

Protocol for Prospective Evaluation of Screening Algorithms integrated in NeutrinoReview

Abstract: The screening phase of systematic reviews is labor-intensive and often represents a major bottleneck in evidence synthesis. To support automation, two large language model (LLM)-based screening algorithms—the 5-Tier and CAL-X approaches—were developed and integrated into the NeutrinoReview web application. Both algorithms leverage the reasoning capabilities of LLMs to function as pre-filtration systems, prioritizing high sensitivity to minimize the risk of excluding relevant studies. While previous evaluations demonstrated promising performance on benchmark datasets, they were retrospective and subject to limitations such as potential data contamination and lack of real-world representativeness. This protocol describes the first **prospective evaluation** of NeutrinoReview’s automated screening, conducted within the ARIA project. The study uses records retrieved for a systematic review on indoor air quality in Italy (1,383 unique citations). Six screening runs were executed across different algorithmic configurations of 5-Tier and CAL-X. The LLM used was LLaMA 3.1-8B hosted at CERN. Preliminary results show workload reductions between 36% and 72%, with sensitivity assessment pending human consensus. Final analyses will compare algorithmic and human screening decisions, evaluate sensitivity and error patterns, and explore combined algorithm configurations. All outputs are time-stamped and embargoed in this Zenodo repository to ensure transparency and trust in this first real-world validation.

Background and Motivation

The screening phase of systematic reviews is both labor-intensive and time-consuming. To facilitate automation, two screening algorithms were developed and are described in detail in previous work [1, 2]. Both algorithms leverage the reasoning capabilities of large language models (LLMs) and are designed to serve as a pre-filtration system. They have been implemented in the proof-of-concept web application [NeutrinoReview](#) [3], which currently operates using a LLaMA-3.1-8B model hosted at CERN.

The 5-tier algorithm was initially evaluated retrospectively using data from four systematic reviews and GPT-4 accessed via the OpenAI API [1]. Subsequently, the CAL-X algorithm was introduced and assessed on 17 datasets using a self-hosted instance of the open-weight model *LLaMA 3.1-8B* [2]. In that work [2], the 5-tier algorithm was also re-evaluated and served as a baseline for comparison.

However, to date, all evaluations have been conducted in retrospective settings using publicly available benchmark datasets. Such evaluations inherently involve several limitations and concerns.

First, there is a risk of data contamination: the datasets used for evaluation originate from systematic reviews published between 2002 and 2023, and were themselves released in April 2023 (*Synergy* Dataset [4]) and May 2023 (Clinical Reviews provided by Guo et. al. [5]). The knowledge cutoffs of the models employed are September 2021 for GPT-4 and December 2023 for *LLaMA 3.1-8B* [6]. Given that large language models are trained on extensive, primarily web-scraped corpora whose exact composition is undisclosed, it remains uncertain whether the evaluation datasets overlap with the models’ training data. Consequently, it cannot be

guaranteed that the models' screening decisions result from genuine reasoning rather than partial memorization of previously encountered content.

Second, retrospective evaluations raise concerns regarding transparency and trust, as datasets could, in principle, be selectively chosen to produce favorable outcomes. In contrast, a prospective evaluation in which the raw results are published under embargo with a verifiable timestamp prior to knowing the ground truth maximizes transparency and is expected to increase trust in the evaluated screening algorithms.

Finally, retrospective evaluations deviate from real-world conditions, as they rely on the eligibility criteria reported in the published systematic reviews, which may differ in phrasing from those originally provided to the human screeners. Such discrepancies could influence the observed performance and limit how well the experimental setting represents real-world screening conditions.

The ARIA project, within which this study is conducted, was originally initiated by the World Health Organization (WHO) and CERN to develop a new model to quantify the risk of SARS-CoV-2 airborne transmission in enclosed spaces, thereby supporting space management decisions during the COVID-19 pandemic. The model relied on several parameters extracted through multiple systematic reviews [7], which required significant resources and time.

Building on this experience, it was decided to develop similar models to quantify the risk associated with other pathogens as well as with indoor air pollutants. A major gap in the field of indoor air quality is the lack of a comprehensive understanding of the global situation. To address this, a strategy was adopted to describe the global situation country by country through systematic reviews.

As the first step in this effort, a systematic review was initiated to describe the current state of indoor air quality (IAQ) in Italy based on measured pollutant concentrations. The registered protocol for this review is available in [8].

To further validate the effectiveness of NeutrinoReview and its integrated LLM-based pre-filtration system, this systematic review will serve as the basis for the present prospective validation study.

