Highlights

Weakly Supervised Active Learning for Abstract Screening Leveraging LLM-Based Pseudo-Labeling

Opeoluwa Akinseloyin,Xiaorui Jiang,Vasile Paladel

- Research highlights item 1
- Research highlights item 2
- Research highlights item 3

Weakly Supervised Active Learning for Abstract Screening Leveraging LLM-Based Pseudo-Labeling

Opeoluwa Akinseloyin[a], Xiaorui Jiang[b,*] and Vasile Paladel[a]

[a]Centre for Computational Sciences and Mathematical Modelling, Coventry University, Puma Way, Coventry, CV1 2TT, Warwickshire, United Kingdom
[b]School of Information, Journalism and Communications, The University of Sheffield, The Wave, 2 Whitham Road, Sheffield, S10 2AH, South Yorkshire, United Kingdom

ARTICLE INFO

Keywords:
Evidence Synthesis
Abstract Screening
Pseudo Labelling
Active Learning
Weakly Supervised Learning
Large Language Model

ABSTRACT

Abstract screening is a notoriously labour-intensive step in systematic reviews. AI-aided abstract screening faces several grand challenges, such as the strict requirement of near-total recall of relevant studies, lack of initial annotation, and extreme data imbalance. Active learning is the predominant solution for this challenging task, which however is remarkably time-consuming and tedious. To address these challenges, this paper introduces a weakly supervised learning framework leveraging large language models (LLM). The proposed approach employs LLMs to score and rank candidate studies based on their adherence to the inclusion criteria for relevant studies that are specified in the review protocol. Pseudo-labels are generated by assuming the top $T\%$ and bottom $B\%$ as positive and negative samples, respectively, for training an initial classifier without manual annotation. Experimental results on 28 systematic reviews from a well-established benchmark demonstrate a breakthrough in automated abstract screening: Manual annotation can be eliminated to safely reducing 42-43% of screening workload on average and maintaining near-perfect recall — the first approach that has succeeded in achieving this strict requirement for abstract screening. Additionally, LLM-based pseudo-labelling significantly improves the efficiency and utility of the active learning regime for abstract screening.

1. Introduction

A systematic review (SR) is an important productivity tool for synthesising the existing literature and evidence of a research topic from all available publications. It is critical for evidence-based medicine. Doing a rigorous SR is very labor-intensive and expensive. One estimate was approximately 67.3 weeks for completing an SR, which cost millions of dollars for academic institutions and pharmaceutical companies every year (Michelson and Reuter, 2019). An SR starts with searching scientific databases, which may return thousands or tens of thousands of candidate studies, most of which are actually irrelevant. Thus it is imperative to screen the candidate studies to decide which meet the inclusion criteria and should be included in the review. Often, screening is first done by checking the abstracts (and titles), called *abstract screening*—one of the most tedious and time-consuming SR steps (Michelson and Reuter, 2019). A cost-effectiveness analysis estimated that each reviewer spends 83 to 125 hours to screen 5000 references at a cost of approximately $17,000 (2013 prices) (Shemilt, Khan, Park, and Thomas, 2016). According to the best-practice guideline, which recommends at most 3 hours per reviewer per day to maintain screening performance and quality (Polanin, Pigott, Espelage, and Grotpeter, 2019), this amounts to 27.5 to 41.7 days just for abstract screening.

Responding to this high demand (Kitchenham and Brereton, 2013), much effort and process have been made in AI-aided automated systematic review tools, with a significant focus on abstract screening using machine learning (van Dinter, Tekinerdogan, and Catal, 2021; Dos Santos, da Silva, Couto, Reis, and Belo, 2023; Ofori-Boateng, Aceves-Martins, Wiratunga, and Moreno-Garcia, 2024). However, prominent challenges exist, hindering the widespread application of such tools. **Challenge 1:** One grand challenge is the *lack of labeled data*. The success of machine learning, particularly deep learning, hinges on the availability of high-quality labeled data, but each SR constitutes a distinct dataset without any initial annotations. Acquiring such labeled data is resource-intensive, especially when expert knowledge is required (Garg and Kalai, 2018). **Challenge 2:** This situation is further exacerbated by the *non-generalisability* of an abstract screener across review topics due to the fact that each SR constitutes a different dataset about a particular review topic. Therefore, models trained on one SR are rarely transferrable to other SRs, making manual annotation unavoidable for every new review.

Active learning has become the dominant strategy to mitigate the reliance on large labeled datasets by selecting the most informative examples to be labeled manually (Rhee, Erdenee, Kyun, Ahmed, and Jin, 2017). However, the success of active learning is determined by the composition strategy and quality of the initial training dataset (Hwang, Choi, and Choi, 2024). **Challenge 3:** This has been proved extremely challenging because abstract screening is an intrinsically extremely imbalanced classification task. The imbalance ratio between relevant/included and irrelevant/excluded studies often can be as high as 1:100. *Extreme*

*Corresponding author. The corresponding author is funded by the Royal Society's International Exchange Scheme (IES/R1231175).
akinseloyo@uni.coventry.ac.uk (O. Akinseloyin); xiaorui.jiang@sheffield.ac.uk (X. Jiang); vasile.palade@coventry.ac.uk (V. Paladel)
ORCID(s): 0000-0003-4255-5445 (X. Jiang); 0000-0002-6768-8394 (V. Paladel)

class imbalance is the major obstacle to efficiently building a high-quality initial training dataset due to the difficulty in discovering samples of the rare class (almost always the included studies) with minimal supervision (i.e., active discovery) (Szűcs and Papp, 2022). Additional difficulty also resides in the optimisation of the initial classifier with limited labeled data, especially without enough samples of the rare class. **Challenge 4:** Adding to all the afore-mentioned challenges is the non-negotiable requirement for *near-total recall* (defined as at least 95% at the minimum) of relevant studies, while maintaining as high precision as possible. Missing a few relevant studies will significantly weaken the reliability of the collected evidence for answering a medical problem and thus invalidate the systematic review.

To address these challenges, this paper proposes a zero-initialised active learning framework by introducing a negligible amount of financial overhead. Our work falls within the paradigm of *unsupervised active learning*, where the absence of sufficient labeled data requires leveraging unsupervised solutions for both selection and prediction (Souza, Rossi, Batista, and Rezende, 2017). Specifically, large Language Models (LLMs) are leveraged to answer predefined selection criteria questions and rank candidate studies based on their abstracts' adherence to the selection criteria. A set of pseudo-labeled studies is generated by assuming the top $T\%$ as positive samples and the bottom $B\%$ as negative samples. These pseudo-labels are used to train an initial classifier. The pseudo-labels make it possible to achieve a more informed starting point for the active learning cycle and continuously improve the effectiveness of the active learning cycle, which results in stronger screening performance and fast convergence.

This study has the following main contributions:

1. We propose a novel method for generating pseudo-labels by ranking candidate studies using LLMs, reducing the burden of manual labeling for abstract screening.
2. We demonstrate how these pseudo-labeled datasets can effectively initialise an active learning cycle, balancing perspectives of uncertainty and diversity.
3. We introduce a framework for zero-initialised unsupervised active learning that naturally addresses challenges of class imbalance and active discovery with little extra effort.
4. We perform extensive experiments on a large number of systematic reviews and evaluate the effectiveness of our framework in various aspects, including improving screening performance, addressing class imbalance and enhancing active learning.

2. Related Work

2.1. Machine Learning for Abstract Screening

There has been nearly two-decade research in using machine learning to automate or semi-automate systematic reviews. In the seminal work (Cohen, Hersh, Peterson, and Yen, 2006), abstarct words, MeSH (Medical Subject Headings) terms and MEDLINE publication tags were used to build document representation and a voting perceptron was trained to predict inclusion/exclusion decisions for abstract screening. Wallace, Trikalinos, Lau, Brodley, and Schmid (2010) extended documentation representation with "bag-of-biomedical-words" features based on UMLS (Universal Medical Language System) terms and trained a support vector machine (SVM) for classification. In 2021, an analysis of 41 studies revealed that SVM and Bayesian Network were the most popular machine learning algorithms and "bag-of-words" and TF-IDF were the most common feature extraction methods (van Dinter et al., 2021). A few tools were developed too, such as Abstrackr (Wallace, Small, Brodley, Lau, and Trikalinos, 2012), EPPI-Reviewer (Thomas, Brunton, and Graziosi, 2010), and SWIFT-Review (Howard, Phillips, Miller, Tandon, Mav, Shah, Holmgren, Pelch, Walker, Rooney et al., 2016). Recently, the deep auto-encoder architecture was employed to learn a strong document representation (Kontonatsios, Spencer, Matthew, and Korkontzelos, 2020). The most prominent challenge to machine learning is the need of enough labeled data for classifier training, which are unavailable at the outset of a systematic review. This challenge is worsened by the extreme data imbalance. The majority of the literature assumed a significant amount (e.g., 50%) of documents annotated with inclusion/exclusion labels (Cohen et al., 2006; Wallace et al., 2010, 2012; Howard et al., 2016), hindering their real-world applicability due to this high demand for unavoidable manual workload.

