Continuous Active Learning with Systematic Reviews in Medicine

Aaron HA Fletcher
School of Computer Science
Sheffield
ahafletcher1@sheffield.ac.uk

Acronym	Full Form
SR	Systematic Review
CAL	Continuous Active Learning
AL	Active Learning
TAR	Technology-Assisted Review
EBM	Evidence-Based Medicine
DTA	Diagnostic Test Accuracy
WSS	Work Saved over Sampling
TF-IDF	Term Frequency-Inverse Document Frequency
SVM	Support Vector Machine
BMI	Base Model Implementation
BERT	Bidirectional Encoder Representations from Transformers
LLM	Large Language Model
RCT	Randomised Controlled Trial
OCEBM	Oxford Centre for Evidence-Based Medicine
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
MeSH	Medical Subject Headings
SAL	Simple Active Learning
SPL	Simple Passive Learning

Table 1: List of Acronyms in Systematic Review Research

Abstract − Continuous Active Learning (CAL) has emerged as a promising technique to enhance the efficiency and accuracy of systematic reviews in medicine. This PhD proposal investigates the application of CAL, specifically focusing on the title and abstract screening substage of systematic reviews. The primary goal is to minimize expert intervention while maintaining high accuracy in document classification, thereby addressing the increasing volume of research and limited resources in healthcare. The PhD research will focus exclusively on a pool-based sampling approach within active learning. The PhD will use datasets such as CLEF-TAR and the Synergy Dataset, which provide real-world scenarios and imbalances typical of systematic reviews. Research gaps to addressed are the lack of long context model integration, such as longformer and big bert into CAL, the limited use of metadata in the title and abstract screening process and utilising encoders as document representations.

Keywords - Systematic Reviews, Continous Active Learning, Technology-Assisted Review, Medical Literature Screening, Evidence-based medicine, BERT, Metadata analysis

Contents

I INTRODUCTION

This document presents a literature review and planned research questions for the author's Ph.D. proposal. The Ph.D. proposal focusses on enhancing the performance of title and abstract selection through the application of continuous active learning in systematic reviews.

The proposal starts by motivating the need for research in this area, highlighting key stages of the systematic review process and the challenges faced. Then it outlines data sets and existing previous work with commentary on what the author feels are limitations of this research body.

The core research questions centre around how documents are represented within the CAL process and are as follows:

1. Introduce novel stopping algorithms for TAR
2. Investigate whether additional metadata, such as citation networks, can enhance model performance.
3. Evaluating the potential benefits of using decoders as embedders rather than encoders in CAL performance.

It also outlines a timeline, potential risks and mitigation strategies, ethical considerations and how the requirements for a professional development plan have been met. The author also mentions other publishable research projects with which he is involved and their potential impact on the Ph.D.

II BACKGROUND LITERATURE

A Systematic Reviews

A systematic review (SR) is one approach, among many, to provide evidence that can be used to support clinical decisions [1]. Evidence-based medicine attempts to incorporate this into clinical practice, by recommending the preferential use of the strongest available evidence in guiding decision making. EBM ranks each approach to support decision making, sometimes referred to as a hierarchy of evidence, where SRs are deemed to provide the strongest evidence to support any clinical decision - for relative rankings of evidence strength, see Table 2.

Level	Type of Evidence
1a	Systematic reviews of randomized controlled trials
1b	Individual randomized controlled trials
2a	Systematic reviews of cohort studies
2b	Individual cohort studies
3a	Systematic reviews of case-control studies
3b	Individual case-control studies
4	Case series
5	Expert opinion

Table 2: A summarised form of the 2009 OCEBM Levels of Evidence [2] There are conflicting thoughts on the absolute ranking of strength for all evidence sources [3, 4]; however, a key commonality is that systematic reviews (of RCTs) are considered the strongest evidence type.

SRs use reproducible systematic methodology to collect existing research, critically assess each study, and synthesise the findings into new research, and aim to provide a complete exhaustive summary of current evidence related to the research question [5].

The need to improve efficiency in SRs is based on two main areas: the increasing volume of research and the resources available within healthcare care. It is known that the amount of research available to be included in these SRs is increasing; with an estimated number of peer-reviewed journals in 2020 being 46,739 (from 14, 694 in 2001), and the total number of articles published increasing threefold, and the total number of clinical research trials increasing twofold - see Figure 2 [6]. However, the merits and disadvantages of the SR process are outside the scope of this Ph.D., but it has been succinctly conveyed in the existing literature [7], and, notwithstanding, SRs represent the best approach available to providing EBM.