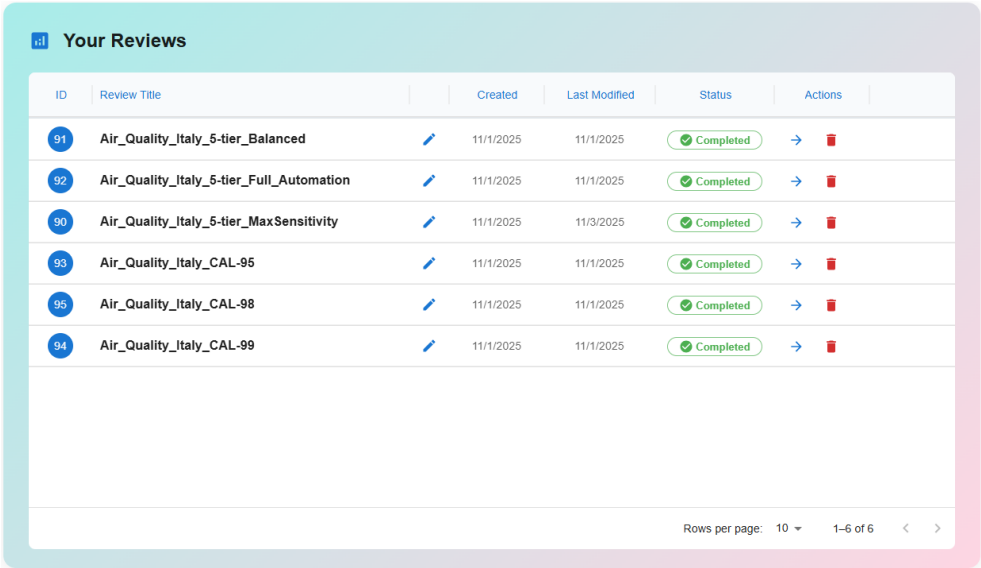
Prospective Evaluation Design

Data

To retrieve the set of candidate records for the systematic review on indoor air quality in Italy, electronic searches were conducted in **PubMed**, **Scopus**, and **Dimensions**. The corresponding search strategies are detailed in the review protocol [8]. After deduplication conducted using the free version of **Rayyan**, a total of **1,383 unique records** were obtained. This list of candidate studies was exported and provided to the team responsible for the prospective evaluation **prior to the start of human screening**.

Review Instances in NeutrinoReview

This list of candidate studies, together with the eligibility criteria outlined in the annexes of the review protocol [8], was subsequently used to create six review instances in NeutrinoReview, as illustrated in the screenshots shown in Figure 1 and Figure 2. All review instances were created and completed as prospective screening executions in NeutrinoReview on November 1, 2025, as detailed in the following section.



The screenshot displays the 'Your Reviews' section of the NeutrinoReview application. It features a table with six rows, each representing a completed review instance. The table columns are ID, Review Title, Created, Last Modified, Status, and Actions. Each row shows a unique ID, a specific review title related to 'Air_Quality_Italy', a creation date of 11/1/2025, and a status of 'Completed' with a green checkmark icon. The 'Actions' column for each row contains a blue arrow icon and a red trash can icon. At the bottom right of the table, there is a pagination control showing 'Rows per page: 10' and '1-6 of 6'.

ID	Review Title	Created	Last Modified	Status	Actions
91	Air_Quality_Italy_5-tier_Balanced	11/1/2025	11/1/2025	Completed	→ 🗑
92	Air_Quality_Italy_5-tier_Full_Automation	11/1/2025	11/1/2025	Completed	→ 🗑
90	Air_Quality_Italy_5-tier_MaxSensitivity	11/1/2025	11/3/2025	Completed	→ 🗑
93	Air_Quality_Italy_CAL-95	11/1/2025	11/1/2025	Completed	→ 🗑
95	Air_Quality_Italy_CAL-98	11/1/2025	11/1/2025	Completed	→ 🗑
94	Air_Quality_Italy_CAL-99	11/1/2025	11/1/2025	Completed	→ 🗑