2.2. Active Learning for Abstract Screening

Active learning is a prominent method for efficiently training a classifier with minimal labelled data by iteratively selecting informative examples from an unlabeled pool and improving the classifier (Lewis, 1995). Due to the cold-start nature of abstract screening, i.e., having no annotated data at the beginning of a review, active learning has been widely applied to gradually enlarging the training dataset by suggesting samples to human reviewers to annotate (Wallace et al., 2010; Miwa, Thomas, O'Mara-Eves, and Ananiadou, 2014; Cormack and Grossman, 2016; Howard, Phillips, Tandon, Maharana, Elmore, Mav, Sedykh, Thayer, Merrick, Walker, Rooney, and Shah, 2020; van de Schoot, de Bruin, Schram, Zahedi, de Boer, Weijdema, Kramer, Huijts, Hoogerwerf, Ferdinands, Harkema, Willemsen, Ma, Fang, Hindriks, Tummers, and Oberski, 2021). Several sampling strategies exist. Wallace et al. (2010) query unlabeled documents using a uncertainty-based method, which prioritise instances closest to the classifier's decision boundary, assuming that solving these hard samples improves model performance. On the contrary, Miwa et al. (2014) and (Howard et al., 2020) adopt a certainty-based strategy, i.e., finding samples farthest from the decision boundary. While traditional methods measure uncertainty using posterior probability (Luo, Schwing, and Urtasun, 2013), recent advances either directly predict uncertainty or predict

data loss to approximate uncertainty (Hwang, Choi, and Choi, 2022). Many of these sampling strategies are also available in existing AI-aided systematic review tools, such as Abstrackr (Wallace et al., 2012), RobotAnalyst (Przybyła, Brockmeier, Kontonatsios, Le Pogam, McNaught, von Elm, Nolan, and Ananiadou, 2018), and ASReview (van de Schoot et al., 2021).

2.3. Constructing the Initial Training Dataset

The construction of the initial training set is a critical but less-explored issue. Diversity-based approaches first apply a clustering algorithm and select the instances that are closest to cluster centres for labelling (Kang, Ryu, and Kwon, 2004; Zhu, Wang, Yao, and Tsou, 2008). On the contrary, border-based methods select the samples near the boundaries between two or more clusters, as they are deemed more confusing to label and thus may provide additional discriminative power (Yuan, Han, Guan, Lee, and Lee, 2011). Yuan et al. (2011) proposed a hybrid strategy that combines the center-based and border-based methods for initial instance selection, which demonstrated outstanding performance compared the individual strategies in the active learning setting. Through a comprehensive evaluation, Xie and Yu (2017) found that the samples near cluster centers often yielded better accuracy than border-based and hybrid strategies. These methods may suffer from class imbalance in the initial training set, as abstract screening data is inherently highly imbalanced. It is challenging to obtain a representative set of positive samples and train an effective classifier. Therefore, some recent approaches created balanced datasets by selecting examples uniformly across clusters (Szűcs and Papp, 2022). Our method solves these challenges; it identifies more positive samples through LLM-based scoring and balances the initial training set using pseudo-labeling. From another angle, our method also falls within the paradigm of unsupervised active learning (UAL), where labeled data is either extremely limited or entirely absent (Souza et al., 2017). In UAL, selection and prediction are based on unsupervised methods, leveraging the cluster assumption that similar instances should share the same label (Wang, Chen, and Zhou, 2012). UAL also faces class imbalance, so our method can be seen as an innovative solution for bootstrapping the zero-initialised UAL process.

2.4. Leveraging LLMs for Abstract Screening

Large language models (LLMs) offer new opportunities for systematic reviews (Luo, Chen, Zhu, Wang, Liu, Lyu, Wang, Wang, and Chen, 2024), but results are somewhat daunting. Guo, Gupta, Deng, Parl, Page, and Naugler (2024) engineered inclusion and exclusion criteria into prompts for GPT-4 to make binary decisions, but the average sensitivity was only 76%. In (Oami, Okada, and Nakada, 2024), all five reviews reported performances and the pooled sensitivity was only 49%. Meanwhile, some light is also shed on the positive side. Matsui, Utsumi, Aoki, Maruki, Takeshima, and Takaesu (2024) applied a 3-layer approach for prompting GPTs to make sequential decisions based on study design, population, and intervention and control, which achieved human-comparable performances on two reviews. This indicates the value of explicitly reasoning over each selection criterion like what we propose in this paper (Sect. 3.1.2). Combining various LLMs' decisions may also suppress individual models' bias and improve recall (Li, Sun, and Tan, 2024; Oami et al., 2024; Sanghera, Thirunavukarasu, El Khoury, O'Logbon, Chen, Watt, Mahmood, Butt, Nishimura, and Soltan, 2025).

A potential way to safely deploy LLMs in the real-world systematic review practice is to rank, instead of classify, candidate studies, e.g., on a five-point Likert scale (Dennstädt, Zink, Putora, Hastings, and Cihoric, 2024; Issaiy, Ghanaati, Kolahi, Shakiba, Jalali, Zarei, Kazemian, Avanaki, and Firouznia, 2024). Our approach has some resemblance to this idea, but we aimed to address the challenges of zero-initialised active learning. Specifically, we score and rank candidate studies in a more nuanced way using LLMs' answers for each selection criterion, generate a small amount of pseudo labels from the ranking to create the initial training data without manual annotation, and demonstrate that a classifier trained on the weak labels generated by pseudo-labeling is able to safely rule out a decent percentage of studies from manual screening without hurting the sensitivity. Additionally, our approach ensures a more balanced and representative dataset, addressing issues of class imbalance and accelerating convergence in active learning.

3. Methodology

This study builds on a previously developed successful framework (Akinseloyin, Jiang, and Palade, 2024; Akinseloyin, Jiang, and Valade, 2025). The primary goals of our methodology are to create a fully automated pipeline to build an initial classifier that is effective at screening workload reduction with minimal to no human intervention and to optimise the active learning regime for abstract screening by leveraging LLMs' capabilities in answering questions about selection criteria, ranking candidate studies, and generating pseudo-labels.

3.1. LLM-Based Screening Prioritisation

3.1.1. Problem Definition

For an SR, suppose the unlabelled document set, i.e., the candidate studies (here titles and abstracts) for screening is denoted by $\mathcal{D} = \{d_1, d_2, \dots, d_N\}$, where d_i is the i-th document and N can be quite large, from thousands to tens of thousands. The abstract screening task aims to train a machine learning model $\mathcal{M}$ that assigns a binary label to each document: $\mathcal{M} : \mathcal{D} \mapsto \mathcal{Y}$, where $\mathcal{Y} \in \{0, 1\}$, "0" means exclusion and "1" means inclusion. To start the active learning cycle for abstract screening, we must obtain a small set of initial labels as the training set and split the dataset into two parts: $\mathcal{L} = \{(d_1, y_1, \dots, (d_l, y_l)\}$ is the set of labeled samples, where $y_j \in \{0, 1\}(\forall j \in \{1, \dots, l\})$, and $\mathcal{U} = \{d_{l+1}, \dots, d_N\}$ is the set of unlabeled samples. It is challenging to obtain a high-quality $\mathcal{L}$ to train a good classifier to start the active learning cycle.

3.1.2. Scoring Using Large Language Models

Each SR specifies a number of selection criteria that every included study must satisfy. Following Akinseloyin et al. (2024), each criterion is transformed into a question, again automatically by LLMs, with three possible answers, "Yes", "No" or "N/A" (not answerable, unsure or neutral). Let $\mathcal{Q} = \{q_1, \dots, q_K\}$ denote the set of inclusion criteria questions. Based on the questions, an LLM is used to score each study. For each document d_i, an LLM is first used to produce the answers for all the questions: $\mathcal{A}_i = \{a_{1,i}, \dots, a_{K,i}\}$, where $a_{k,i}$ is the answer for the k-th question q_k on the i-th document d_i. A "Yes" answer starts with the answer word "Positive" followed by an LLM-generated reason for the answer, while the "No" and "N/A" answers start with "Negative" and "Neutral" respectively. Then, a score is assigned to measure the adherence of document d_i to the k-th criterion, denoted by $\overline{s}(d_i, q_k)$, based on its answer $a_{k,i}$ and the reasoning text for question q_k. See Appendix A.1 for details of the prompt and an example. In this paper, this initial score is the probability that an answer text has a positive sentiment by a BART-based sentiment analyser (Muthukumar, 2021).