A.1 The systematic review process

To understand how we might aim to improve SRs, we first need to outline the stages through which a SR progresses. Typically, the process is broken down into five distinct phases, as outlined in Table 3. My PhD will focus on stage 2: Identifying relevant work, which can be further granularised into several substages:

- Inclusion/exclusion criteria generation

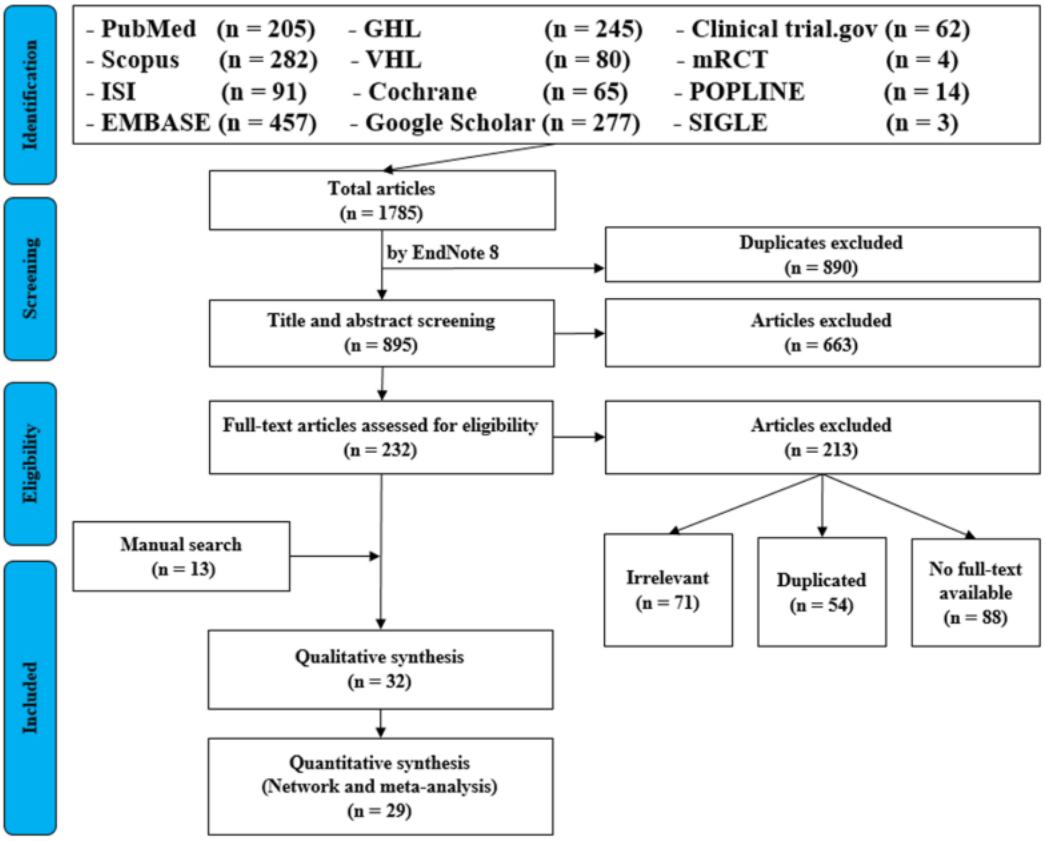


Figure 1: PRISMA flow diagram of studies' selection and screening process: Copied from [8]

- Search strategy development
- Database searching
- Protocol writing
- Title and abstract screening
- Full-text download and screening
- Manual search

Figure 1 illustrates these substages. Specifically, this PhD will concentrate solely on the title and abstract screening substage of the "identifying relevant work" phase. At this point in the identification of works process, preliminary work has been identified through a Boolean search, providing a large list of potentially included research. Traditionally, the titles and abstracts of these works are then manually evaluated by 2-3 reviewers to decide whether they should be included or excluded based on predetermined criteria, reducing the additional work that occurs within the full text download and screening substage. This process is similar to information retrieval.

A.2 Efficiency within the title and abstract Screening process

Abstract screening averages 0.13-2.88 abstracts per minute [**felizardo˙visual˙2013**, 9, 10]. Conflict resolution, which is often necessary when multiple reviewers are used (which is preferred), takes on average 5 minutes. Screening

Stage	Purpose
1	**Framing questions for a review:** The research question is structured and explicitly formulated.
2	**Identifying relevant work:** A wide range of databases are searched to identify research to be included. Potential research is first identified, screened, eligibility checked, and then a decision is made on the inclusion of that research [8].
3	**Assessing the quality of studies:** Research is tested for quality, such as minimum research design, and subjected to higher quality assessment checks, including tests for research heterogeneity.
4	**Summarizing the evidence:** Data synthesis occurs with tabulation of study characteristics and quality. Statistical testing is performed at this stage.
5	**Interpreting the findings:** Any issues highlighted in the previous steps should be addressed. Generate recommendations guided by reference to the strength of the evidence.