Rows per page: 10 1-6 of 6 < >

Figure 1: Reviews created in NeutrinoReview

Eligibility Criteria

+ Add Criterion

Label	Included	Excluded	Actions
Population / Setting	Studies conducted in Italy, or multicountry studies providing disaggregated Italian data, including all Italian regions and provinces. Any indoor environments where people spend time: homes, schools, offices, hospitals, public buildings, and general indoor workplaces.	Studies conducted outside Italy or without Italian-specific results. Studies limited to experimental chambers not representing real-life indoor exposure.	 
Exposure (Intervention)	Measurement of indoor air pollutants through field monitoring or observational campaigns, including both short- and long-term measurements.	Studies relying solely on modelled, simulated, or estimated pollutant concentrations with no field data.	 
Comparator	Indoor-outdoor comparisons or between-building comparisons are eligible, provided indoor data are separable and clearly reported.	Studies reporting only outdoor data or mixed results where indoor values cannot be separated.	 
Outcomes	Quantitative pollutant concentration data (mean, median, range, SD, or IQR) expressed in recognized units ($\mu\text{g}/\text{m}^3$, ppm, Bq/m ³ , particles/cm ³ , etc.). Pollutants include PM ₁₀ , PM _{2.5} , ultrafine particles (UFPs), TVOCs, formaldehyde, CO ₂ , CO, NO ₂ , O ₃ , SO ₂ , radon, PAHs.	Studies presenting qualitative findings only (e.g., "high" or "low" exposure) or lacking essential information on methods, units, or duration. Studies focusing solely on biological contaminants (e.g., mold, bacteria, viruses, allergens) or comfort parameters (temperature, humidity) without pollutant data.	 
Study design	Observational or monitoring-based field studies (cross-sectional, longitudinal, or survey-based). Includes grey literature (ARPA, ISPRA, ISS, Ministry of Health, ENEA, or regional reports) with transparent methodology.	Editorials, commentaries, letters, policy briefs, or non-reviewed documents lacking methodological transparency. Excludes pure laboratory or experimental chamber studies.	 
Language and publication period	Publications in English or Italian, from January 2000 onward.	Publications in other languages, or before 2000 unless providing historical baseline data for comparison.	 
Data quality and duplication	Studies with complete reporting of methodology and results; when multiple reports overlap, include the most complete and recent version.	Studies with incomplete or duplicate datasets (less complete versions excluded).	 

Figure 2: Eligibility Criteria in NeutrinoReview

LLM based Prefiltration

While NeutrinoReview [3] offers a feature to directly retrieve studies from selected academic databases, the upload functionality was used to add the provided candidate studies to the six review instances. As the uploaded dataset already consisted of unique records, NeutrinoReview did not detect any duplicates.

Subsequently, LLM-based screening was performed in NeutrinoReview, using a unique configuration for each of the six review instances. NeutrinoReview provides two screening algorithms (5-Tier and CAL-X) each available with three configurable settings. The implementation of these algorithms within NeutrinoReview is described below, while technical details on the underlying algorithms are provided in the primary publications introducing them [1, 2].

In the 5-Tier algorithm, the LLM classifies candidate studies into categories 1 to 5, with decreasing relevance from 1 (highly relevant) to 5 (irrelevant). The **Max Sensitivity** setting automatically excludes only studies assigned to category 5, while all others require subsequent human screening. When selecting the **Balanced** automation level, studies categorized as 4 or 5 are automatically excluded, reducing the number of records requiring human review but increasing the risk of excluding potentially relevant studies. The **Full-Automation** setting—intended solely for evidence synthesis in less critical contexts and not for systematic reviews—further increases automation by excluding all studies except those classified as category 1 or 2 from human screening.

In contrast to the 5-Tier algorithm, the CAL-X algorithm instructs the LLM to return a binary response representing an include or exclude decision. In addition to the textual response, the LLM provides the next-token likelihood for the first token of the response, which is interpreted as the model’s confidence in that decision. This score is used to calibrate the algorithm towards a specific sensitivity. Two predefined settings, as well as an option to define custom thresholds, are available in NeutrinoReview. The **Low Workload** option is calibrated to achieve an average sensitivity of approximately 95%. While this setting may be suitable for rapid reviews, it is not appropriate for formal systematic reviews. The **High Sensitivity** option is tuned to reach a sensitivity of around 99% and is designed to satisfy Cochrane’s requirements for replacing human screeners in high-quality systematic reviews. Additionally, a third experiment using the CAL-X algorithm was conducted by manually setting the threshold to 80, corresponding to an expected sensitivity of 98%.

Bypassing Manual Screening and Exporting Results

Since the current proof-of-concept implementation of NeutrinoReview does not yet include a feature to export data immediately after LLM screening, manual screening execution for each record was required. To avoid the need for manually including several hundred papers across all six test cases, a JavaScript function was developed to automate the button-click process and executed via the browser console. The corresponding code is provided in `.auto_include_bot.js`.

Once the manual screening was completed, the results from all NeutrinoReview instances were exported. These files are uploaded to this Zenodo repository under embargo and will be released after the human screeners have submitted their final decisions, ensuring full transparency and eliminating concerns regarding potential bias in either the LLM or human screening resulting from prior knowledge of the other’s decisions.

All experiments were executed using NeutrinoReview at commit 007f43d1 [3].