Using LLM-generated answers alone does not produce the optimal results (Akinseloyin et al., 2024; Guo et al., 2024; Oami et al., 2024) (Also see A.2 for result comparisons), so the following hypothesis was engineered to re-rank LLM-based results: Given an included study, there must be a certain part of it that matches the corresponding inclusion criterion, so it is natural to assume a high semantic relevance between the study and the criterion question. Therefore the initial answer score is adjusted by averaging it with the semantic relevance between the study and the corresponding question:

$$s(d_i, q_k) = (1 - \alpha) \cdot \overline{s}(d_i, q_k) + \alpha \cot r(d_i, q_k), \tag{1}$$

where $r(d_i, q_k))$ is the semantic relevance between the i-th document and k-th criterion question and $\alpha \in (0, 1)$ is the controlling parameter. In this paper, $r(d_i, q_k))$ is approximated by the cosine similarity between the text embeddings of document d_i and criterion question q_k. The initial score of each document is defined as the average of its scores with respect to all questions:

$$\overline{s}(d_i, \mathcal{Q}) = \frac{1}{K} \sum_{k=1}^{K} s(d_i, q_k). \tag{2}$$

Because an included study is expected to meet all selection criteria, we assume a high semantic relevance between it and the paragraph of inclusion criteria in the SR protocol, denoted by $\mathcal{Q}$. So, the document score is can be further adjusted using this document-level semantic relevance as follows:

$$s(d_i, \mathcal{Q}) = (1 - \beta) \cdot \overline{s}(d_i, \mathcal{Q}) + \beta \cdot r(d_i, \mathcal{Q}), \tag{3}$$

where $r(d_i, \mathcal{Q})$ is the cosine similarity between the text embeddings of the document and selection criteria paragraph and $\beta \in (0, 1)$ is another controlling parameter.

3.1.3. Ensemble of Large Language Models

Following Akinseloyin et al. (2025), an ensemble of LLMs, denoted as $\mathcal{G} = \{g_1, \dots, g_M\}$ with each g_j ($j \in \{1, 2, \dots, M\}$) being an LLM, was employed to reduce the potential bias of the individual LLMs (Also see Appendix A.2 for further analysis). For each SR with the selection criteria questions (i.e., selection criteria paragraph) $\mathcal{Q}$ and each candidate study d_i, the ensemble averages the scores assigned by each individual LLM g according to the selection criteria, denoted as $s_g(d_i, \mathcal{Q})$, according to the following equation. This method has been proved extremely effective for candidate study prioritisation (Akinseloyin et al., 2025) while maintaining minimal additional costs (Also see Sect. 5.6).

$$s(d_i, \mathcal{Q}) = \frac{1}{M} \sum_{\forall g \in \mathcal{G}} s_g(d_i, \mathcal{Q}). \tag{4}$$

3.2. Weakly Supervised Active Learning

3.2.1. Handle Data Imbalance by Pseudo-Labeling

Systematic reviews often exhibit a high degree of class imbalance, where the relevant documents ($y = 1$) are significantly outnumbered by the irrelevant ones ($y = 0$). Existing methods struggle to initialise a classifier effectively under such high imbalance. In this paper, we propose to generate and use pseudo-labels according to LLM-based document scores obtained using the method described in prior sections.

Suppose we sort the documents in descending order of their scores, forming a ranked list

$$\mathcal{R} = \{d_{(1)}, d_{(2)}, \dots, d_{(N)}\},$$

where $d_{(i)}$ represents the top i-th document according to LLM-assigned scores such that $\forall i <= j$, we have $s(d_{(i)}, \mathcal{Q}) >= s(d_{(j)}, \mathcal{Q})$. The ranking procedure prioritises the documents that are most likely relevant, which significantly increases the chance of identifying positive samples from the top of the ranked list $\mathcal{R}$. We propose to choose the top $T\%$ of the ranked documents as (pseudo-labelled) positive samples, denoted as $\mathcal{P}$, and correspondingly the bottom $B\%$ as (pseudo-labelled) negative samples, denoted as $\mathcal{N}$. More formally, they are created as follows:

$$\begin{aligned} \mathcal{P} &= \{d_{(i)} \mid d_{(i)} \in \mathcal{R}, i \leq \lfloor T\% \times N \rfloor\}, \\ \mathcal{N} &= \{d_{(i)} \mid d_{(i)} \in \mathcal{R}, i > \lfloor (1 - B\%) \times N \rfloor\}. \end{aligned} \tag{5}$$

Using the pseudo-labels, we are able to initialise the training set as follows in order to train an initial classifier and initialise the active learning cycle:

$$\overline{\mathcal{L}} = \{(d, y) \mid d \in \mathcal{P} \cup \mathcal{N}, y = 1 \text{ if } d \in \mathcal{P}, y = 0 \text{ if } d \in \mathcal{N}\}. \tag{6}$$

Because the initial training set is built using pseudo-labels, the complete dataset $\mathcal{D}$ is not annotated with human-assigned labels.

Algorithm 1 Active Learning Framework for Abstract Screening Based on Pseudo-Labeling

Input: Document set $\mathcal{D}$, an ensemble of one or several LLMs $\mathcal{G}$.
Output: Trained classifier f_θ.

1: Score documents using LLMs according to Eq. (3-4).
2: /*The following two steps differ from normal active learning.*/
3: Build the initial pseudo-labeled training set $\overline{\mathcal{L}}$ according to Eq. (6).
4: **Initialise**: Set $i \leftarrow 0$; initialise the set of labeled samples and the set of unlabeled samples before active learning starts: $\mathcal{L}^{(0)} \leftarrow \overline{\mathcal{L}}$, $\mathcal{U}^{(0)} \leftarrow \mathcal{D}$.
5: **repeat**
6: **Train**: Train ($i = 0$) or re-train a classifier $f_\theta^{(i)}$ using $\mathcal{L}^{(i)}$.
7: **Predict:** Use classifier $f_\theta^{(i)}$ to predict the samples in $\mathcal{U}^{(i)}$.
8: **Query**: Select a batch of unlabeled samples $\mathcal{B} \subset \mathcal{U}^{(i)}$ according to Eq. (8).
9: **Annotate**: For each sample $d \in \mathcal{B}$ assign its true label y^*.
10: /*The following step differs from normal active learning.*/
11: **Update**: $\mathcal{L}^{(i+1)} \leftarrow \mathcal{L}^{(i)} \setminus \{(d, y) | \forall d \in \mathcal{B}\} \cup \{(d, y^*) | \forall d \in \mathcal{B}\}$; $\mathcal{U}^{(i+1)} \leftarrow \mathcal{U}^{(i)} \setminus \mathcal{B}$.
12: **until** a stopping condition is met, e.g., achieving an estimated 95% recall threshold (Howard et al., 2020).

3.2.2. Initialising and Running Active Learning Cycle

After obtaining the pseudo-labeled training set, an initial probabilistic classifier f_θ is trained on $\overline{\mathcal{L}}$. Then batch-based active learning starts: use the classifier to select a batch of the most informative unlabeled samples, send them for human annotation, rebuild the training set with human-assigned labels, and repeat the process iteratively until a predefined stopping condition is met. The informativeness of an unlabeled document d is measured using uncertainty sampling according to the following equation (Lewis, 1995):

$$u(d) = 1 - \max(f_\theta(d), 1 - f_\theta(d)), \tag{7}$$

where $f_\theta(d)$ returns the posterior probability for document d. A batch $\mathcal{B}$ of k most uncertain samples is chosen as follows:

$$\mathcal{B} = \text{argmax}_{\mathcal{B}' \subset \mathcal{U}, |\mathcal{B}'|=k} \sum_{d \in \mathcal{B}'} u(d). \tag{8}$$

Algorithm 1 summarises how our framework initialises and runs the active learning cycle, called *weakly supervised active learning*, for abstract screening leveraging pseudo-labelling. The overall process is similar to traditional active learning with three notable differences. Firstly, Algorithm 1 uses LLM-assigned pseudo-labels to initialise a training set to start the active learning cycle (Step 3). Secondly, during the active learning cycle, the initial pseudo-labelled training set will be improved by new samples, so $\mathcal{L}^{(0)}$ is initialised by $\overline{\mathcal{L}}$ (Step 4), which will be iteratively updated. Thirdly, after the active learning cycle is started, the newly sampled documents are labeled and used to improve the training set for re-training: (i) if a sample is not in the pseudo-labelled training set, then simply add it to the training set; (ii) otherwise, replace the label of the original sample with the label assigned by human annotator (Step 11). Note that in Step 11, each selected sample (d, y) will only be visited once as it will be removed from the unlabeled set. A noteworthy feature of our weakly supervised active learning framework is that it prevents class imbalance from regaining dominance because the initial set of pseudo-labeled samples (i.e. $\overline{\mathcal{L}}$) are always part of the training data. In addition, out framework allows active learning to pick more positive samples during the whole active learning cycle and converge much faster than traditional methods, which is a second factor to alleviate class imbalance.