Table 3: Stages of a Systematic Review

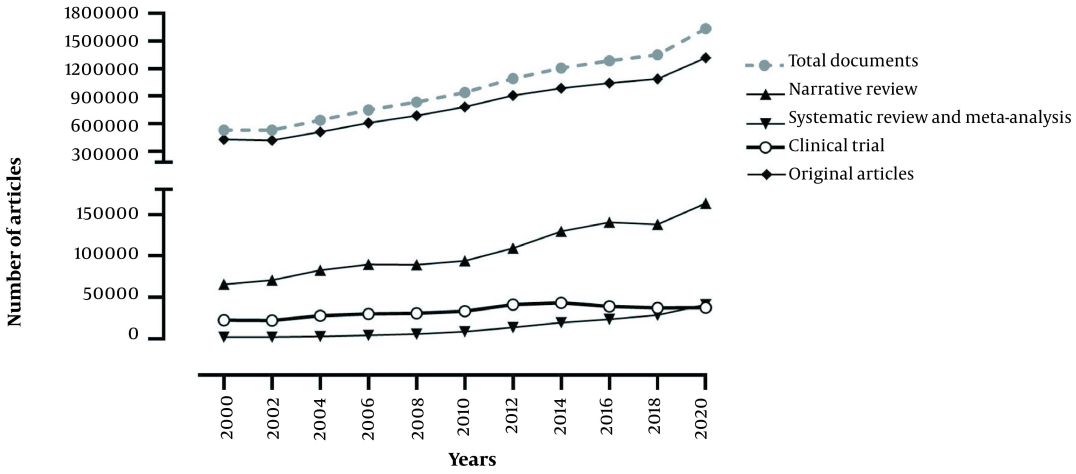


Figure 2: Increasing publications over that past two decades [6]

a full-text article takes 4 minutes on average [9]. Given a recently released Cochrane SR on the use of preoperative statin therapy in adults undergoing cardiac surgery, we can estimate that the total time to review all the text and abstracts for this would take 7.43 hours at best and 164.64 hours at worst [11]. Factored into an inflation adjusted average research cost per minute (£1.598), expected costs **for just this substage alone** could be expected to be £721.97 to £15,785.68 [12].

B Active Learning

Active Learning (AL) presents a promising approach to address one of the most resource-intensive stages of the SR: title and abstract screening. In the context of SRs, where the volume of potentially relevant literature is vast and growing, AL offers a method to significantly reduce the manual workload while maintaining high accuracy in document selection. Traditional systematic review methods require domain experts to manually screen all titles and abstracts identified in the initial search phase. This process is time-consuming and costly, especially given the increasing volume of published research. AL aims to optimise this process by intelligently selecting which documents should be reviewed by human experts, potentially saving significant time and resources. By applying AL techniques to the screening process, we can:

- Prioritise potentially relevant documents for expert review
- Reduce the overall number of documents that require manual screening
- Potentially identify relevant documents that might be missed in manual screening due to human fatigue or error
- Accelerate the overall SR process without compromising on quality

Deep-learning models have traditionally relied on large-labelled datasets for training. However, this approach contrasts sharply with real-world scenarios, particularly in specialised domains such as medicine. Although data collection is relatively straightforward in these fields, labelling is often time-consuming and requires expert knowledge [13, 14]. This disparity presents a significant challenge to optimise model performance with a limited number of labelled examples. This challenge is particularly relevant to the screening process in SRs, where we currently ask experts to screen all titles and abstracts returned from the identification phase. However, we would want to move to a scenario where minimal expert screening is sought from research returned from the identification phase, whose screening can then be safely extrapolated to a larger pool.

Active learning (AL) studies how to do just this. Through AL terms, it attempts to use a sampling policy π to select samples $\mathbf{T}C, i$ from an unlabelled dataset $\mathbf{T}U, i$ and pass them to an oracle for labelling and added to a known dataset $\mathbf{T}_{K,i}$. Technology-Assisted Review (TAR), also known as Computer-Assisted Review or Predictive Coding, is a process that uses machine learning to assist in document review tasks. In the context of SRs, the TAR applies AL principles to streamline the selection process of titles and abstracts to screen. By iteratively training a machine learning model on human-labelled examples, TAR can prioritise potentially relevant documents for expert review, significantly reducing manual workload and, in some cases, exceeding human ability [15]. This approach aligns closely with the goals of AL in SRs, as it aims to maximise the efficiency of expert input while maintaining high accuracy in document classification.