Preliminary Results

Table 2 provides an overview of the executed experiments and the results currently available. In NeutrinoReview, automated screening functions as a pre-filtration system, automatically excluding a subset of records, while all records classified as “include” remain subject to human screening. Consequently, **workload reduction** is defined as the proportion of records automatically excluded by the system. **Sensitivity** represents the most critical metric for any screening automation, as it measures the proportion of relevant studies correctly identified as such. Its calculation requires a ground truth, which in this prospective study corresponds to the consensus screening decisions of the human reviewers. Accordingly, sensitivity values will be reported in the final publication. Execution time was measured manually using a stopwatch, starting when the “Start Automated Screening” button was pressed and stopping once the interface displayed the final results.

ID	Algorihtm	No Excluded	No remaining for Human Screening	Workload Reduction	Sensitivity	Execution Time
90	5-Tier (MaxSensitivity)	699	684	51%	-	2 min 33 s
91	5-Tier (Balanced)	731	652	53%	-	2 min 29 s
92	5-Tier (Full Automation)	994	389	72%	-	2 min 31 s

93	CAL-95 (t=63)	893	490	65%	-	2 min 26 s
94	CAL-99 (t=89)	498	885	36%	-	2 min 29 s
95	CAL-98 (custom threshold t=80)	705	678	51%	-	2 min 29 s

Table 2: Executed experiments and already available results

Planned Analysis

Once the human screening results become available, the analysis will proceed by comparing the LLM-generated screening decisions with the gold standard. This comparison will allow for quantifying the number of studies wrongly excluded by each algorithm and for calculating the corresponding sensitivity values.

Additionally, a multi-algorithm simulation will be conducted by combining the 5-Tier algorithm (using the **Max Sensitivity** setting) with the **CAL-X High Sensitivity** configuration. In this setup, studies will be automatically excluded only if both algorithms classify them as “exclude,” while all others will remain subject to human screening. This strategy is expected to further increase overall sensitivity.

Following the main performance analysis, a detailed investigation of misclassifications in the high-sensitivity settings will be carried out to identify potential root causes of false exclusions. Moreover, key disagreements between LLM and human decisions will be reviewed in collaboration with the human screeners to gain deeper insights into the reasoning processes underlying both automated and human screening.

References

- [1] E. Sandner, B. Hu, A. Simiceanu, L. Fontana, I. Jakovljevic, A. Henriques, et al., “Screening Automation for Systematic Reviews: A 5-Tier Prompting Approach Meeting Cochrane’s Sensitivity Requirement,” in *Proc. 2024 2nd Int. Conf. Foundation and Large Language Models (FLLM)*, pp. 150–159, IEEE, Nov. 2024, doi: [10.1109/FLLM63129.2024.10852425](https://doi.org/10.1109/FLLM63129.2024.10852425)
- [2] E. Sandner, M. Negovetic, K. Kavita, L. Fontana, I. Jakovljevic, A. Henriques, et al., “Cal-X: Enhancing Systematic Review Screening with LLMs and Next-Token Likelihood Calibration,” in *Proc. 2025 3rd Int. Conf. Foundation and Large Language Models (FLLM)*, accepted for publication, IEEE, Nov. 2025.
- [3] CERN, *NeutrinoReview: Prototype of a Semi-Automated Systematic Review Management Tool*, CAIMIRA WP4, GitLab repository. Available: <https://gitlab.cern.ch/caimira/caimira-wp4/neutrinoreview>. Accessed: Nov. 1, 2025.
- [4] J. De Bruin, Y. Ma, G. Ferdinands, J. Teijema, and R. Van de Schoot, “SYNERGY – Open Machine Learning Dataset on Study Selection in Systematic Reviews,” *DataverseNL*, V1, 2023, doi: [10.34894/HE6NAQ](https://doi.org/10.34894/HE6NAQ).
- [5] E. Guo, M. Gupta, J. Deng, Y.-J. Park, M. Paget, and C. Naugler, “Automated Paper Screening for Clinical Reviews Using Large Language Models,” *Mendeley Data*, V1, 2023, doi: [10.17632/np79tmhkh5.1](https://doi.org/10.17632/np79tmhkh5.1).
- [6] Hao00Wang. (2025). *LLM Knowledge Cut-off Dates Summary* [Repository]. GitHub. <https://github.com/Hao00Wang/llm-knowledge-cutoff-dates>
- [7] World Health Organization, “*Indoor Airborne Risk Assessment in the Context of SARS-CoV-2: Description of Airborne Transmission Mechanism and Method to Develop a New Standardized Model for Risk Assessment*,” Geneva: World Health Organization, 2024. Available: <https://iris.who.int/handle/10665/376346>. License: CC BY-NC-SA 3.0 IGO.
- [8] L. Fontana and G. Buonanno, “Indoor air quality in Italy: a systematic review of measured indoor pollutant concentrations across residential, public, and occupational environments,” *OSF Preprints*, Oct. 24, 2025, doi: [10.17605/OSF.IO/92A3G](https://doi.org/10.17605/OSF.IO/92A3G).