4. Experimental Setup

4.1. LLM Setup

For LLM-based scoring and ranking, we used three cheap LLMs by some of the most famous companies in the LLM industry: GPT-4o mini by OpenAI (more precisely, **gpt-4o-mini-2024-07-18**), Gemini 1.5 Flash by Google AI (more precisely, **gemini-1.5-flash-001**), and Claude 3 Haiku by Anthropic (more precisely, **claude-3-haiku-20240307**). The APIs of all three LLMs allow setting the temperature to 0 to ensure replicability, so a temperature of zero was employed throughout our experiments to maintain the stability of the generated responses. For other LLM options, the default values were used. The text embedding model for generating the representation vectors for candidate studies was OpenAI's **text-embedding-3-large**. In our experiments, both α and β are set to 0.5 for calcualting the re-ranked document scores.

4.2. Datasets

We use 28 datasets from the CLEF eHealth 2019 Task 2, a well-established benchmark for technology-assisted review in empirical medicine (Kanoulas, Li, Azzopardi, and Spijker, 2019). 20 SRs about clinical intervention trials (**Intervention**) and 8 SRs about diagnostic test accuracy (**DTA**) are used for evaluation, with each SR constituting a distinct dataset. Each SR contains a pool of candidate studies (titles and abstracts) and the ground-truth include/exclude labels. Dataset statistics can be found in (Kanoulas et al., 2019).

4.3. Evaluation Metrics

Traditional Metrics for Abstract Screening The abstract screening task entails labeling *all* documents in a finite pool so that no potentially relevant study is missed. Consequently, the most critical requirement for abstract screening is maintaining *(near) total recall* (or near perfect sensitivity): failing to retain relevant studies will severely undermine the validity of an entire review (Kanoulas et al., 2019). Precision and accuracy are only useful when (near) total recall (of positive samples, i.e., included studies) is achieved. In the field of automated systematic review, the dominant performance metric for abstract screening is Work Saved over Sampling (Cohen et al., 2006), which is defined below.

- **Work Saved over Sampling at Recall Level R% (WSS@R%)**:

$$\text{WSS@}R\% = 1 - \frac{TN + FN}{N} - (1.0 - R\%),$$

where TN and FN are the true and false negatives, respectively, at recall level R%, and N is the total number of studies. In practice,WSS@95% is often used, which quantifies the percentage of the dataset that can be *skipped* from manual screening once 95% of the relevant studies are found.

Finite-Pool Active Learning Metrics Following Miwa et al. (2014), we also incorporate the following metrics tailored to finite-pool active learning (Wallace, Small, Brodley, and Trikalinos, 2010):

- **Yield** captures the fraction of relevant documents *ultimately* identified (at the end of an iteration of the active learning cycle). Formally:

$$\text{yield} = \frac{\text{TP}_L + \text{TP}_U}{(\text{TP}_L + \text{FN}_U) + \text{TP}_U},$$

where TP_L and TP_U are the true positives among the labeled and unlabeled sets, respectively, and FN_U denotes false negatives in the unlabeled pool.

- **Utility**$_\beta$, a weighted combination of yield burden:

$$\text{utility}_\beta = \frac{\beta \cdot \text{yield} + (1 - \text{burden})}{\beta + 1}, \tag{9}$$

where burden($|\mathcal{L}|/(|\mathcal{L}| + |\mathcal{U}|)$ and it measures the proportion of studies that the reviewer must manually inspect, and the parameter β reflects how critical recall is compared to minimising workload. A a lower burden indicates less human effort and a larger β means prioritising recall (sensitivity) over workload reduction.

4.4. Baselines and Proposed Approach

Most sampling strategies are applied to selecting samples after the active learning cycle has been started, but the focus of the current paper is on improving the quality of the initial training data. Regarding initialising a classifier *without* pre-labeled data, only a few methods exist, which are compared to as baselines. They include:

- `Centroid` (Kang et al., 2004): The Centroid method selects the samples near cluster centroids (by k-means, $k = 2$) to initialise the classifier.
- `Border` (Yuan et al., 2011): The Border method chooses the samples around cluster boundaries.
- `Hybrid` (Yuan et al., 2011): The Hybrid method uses Centroid and Border to each select half of the samples.
- The `Random` method randomly initialises a training set.

There are two variants of our approach. The details of used LLMs are described in Appendix **??**.

- **LLM-Based Pseudo-Labeling (`LLM-Pseudo`)**: Candidate studies are scored according to Eq. (4). The top T% are pseudo-labeled as positive samples and the bottom B% as negative according to Eq. (5), circumventing any need for manual annotations at the outset.
- **LLM-Ranking & Expert Annotation (`LLM-True`)**: This is a variant of our approach by using the *gold-standard* labels of the top T% and bottom B% documents based on LLM ranking.

4.5. The Active Learning Protocol

To train the initial classifier, each baseline method and our own approaches use 5% of the whole dataset to form the initial training set (i.e., $T = B = 2.5$). In the LLM-Pseudo method, the top and bottom 2.5% are assigned the positive (included) and negative (excluded) pseudo-labels, while in the LLM-True method the ground-truth labels are used, which indicates that this 5% need to be annotated by human experts. In the Centroid, Border and Hybrid methods, 2.5% of the documents from each of the two clusters are used and annotated by human reviewers.

Active learning then proceeds according to Algorithm 1 described in Sect. 3.2.2, where new documents are queried based on *certainty-based sampling*. Throughout the cycle, we track both traditional metrics (recall, accuracy, WSS@95%) and finite-pool metrics (yield, utility$_\beta$). Note that, our approach is not only used for constructing the initial training set (Step 4 in Algorithm 1) and initialising the classifier, but also used to improve the training set by adding more labelled samples and positive samples, and by correcting the initial training set's labels (Step 11 in Algorithm 1).

Note that the current paper focuses on the improving the active learning cycle from angle of better-quality training data, so we follow the common choices in prior research in abstract screening based on active learning (Wallace et al., 2010; Miwa et al., 2014; van Dinter, Catal, and Tekinerdogan, 2021). To be precise, a linear support vector machine (SVM) and certainty-based sampling are used during the active learning cycle. According to Ofori-Boateng, Trujillo-Escobar, Aceves-Martins, Wiratunga, and Moreno-Garcia (2024), document embedding is the most effective feature engineering approach, so GPT embeddings (Sect. 4.1) are used in our experiments. Although our framework significantly alleviates class imbalance, we still apply weights on

Method	Pos-Neg Ratio	Recall	Accuracy	WSS@95%
Border	0.003 ± 0.005	0.001 ± 0.001	0.092 ± 0.119	0.009 ± 0.015
Centroid	0.191 ± 0.123	0.549 ± 0.176	0.766 ± 0.128	0.470 ± 0.152
Hybrid	0.073 ± 0.044	0.375 ± 0.173	0.634 ± 0.181	0.358 ± 0.162
Random	0.066 ± 0.038	0.243 ± 0.088	0.648 ± 0.109	0.259 ± 0.109
LLM-True	0.316 ± 0.118	0.298 ± 0.175	0.879 ± 0.072	0.355 ± 0.142
LLM-Pseudo	1.000 ± 0.000	$\mathbf{0.998 \pm 0.002}$	0.474 ± 0.074	$\mathbf{0.664 \pm 0.097}$

Table 1
Initial classifier's performance on Intervention.

Method	Pos-Neg Ratio	Recall	Accuracy	WSS@95%
Border	0.001 ± 0.002	0.000 ± 0.000	0.117 ± 0.202	0.056 ± 0.098
Centroid	0.166 ± 0.127	0.585 ± 0.242	0.871 ± 0.099	0.534 ± 0.232
Hybrid	0.170 ± 0.133	0.635 ± 0.244	0.692 ± 0.210	0.522 ± 0.217
Random	0.114 ± 0.138	0.297 ± 0.151	0.722 ± 0.161	0.407 ± 0.174
LLM-True	0.197 ± 0.105	0.144 ± 0.209	0.880 ± 0.110	0.437 ± 0.209
LLM-Pseudo	1.000 ± 0.000	$\mathbf{0.994 \pm 0.006}$	0.515 ± 0.054	$\mathbf{0.675 \pm 0.157}$

Table 2
Initial classifier's performance on DTA.

the positive and negative classes when training or re-training the SVM, based on this equation: $w_j = N/(K \times N_j)$, where w_j is the weight for the j-th class, $K = 2$ is the total number of classes, N is total number of samples in the training set, and N_j is the size of the j-th class. Although a large volume of literature exists for improving other aspects of active learning, they are not the foci of the current study and are not explored as the investigation of these factors will constitute a separate study.