Different approaches can be taken to AL, such as membership query synthesis [16], stream-based selective [17] and pool-based sampling [18]. These approaches are divided on how much of the unlabelled dataset ($\mathbf{T}U, i$) that a model has access to when utilising a policy π for the selection of data points to be labelled $\mathbf{T}C, i$. In pool-based sampling, the entire unlabelled dataset, $\mathbf{T}U, i$, is evaluated in the selection of $\mathbf{T}C, i$, in stream-based selective sampling, datapoints are evaluated one at a time, and in membership query synthesis, synthetic data are generated from an underlying natural distribution. As we are concerned solely with the subprocess of title and abstract screening, it aligns strongly with pool-based sampling as the entire unlabelled dataset is known ahead of time and will be the approach used within this PhD. Figure 3 outlines the AL cycle with pool-based querying.

The oracle (O) within this PhD will denote a human-verified label resulting from screening potential research for inclusion within the SR within the title and abstract screening stage. Domain experts typically perform this. More

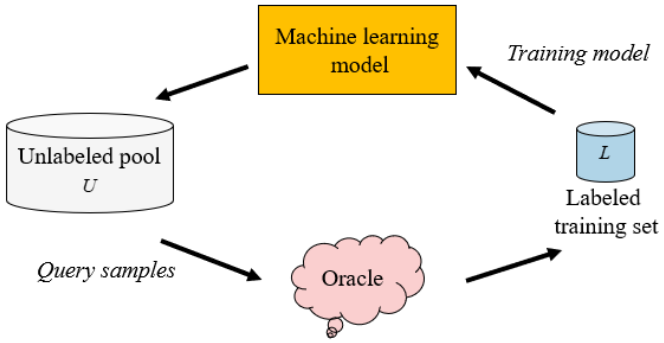


Figure 3: Overview of a pool-based query strategy for AL, replicated[19]

.

Notation	Explanation	Notes
$\mathbf{T}$	Total dataset	e.g. Research gathered after Identification phase of the selection process.
i	Iteration	A single cycle within the active learning process.
$\mathbf{T}_{K,i}$	Known datapoints per iteration	e.g. research that has been screened by a reviewer
$\mathbf{T}_{U,i}$	Unknown datapoints per iteration	e.g. research that has not been screened by a reviewer
$\mathbf{T}_{C,i}$	A subset of $\mathbf{T}_{U,i}$ to be labelled	chosen by a policy, datapoints to be screened by a reviewer.
π	Policy	How $\mathbf{T}_{U,i}$ is selected, e.g. uncertainty, random, certainty, diversity sampling
O	Oracle	Often a domain expert, who assigns labels to unscreened research.
T_R	Total Relevant Documents	All research that should be included in a systematic review.
T_{IR}	Total Irrelevant Documents	All research that should not be included in a systematic review.

Table 4: Notation used for active learning within this review

concretely, O can be considered a function $O(x) = y$ where X is a representation, such as embedding the research title and abstract, and y is the assigned category (included or excluded). We assumed that for each datapoint (x), O provides a single judgement (y), which is always correct and do not concern ourselves with any potential intercoder agreement or bias within that decision process [20].

The broader literature has evaluated the effects of the choice of π. Traditionally, π used the sigmoid response of the final layer of a model as a proxy of confidence, which is not a reliable measure, as these responses tend to be overly confident [21]. The use of the softmax response has been shown, in some cases, to be worse than random sampling [22]. However, the effect of the choice of π in combination with the final output layer in this specific domain (i.e. the SR process) is unknown, and this remains an active area of research which will not be covered during the PhD. The author uses techniques such as temperature scaling to effectively combat overly confident soft-max responses [23].

In the author's mind, it is unclear the exact difference between continuous/online active learning (CAL) and AL, and indeed, it seems that much of the current literature refers to CAL when, in fact, it means AL. Some authors refer to eliminating models between iterations and the process occurring in discrete rounds as AL [24]. Cormack differentiates the two based on objectives, with the aim of CAL being to find and review as many of the responsive documents as possible, as quickly as possible, and AL is to produce the best classier possible, considering the level of training effort (which are subtlely different objectives) [25]. For this research, we will use continuous to denote the incremental streaming of newly available information to any model.

