080 publication on 11 June 2026.
1081
1082
1083
1084
1085 [19] Sandner E, Fontana L, Simniceanu A, Jakovljevic I, Henriques A, Wagner A,
1086 et al. Protocol for Prospective Evaluation of Screening Algorithms Integrated
1087 in NeutrinoReview. *Zenodo*. 2026;Version v2. [https://doi.org/10.5281/zenodo.](https://doi.org/10.5281/zenodo.17554629)
1088 [17554629](https://doi.org/10.5281/zenodo.17554629).
1089
1090
1091
1092
1093 [20] Fontana L, Buonanno G. Indoor air quality in Italy: a systematic review of
1094 measured indoor pollutant concentrations across residential, public, and occupa-
1095 tional environments. *OSF*. 2025 Oct;Registered 24 October 2025; Generalized
1096 Systematic Review Registration. <https://doi.org/10.17605/OSF.IO/92A3G>.
1097
1098
1099
1100 [21] Meta. Llama-3.1-8B-Instruct [model card]. Hugging Face; 2024 [cited 15
1101 May 2026]. Available from: <https://huggingface.co/meta-llama/Llama-3>.
1102
1103
1104

1-8B-Instruct. Model release date: July 23, 2024.	1105
	1106
[22] Sandner E, Fontana L, Simniceanu A, Jakovljevic I, Henriques A, Wagner A, et al.	1107
Prospective validation of NeutrinoReview for LLM-assisted systematic review	1108
screening: an indoor air quality case study (supplementary material). Zenodo.	1109
2026 Jun;Version v3; supplementary material. https://doi.org/10.5281/zenodo.	1110
20529631 .	1111
	1112
	1113
	1114
	1115
	1116
	1117
	1118
	1119
	1120
	1121
	1122
	1123
	1124
	1125
	1126
	1127
	1128
	1129
	1130
	1131
	1132
	1133
	1134
	1135
	1136
	1137
	1138
	1139
	1140
	1141
	1142
	1143
	1144
	1145
	1146
	1147
	1148
	1149
	1150

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarymaterial.zip](#)