5. Results and Discussion

5.1. Mitigating Class Imbalance

Table 1 and 2 present the average ratio of positive to negative studies (the "Pos-Neg Ratio" columns) by six initialisation methods across two review types (Intervention and DTA). The Border and `Random` methods identify extremely small fractions of positive studies (e.g., on average 0.003 and 0.066, respectively, for Intervention), underscoring how conventional approaches are susceptible to exacerbating class imbalance. `Centroid` and `Hybrid` perform slightly better but still struggle to ensure a sufficient number of relevant studies in the initial sample set. `LLM-True` leverages the actual labels assigned by human reviewers, so it achieves higher ratios (0.316 and 0.197); however notable imbalance still exists.

In contrast, our `LLM-Pseudo` method sets the positive-to-negative ratio to 1.0 by design, implicitly mitigating class imbalance. Balanced sampling plays a crucial role in improving active learning for systematic reviews, which aims to achieve total recall of relevant studies. By guaranteeing a sufficient number of positive examples early on, our approach allows the model to converge more rapidly and accurately, thus reducing the overall screening workload. Furthermore, because it derives pseudo-labels from LLM-based relevance ranking, no human annotation is needed at the outset, thereby further minimising manual effort.

5.2. The Screening Power of the Initial Classifier

Tables 1 (Intervention) and 2 (DTA) summarise the recall, accuracy, and WSS@95% of the classifiers initialised by six methods. Since systematic reviews strive to avoid missing relevant studies, methods like `Border` and `Random` are immediately unsuitable due to their low recall levels. `Centroid` and `Hybrid` provide moderate improvements but still fail to meet the requirement of above 95% recall. Although the `LLM-True` offers relatively high accuracy (e.g., 0.879 in Intervention), its recall remains insufficient. By contrast, `LLM-Pseudo` achieves near-perfect recall: on avergae 0.998 for Intervention and 0.994 for DTA. More importantly, `LLM-Pseudo` meets this critical threshold on all 28 reviews, indicating the universal applicability and trustworthy deployment of our approach to the real-world systematic review practice. `LLM-Pseudo` also achieves the highest WSS@95%, indicating a substantial amount of potential workload reduction. Note that all other methods require human reviewers to manually label at least 5% of the candidate studies at the outset. This amount of unavoidable workload causes a small further reduction to the reported WSS@95% values of these methods in the tables.

Method	Intervention	DTA
LLM-Pseudo Theoretical WSS@95%	0.664 ± 0.108	0.675 ± 0.097
LLM-Pseudo Actual WSS	0.417 ± 0.098	0.432 ± 0.157
Cali-LLM-Ens Actual WSS	0.28	0.35

Table 3
Comparison of Theoretical WSS and Actual WSS.

Our `LLM-Pseudo` method can be deemed a breakthrough in semi-automated abstract screening. It achieves average accuracies of 47.4% and 51.5% on Intervention and DTA, respectively. Although its accuracy is lower than other baselines, `LLM-Pseudo` is the only method that has the potential to be accepted by human reviewers due to the strict requirement of near-total recall. From the near-total recall and low positive-to-negative ratio, we can conclude that errors primarily manifest as false positives (i.e., including irrelevant studies). This characteristic enables the safe elimination of a substantial portion of candidate studies from manual screening based solely on the initial classifier's verdicts, achieving an average reduction of 41%-43% on all 28 reviews (refer to Sect. 5.3 for more in-depth discussions). These results highlight the benefits of our LLM-based pseudo-labeling approach, which completely eliminates the need for upfront manual annotation and implicitly mitigates the extreme class imbalance.

5.3. Theoretical v.s. Actual Workload Savings

Most prior studies, if not all, report the *Theoretical WSS@95%*, which represents an oracle-like scenario, where the WSS@95% is computed by adjusting the threshold for the classifier's posterior probabilities and stopping exactly at the point where 95% of positive studies are retrieved. However, in practical screening scenarios, one cannot know a priori when precisely 95% recall has been achieved without having full knowledge of the dataset. Therefore, theoretical WSS inflates the actual workload saving. Our approach empirically guarantees 95% recall on all SRs, which allows us to calculate the *Actual WSS* by replacing R in the WSS equation

with the actual recall obtained by the initial classifier. WSS inflation is also caused by the fact that prior studies require a fixed fraction of the dataset to be manually annotated. In this case, WSS@95% is only calculated on the remaining unannotated portion. In contrast, our methodology eliminates the need for any initial annotation, thereby saving screening load across the entire dataset.

Table 3 compares Theoretical WSS@95% and Actual WSS on both review categories. Here we also compare with a recent "zero-shot" method which makes probabilistic decisions by instructing LLMs (Wang, Scells, Zhuang, Potthast, Koopman, and Zuccon, 2024), although it is not entirely zero-shot because the probability threshold is calibrated on a separate large set of reviews as training data. The best results of their ensemble method that meets the minimal requirement for recall are retrieved and compared, named **`Cali-LLM-Ens`** in Table 3. For LLM-Pseudo, the average Actual WSS values are 0.417 for Intervention and 0.432 for DTA. They are understandably lower than the Theoretical WSS@95% values (0.664 and 0.675, respectively), but significantly higher than `Cali-LLM-Ens`. These results are meaningful in two aspects. On the one hand, on average 41%-43% of screening load can be safely saved on both review categories without any human annotation, which can be considered as a breakthrough.

5.4. Impact of Pseudo-Labeling on Active Learning

This section demonstrates how pseudo-labeling benefits active learning through several simulation studies, focusing on one systematic review (review ID: CD011768; category: Intervention; Size: 8963; Number of included studies: 53). The widely-adopted protocol for simulation studies in prior research (Wallace et al., 2010; Miwa et al., 2014; Cormack and Grossman, 2016; Howard et al., 2020; van de Schoot et al., 2021) was obeyed. Before active learning is started, all six methods (Sect. 4.4) are used for creating the initial training set and training the initial classifier. Then, certainty-based sampling was applied to select the instance the model is most confident is positive, i.e., $|\mathcal{B}| = 1$. This approach is particularly effective in abstract screening when prioritising high-confidence positives increases the representation of relevant studies in the labeled set, which in turn helps the classifier learn more discriminative features for identifying relevant literature (Miwa et al., 2014; van de Schoot et al., 2021) in future iterations.

Figure 1 illustrates how positive samples (relevant studies) are accumulated throughout the active learning cycle. The `Border` method (green line) struggles to find any positives early on, while `Random` (blue) and `Centroid` (orange) increase recall at a slower pace. By contrast, `LLM-Pseudo` (brown) rapidly detects relevant studies and achieves the 95% recall threshold (pink dashed horizontal line) much faster than all other methods. This swift ascent is critical in systematic reviews, where missing even a few essential citations can compromise the overall findings. This LLM-driven strategy consistently targets the most informative

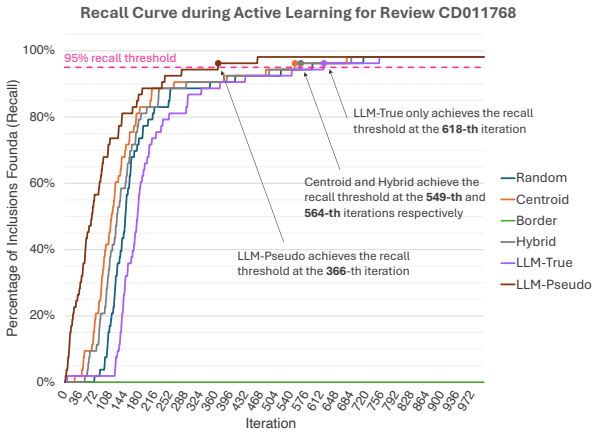

Figure 1: Pseudo-labeling makes active learning converge faster to reach the required recall level: A case study.

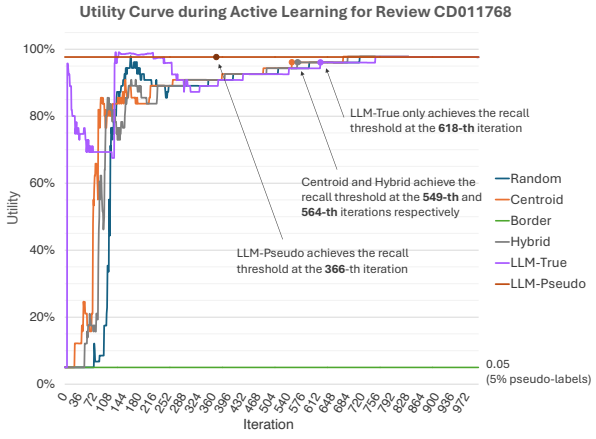

Figure 2: Pseudo-labeling improves the utility of active learning in abstract screening: A case study.

samples, minimising the effort required to reach a near-complete recall of relevant studies.

Figure 2 shows each approach's utility under a high penalty for missing positives ($\beta = 19$). The `Border` method (green line) remains near zero, indicating it fails to quickly locate enough relevant documents. `Random`, `Centroid`, `Hybrid`, and `LLM-True` (blue, orange, grey, and purple, respectively) all rise sharply but occasionally fluctuate, reflecting dips when less-informative samples are chosen. By contrast, `LLM-Pseudo` (brown) rapidly attains near-perfect utility and sustains that level, demonstrating an ability to identify critical documents early while minimising reviewer workload. Overall, these trends highlight how LLM-based pseudo-labeling effectively balances sensitivity and annotation effort. This efficiency is further quantified by examining how fast each method first achieves the target recall threshold of 95%. As annotated in Figure 1, `LLM-Pseudo` reaches this threshold by the 366-th iteration, substantially earlier than the `Centroid` (549-th) and `Hybrid` (564-th) methods, resulting in an improvement of over 1/3 on the efficiency of active learning. Reaching high recall earlier reduces the number of samples that need to be manually labeled, thus lowering

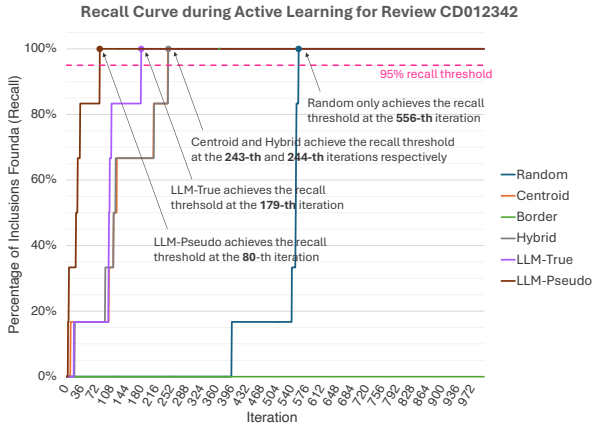

Figure 3: Pseudo-labeling makes active learning converge faster to reach the required recall level: A second case study.

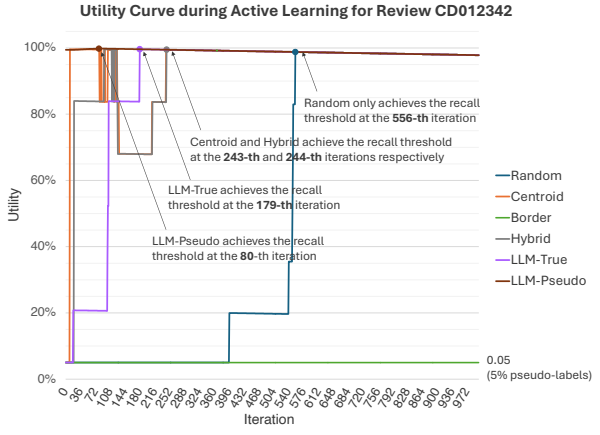

Figure 4: Pseudo-labeling improves the utility of active learning in abstract screening: A second case study.

the *burden* component of the utility metric (Eq. 9). As a result, `LLM-Pseudo` maintains higher utility values throughout the active learning process as shown in Figure 2, though subject to a negligible decline due to more samples being annotated. This better trade-off between recall and human effort reinforces its practical benefit in resource-constrained review settings where both sensitivity and efficiency are paramount.

An additional case study on the Intervention review CD012342 is given in Figure 3 and 4. `Centroid` and `Hybrid` significantly improve the efficiency of active learning compared to `Random`. They reach the 95% recall threshold at the 243-th and 244-th iterations respectively, `LLM-Pseudo` allows active learning to converge much earlier, at the 80-th iteration, resulting in a signfiicant speedup of about 2/3 on this particular review.

5.5. Why is Pseudo-Labeling Effective?

An interesting question is why pseudo-labeling is so effective. To answer this question, we evaluate the effect of varying the top $n\%$ threshold (in our experiments $n = T + B$ and $T = B$) on the pseudo-labeled training set. In Figure 5, by varying $n\%$ from 5% to 50% (i.e., $T\%$ from

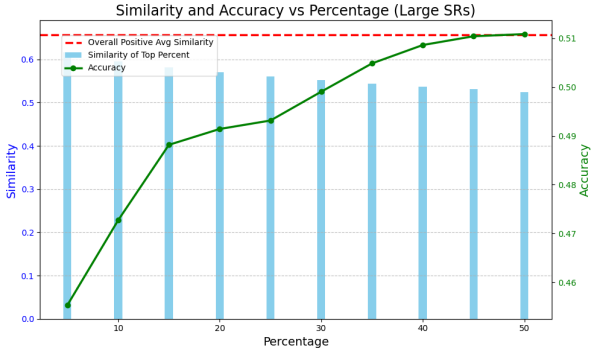

Figure 5: Impact of the pseudo-labeled size ($n\%$) on the quality of the initial training set, explaining why pseudo-labeling works.

2.5% to 25%), we compare the average cosine similarity among the true positive samples (the red dotted line) to the average similarity between the pseudo-labeled positive and the true positive samples (the blue bars) as well as the corresponding initial classifier's accuracy (the green line), both averaged across all 28 reviews. For small values of $n\%$ (e.g., 5–10%), the pseudo-labeled positive samples exhibit a similarity close to or exceeding the average similarity of true positives, suggesting that the pseudo-labeling approach is effective at capturing highly semantically similar, thus potentially relevant documents to the review topic. As $n\%$ increases, the similarity between the pseudo-labeled positive and true positive samples decreases, reflecting a dilution effect where more and more irrelevant samples are included in the top $T\%$ and pseudo-labeled as positive. However, accuracy improves steadily as the classifier incorporates more data, reaching a plateau beyond 30%.

This analysis highlights the trade-off inherent in pseudo-labeling: A smaller $n\%$ ensures higher precision by focusing on the most relevant studies, but risks omitting potentially useful data. Conversely, a larger $n\%$ allows for broader coverage of relevant documents at the cost of introducing a higher proportion of less relevant documents. Determining the optimal $n\%$ is critical for balancing precision and recall, ensuring the effectiveness of pseudo-labeling in reducing manual workload without compromising classification performance.

5.6. Cost Effectiveness Analysis

Table 4 shows the calculation of the costs of the individual LLMs and the ensemble, where "Price" is set by LLM providers for every million tokens (MTks), the "Size" columns show the total sizes of the inputs to LLMs and the outputs geenrated by LLMs for all the abstracts of the 28 SRs, the "Cost" columns show the total costs spent on each LLM. Correspondingly, the "Avg. Size" and "Avg. Cost" columns show the average size (number of tokens) and costs for the LLM inputs and outputs per abstract. The total cost for LLM APIs used in our methodology is only $66.21 for 66677 abstracts in total. Abstract wise the cost is negligible.

Table 3 shows that our LLM-Pseudo approach can safely reduce on average 41.7% and 43.2%, respectively, of the

Models	LLM Inputs			LLM Outputs			Sum	
	Price	Size	Cost	Price	Size	Cost	Total Cost	Avg. Cost
Gemini 1.5 Flash	$0.075	28.32 MTks	$2.12	$0.3	18.50 MTks	$5.55	$7.67	$1.93E-04
GPT-4o Mini	$0.15	28.32 MTks	$4.25	$0.6	20.93 MTks	$12.56	$16.81	$4.22E-04
Claude 3 Haiku	$0.25	28.32 MTks	$7.08	$1.25	27.72 MTks	$34.65	$41.73	$1.05E-03
Averaging (Ensemble)	-	-	$13.45	-	-	$52.76	$66.21	$ 1.66E-03

Table 4
Cost effectiveness analysis of LLM-based pseudo labelling.

screening load on Intervention (in total 39847 abstracts) and DTA (26830), eliminating the need for any human intervention. They amount to $39847 \times 41.7\% \approx 16993$ studies about Intervention and $26830 \times 43.2\% \approx 11591$ studies about DTA, in total about $(16993 + 11591 =)28584$ candidate studies saved from manual screening.

The Cochrane Handbook for Systematic Reviews of Interventions suggests an "estimated reading rate of one or two abstracts per minute" (Lefebvre, Glanville, Briscoe, Featherstone, Littlewood, Metzendorf, Noel-Storr, Paynter, Rader, Thomas, and Wieland, 2024), which we believe is applicable to systematic reviewers of very high domain expertise. According to this reading rate, the minimal savings made by our approach will be to $(28584 \div 2 \div 60 =)238.2$ hours to 476.4 hours, which are approximately 31.76 to 63.52 days per reviewer (calculated based on the UK standard of 7.5 working hours per day). Suppose the best practice guideline for large-scale abstract screening (Polanin et al., 2019) is followed, which suggests at most 2-3 hours per reviewer per day, the savings will be about $(238.2 \div 3 =)79.4$ days to $(476.4 \div 2 =)238.2$ days per reviewer. This is indeed a significant amount of savings at a very low cost! Note that, the Actual WSS reported in Table 3 establishes the lower bound, so we envision much better automated screening performance in future work that is built upon the current study.

It will be harder to estimate the financial savings brought by the LLM-based pseudo-labeling approach. However, a conservative estimation can be made by supposing a postgraduate research assistant does abstract screening and assuming a low-end annual salary at about £22,000 (according to the statistics on Grassdoor, roughly $29526 according to the exchange rate on 22 May 2025), a total number of 200 working days and 7.5 work hours per day according to the normal UK standard, without considering the overhead for national insurance and pension, etc. The hourly salary for a postdoc-level reviewer is about $\$29526 \div (7.5 \times 200) \approx \$19.68$ per hour. So, the estimated amount of financial savings per reviewer can be from $(238.2 \times 19.68) \approx \$4687.78$ to $(476.4 \times 19.68) \approx \$9375.55$, at the cost of paying less than $66.21 for API calls, approximately 1/140 of the manual labelling cost. If the reviewer is a postdoctoral researcher with a base annual salary of £37000 (roughly $49603), the financial savings will be increased to as high as $15754.55 per reviewer.

6. Conclusion

This paper introduces a zero-initialised active learning framework that leverages LLM-based pseudo-labeling to overcome the lack of initial labeled data in systematic reviews. By automatically ranking and designating the top $T\%$ as positive and the bottom $B\%$ as negative samples, the method ensures a balanced, high-quality initial training set without manual intervention. Experiments on 28 systematic reviews of a well-established benchmark for technology-assisted review demonstrate near-perfect recall of relevant studies and significantly higher workload reduction compared to conventional methods. Particularly, pseudo-labeling enables training a classifier that can safely reduce manual screening workload by more than 40% without the need for any human labeling at minimal LLM API costs. It is currently the only approach proven to achieve the minimum 95% recall threshold required for automated abstract screening across a large number of systematic reviews. Moreover, the results highlight the value of combining LLM-based pseudo-labeling with active learning to accelerate abstract screening, leading to faster convergence and improved utility. In future work, further validation studies will be conducted and additional strategies will be explored to enhance LLMs' question-answering capabilities and to optimise the scoring, ranking and pseudo-labeling methods.

A. Appendices

A.1. LLM Prompt and Example

Figure 6 shows the prompt we used for experiments. To elicit the reasoning capability of LLMs, we also prompted an LLM to describe how he derived the answer, which fills the "Reasoning path" section of the structured output in a predefined format shown in the figure, as well as how confident the LLM was about his answer, which is a float number between 0 and 1 filled in the "Confident level" section of the structured output. The "Extra information" section returned by an LLM typically included elements such as abbreviations or contextual nuances that the model might not fully understand. This helped in identifying areas where the model's predictions could be further clarified or corrected, particularly when errors occured.

Figure 7 shows an example, including the title and abstract of the abstract of a candidate study, and an inclusion criterion (converted to a question), and Figure 8 shows the corresponding output generated by an LLM using the prompt presented in Figure 6.

You are a researcher screening titles and abstracts of scientific papers for the systematic review '**{review_title}**'.

Analyse the abstract below within the brackets and answer the question below. Taking a step-by-step approach towards reasoning and answering the question.

Question: **{question}**

Keep your answers as short as possible. The answer should be in either a positive, neutral, or negative sentiment format The answer must contain your answer, how you got to your answer (reasoning path), confidence of your answer from 0 to 100 and finally what extra information would make you more confident in your answer. Format of the answer should be like the example below in text(string) not JSON or Code:

('Answer': ,
'Reasoning path': ,
'Confidence level': ,
'Extra Information':
)

Figure 6: Prompt for answering inclusion criteria questions.

Model	MAP	R@50%	WSS@95%
GPT-4o_mini	0.4461	96.97%	0.6558
Gemini 1.5 flash	0.4665	97.78%	0.6740
Claude Haiku	0.4144	97.03%	0.6321
Ensemble	**0.4703**	**97.87%**	**0.6961**

Table 5
Performance comparisons of LLMs and the ensemble.

A.2. Additional Results

Table 5 compares the performances of the LLM ensemble (for scoring and ranking candidate studies) against the three individual LLMs as baselines across three key metrics: Mean Average Precision (MAP), Recall at 50% screening (R@50%), and WSS@95. The ensemble method consistently outperformed the individual baseline models across all metrics, achieving the highest MAP (0.470), R@50% (97.87%), and theoretical WSS@95 (69.61%). These results underscore the value of model ensembling in enhancing systematic review screening performance by capturing the complementary strengths of individual models.

Question:

1. Does the study involve full-text diagnostic test accuracy research on any individual index test or a combination of tests against a reference standard?

Title:

Neck circumference to thyromental distance ratio: a new predictor of difficult intubation in obese patients.

Abstract:

Background: This study was performed to assess whether intubation is more difficult in obese patients and to assess the ability of a new index: the ratio of the neck circumference to thyromental distance (NC/TM), to predict difficult intubation in obese patients.

Methods: The incidence of difficult tracheal intubation in 123 obese (BMI⩾27.5 kg m(-2)) and 125 non-obese patients was compared. Difficult intubation was determined using the intubation difficulty scale (IDS⩾5). The NC/TM ratio was calculated and its ability to predict difficult intubation in obese patients was compared with that of established predictors including high BMI, the Mallampati score, the Wilson score, NC, width of mouth opening, sternomental distance, TM, and a previous history of difficult intubation.

Results: Difficult intubation was more frequent in obese patients than in non-obese patients (13.8% vs 4.8%; P=0.016). Multivariate analysis revealed that the Mallampati score, the Wilson score, and NC/TM independently predicted difficult intubation in obese patients. Among these three indices, NC/TM showed the highest sensitivity and a negative predictive value, and largest area under the curve on an ROC curve.

Conclusions: Difficult intubation was more common in obese patients and the NC/TM was a better method for predicting difficult intubation than other established indices.

Figure 7: An example of an inclusion criterion question on a candidate study (abstract).

References

M. Michelson, K. Reuter, The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials, Contemporary Clinical Trials Communications 16 (2019) 100443.

I. Shemilt, N. Khan, S. Park, J. Thomas, Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews, Systematic Reviews 5 (2016) 1–13.

J. R. Polanin, T. D. Pigott, D. L. Espelage, J. K. Grotpeter, Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses, Research Synthesis Methods 10 (2019) 330–342.

B. Kitchenham, P. Brereton, A systematic review of systematic review process research in software engineering, Information and Software Technology 55 (2013) 2049–2075.

R. van Dinter, B. Tekinerdogan, C. Catal, Automation of systematic literature reviews: A systematic literature review, Information and Software Technology 136 (2021) 106589.

Á. O. Dos Santos, E. S. da Silva, L. M. Couto, G. V. L. Reis, V. S. Belo, The use of artificial intelligence for automating or semi-automating biomedical literature analyses: a scoping review, Journal of Biomedical Informatics 142 (2023) 104389.

R. Ofori-Boateng, M. Aceves-Martins, N. Wiratunga, C. F. Moreno-Garcia, Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review, Artificial Intelligence Review 57 (2024) 200.

V. Garg, A. T. Kalai, Supervising unsupervised learning, Advances in Neural Information Processing Systems 31 (2018).

P. K. Rhee, E. Erdenee, S. D. Kyun, M. U. Ahmed, S. Jin, Active and semi-supervised learning for object detection with imperfect data, Cognitive Systems Research 45 (2017) 109–123.

(
"Answer": "Negative",
"Reasoning path": "The abstract describes a retrospective study assessing the prevalence of vascular disease in patients with bronchiectasis, without any mention of inhaled corticosteroids or a randomized controlled trial design. It focuses on vascular disease risk factors rather than treatment comparisons. ",
"Confidence level": 95,
"Extra Information": "Access to the full text of the study to confirm the absence of any treatment-related comparisons would increase confidence."
)

Figure 8: An example of the answers generated by an LLM for an inclusion criterion question.

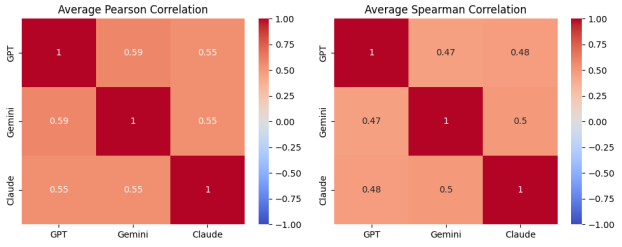

Figure 9: Correlations between three baseline LLMs.

S. Hwang, J. Choi, J. Choi, Strategic construction of initial datasets for active learning: Leveraging self-supervised learning (2024).
G. Szűcs, D. Papp, Zero initialised unsupervised active learning by optimally balanced entropy-based sampling for imbalanced problems, Journal of Experimental & Theoretical Artificial Intelligence 34 (2022) 781–814.
V. M. Souza, R. G. Rossi, G. E. Batista, S. O. Rezende, Unsupervised active learning techniques for labeling training sets: an experimental evaluation on sequential data, Intelligent Data Analysis 21 (2017) 1061–1095.
A. M. Cohen, W. R. Hersh, K. Peterson, P.-Y. Yen, Reducing workload in systematic review preparation using automated citation classification, Journal of the American Medical Informatics Association 13 (2006) 206–219.
B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, C. H. Schmid, Semi-automated screening of biomedical citations for systematic reviews, BMC Bioinformatics 11 (2010) 1–11.
B. C. Wallace, K. Small, C. E. Brodley, J. Lau, T. A. Trikalinos, Deploying an interactive machine learning system in an evidence-based practice center: abstrackr, in: Proceedings of the 2nd ACM SIGHIT international health informatics symposium, 2012, pp. 819–824.
J. Thomas, J. Brunton, S. Graziosi, Eppi-reviewer 4.0: software for research synthesis, EPPI-Centre Software. London: Social Science Research Unit, Institute of Education (2010).
B. E. Howard, J. Phillips, K. Miller, A. Tandon, D. Mav, M. R. Shah, S. Holmgren, K. E. Pelch, V. Walker, A. A. Rooney, et al., Swift-review: a text-mining workbench for systematic review, Systematic Reviews 5 (2016) 1–16.
G. Kontonatsios, S. Spencer, P. Matthew, I. Korkontzelos, Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews, Expert Systems with Applications: X 6 (2020) 100030.
D. D. Lewis, A sequential algorithm for training text classifiers: Corrigendum and additional data, in: Acm Sigir Forum, volume 29, ACM New York, NY, USA, 1995, pp. 13–19.
M. Miwa, J. Thomas, A. O'Mara-Eves, S. Ananiadou, Reducing systematic review workload through certainty-based screening, Journal of Biomedical Informatics 51 (2014) 242–253.
G. V. Cormack, M. R. Grossman, Scalability of continuous active learning for reliable high-recall text classification, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM'16), ACM, 2016, p. 1039–1048.
B. E. Howard, J. Phillips, A. Tandon, A. Maharana, R. Elmore, D. Mav, A. Sedykh, K. Thayer, B. A. Merrick, V. Walker, A. Rooney, R. R. Shah, Swift-active screener: Accelerated document screening through active learning and integrated recall estimation, Environment International 138 (2020) 105623.
R. van de Schoot, J. de Bruin, R. Schram, P. Zahedi, J. de Boer, F. Weijdema, B. Kramer, M. Huijts, M. Hoogerwerf, G. Ferdinands, A. Harkema, J. Willemsen, Y. Ma, Q. Fang, S. Hindriks, L. Tummers, D. L. Oberski, An open source machine learning framework for efficient and transparent systematic reviews, Nature Machine Intelligence 133 (2021) 125–133.
W. Luo, A. Schwing, R. Urtasun, Latent structured active learning, Advances in Neural Information Processing Systems 26 (2013).
S. Hwang, J. Choi, J. Choi, Uncertainty-based selective clustering for active learning, IEEE Access 10 (2022) 110983–110991.
P. Przybyła, A. J. Brockmeier, G. Kontonatsios, M.-A. Le Pogam, J. McNaught, E. von Elm, K. Nolan, S. Ananiadou, Prioritising references for systematic reviews with robotanalyst: a user study, Research Synthesis Methods 9 (2018) 470–488.
J. Kang, K. R. Ryu, H.-C. Kwon, Using cluster-based sampling to select initial training set for active learning in text classification, in: Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004. Proceedings 8, Springer, 2004, pp. 384–388.
J. Zhu, H. Wang, T. Yao, B. K. Tsou, Active learning with sampling by uncertainty and density for word sense disambiguation and text classification, in: Proceddings of the 22nd International Conference on Computational Linguistics (COLING'08), 2008, pp. 1137–1144.
W. Yuan, Y. Han, D. Guan, S. Lee, Y.-K. Lee, Initial training data selection for active learning, in: Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, 2011, pp. 1–7.
S. Xie, P. S. Yu, Active zero-shot learning: a novel approach to extreme multi-labeled classification, International Journal of Data Science and Analytics 3 (2017) 151–160.
Y. Wang, S. Chen, Z.-H. Zhou, New semi-supervised classification method based on modified cluster assumption, IEEE Transactions on Neural Networks and Learning Systems 23 (2012) 689–702.
X. Luo, F. Chen, D. Zhu, L. Wang, H. Liu, M. Lyu, Y. Wang, Q. Wang, Y. Chen, Potential roles of large language models in the production of systematic reviews and meta-analyses, Journal of Medical Internet Research 26 (2024) e56780.
E. Guo, M. Gupta, J. Deng, Y.-J. Parl, M. Page, C. Naugler, Automated paper screening for clinical reviews using large language models: Data analysis study, Journal of Medical Internet Research 26 (2024) e48996.
T. Oami, Y. Okada, T.-a. Nakada, Performance of a large language model in screening citations, JAMA Network Open 7 (2024) e2420496.
K. Matsui, T. Utsumi, Y. Aoki, T. Maruki, M. Takeshima, Y. Takaesu, Human-comparable sensitivity of large language models in identifying eligible studies through title and abstract screening: 3-layer strategy using gpt-3.5 and gpt-4 for systematic reviews, Journal of Medical Internet Research 26 (2024) e52758.
M. Li, J. Sun, X. Tan, Evaluating the effectiveness of large language models in abstract screening: a comparative analysis, Systematic Reviews 13 (2024) 219.

R. Sanghera, A. J. Thirunavukarasu, M. El Khoury, J. O'Logbon, Y. Chen, A. Watt, M. Mahmood, H. Butt, G. Nishimura, A. A. S. Soltan, High-performance automated abstract screening with large language model ensembles, Journal of the American Medical Informatics Association 32 (2025) 893–904.

F. Dennstädt, J. Zink, P. M. Putora, J. Hastings, N. Cihoric, Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain, Systematic Reviews 13 (2024) 158.

M. Issaiy, H. Ghanaati, S. Kolahi, M. Shakiba, A. H. Jalali, D. Zarei, S. Kazemian, M. A. Avanaki, K. Firouznia, Methodological insights into chatgpt's screening performance in systematic reviews, BMC Medical Research Methodology 24 (2024) 78.

O. Akinseloyin, X. Jiang, V. Palade, A question-answering framework for automated abstract screening using large language models, Journal of the American Medical Informatics Association 31 (2024) 1939–1952.

O. Akinseloyin, X. Jiang, V. Valade, A llm-based multi-agent collaborative approach for screening prioritization towards automated systematic reviews, medRxiv (2025).

N. Muthukumar, Few-shot learning text classification in federated environments, in: 2021 Smart Technologies, Communication and Robotics (STCR), IEEE, 2021, pp. 1–3.

E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, Clef 2019 technology assisted reviews in empirical medicine overview, in: CEUR Workshop Proceedings, volume 2380, 2019, p. 250.

B. C. Wallace, K. Small, C. E. Brodley, T. A. Trikalinos, Active learning for biomedical citation screening, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp. 173–182.

R. van Dinter, C. Catal, B. Tekinerdogan, A decision support system for automating document retrieval and citation screening, Expert Systems with Applications 182 (2021) 115261.

R. Ofori-Boateng, T. G. Trujillo-Escobar, M. Aceves-Martins, N. Wiratunga, C. F. Moreno-Garcia, Enhancing systematic reviews: An in-depth analysis on the impact of active learning parameter combinations for biomedical abstract screening, Artificial Intelligence in Medicine 157 (2024) 102989.

S. Wang, H. Scells, S. Zhuang, M. Potthast, B. Koopman, G. Zuccon, Zero-shot generative large language models for systematic review screening automation, in: Advances in Information Retrieval (ECIR'24), Springer Nature Switzerland, 2024, pp. 403–420.

C. Lefebvre, J. Glanville, S. Briscoe, R. Featherstone, A. Littlewood, M.-I. Metzendorf, A. Noel-Storr, R. Paynter, T. Rader, J. Thomas, L. S. Wieland, L., Cochrane, 2024.