B.1 Active Learning Key Literature

There exists a plethora of research within the AL area; however, due to the specific focus on medicine within this PhD, I will focus on existing literature within the medical domain and some key ones from others. It is valid to delineate between research within differing domains within AL (e.g., e-discovery in the Legal Domain, sentiment

analysis on social media, or image classification in computer vision) as each domain presents unique challenges and characteristics that influence the application of AL techniques. In the medical domain, particularly in SRs, AL must contend with highly specialised vocabulary, complex interrelationships between concepts, and the critical importance of high recall to ensure that relevant studies are not missed. The emphasis of the medical domain on evidence-based practice and the potential impact on patient care requires a more stringent approach to AL. Unlike other domains, where missing a small percentage of relevant items might be acceptable in systematic medicine reviews, overlooking a crucial study could have significant consequences. This requirement for near-perfect recall and the need to process large volumes of literature efficiently create a unique set of demands for medical TAR AL algorithms. This literature review will assess each work for their respective contributions, the datasets used, the evaluation metrics (and scores achieved), the models used, and the representation of data points used in each approach.

An early contribution to this field demonstrated that automated classification of document citations can be used to reduce the time reviewers spend screening evidence for inclusion in SRS of drug class efficacy [26]. The researchers used a novel data set created from annotated reference files from 15 SRs of drug classes. The features were extracted from the research articles using the "bag-of-words" approach for the title and abstract, the MeSH terms, and the MEDLINE publication type. The features were one-hot encoded and selected using the chi-square test to drop insignificant features. Finally, this input was used to train a perceptron model and was evaluated using precision, recall, and the F measure in a range of sample weighting and the WSS@95%. This work's significant contribution was using machine learning approaches to address this screening issue.

A significant contribution to the field came from a simulation study that mimics the process of a human reviewer screening records while interacting with an active learning model [27]. This work used six previously labelled SRs. It looked at four classification techniques (naive Bayes, logistic regression, support vector machines, and random forest). Two feature extraction strategies (TF-IDF and doc2vec) were evaluated based on the work saved on sampling and recall. The title and abstracts were used to generate these inputs. It showed that in a simulated approach, the models reduced the number of publications screened from 91.7 to 63. 9% (WSS@95). The naive Bayes and TF-IDF models yielded the best overall results in this study. This study is limited due to the smaller datasets used and the feature extraction approaches used (which, for the year of publication, other potential superior choices, such as contextual embeddings, could have been explored). It introduced some new evaluation metrics, such as TTD and ATD. This research only superficially evaluates the variability of these approaches in SRs, reporting the range of WSS@95 and not attempting to consider factors within SRs that may have led to this variability.

More recent work reported a protocol denoted "CAL" (Continuous Active Learning), initially performed on legal datasets [28]. The process involves selecting an initial set of seed documents, typically using keyword search, which are then reviewed and coded. This training set is used to train a model that scores each document based on its response (relevant) likelihood. The top-scoring documents that have not been coded are then reviewed and coded. The extended training set is then used to retrain the model, and this process continues until "enough" of the responsive documents are found. The key difference between this approach and previous ones, such as SAL (Simple Active Learning) and SPL (Simple Passive Learning), is their selection strategy: CAL uses relevance feedback (selecting highest-scoring documents), SAL uses uncertainty sampling, and SPL uses random selection. CAL achieved better results on the recall@75 The study primarily used SVM (Sofia-ML implementation of Pegasos SVM) for all protocols. It mentions briefly that it replicated most experiments using logistic regression, achieving similar results to SVM. It also tested with Nave Bayes, which achieved generally inferior results overall but maintained the same relative effectiveness among the protocols. This paper is essential to outline the CAL process and demonstrate its effectiveness. However, it had some limitations: It used a fixed batch size of 1,000 documents for efficiency, though they noted slightly better results with a batch size of 100 for CAL. Although multiple classifiers were tested, detailed

performance comparisons between models were not reported. The study did not explore extensively the effects of feature engineering methods. The human factors in review accuracy were not fully addressed in the simulation. Despite these limitations, the article provided a strong foundation for understanding and further investigating TAR protocols in SR TAR.

The follow-up was to introduce the autonomous TAR process (Auto TAR), which showed better performance than CAL [25]. The algorithm is outlined in Figure 4. The AUTO TAR process differs from CAL through:

- AUTO TAR uses a single seed document, and CAL uses a 1,000 document set.
- AUTO TAR uses word-based features TF-IDF, and CAL uses binary byte-4 grammes.
- AUTO TAR exponentially increases batch sizes, starting with one and increasing by 10% each iteration; CAL's batch size is fixed at 1,000

Finally, this approach was augmented to use BM25 (a saturated form of TF-IDF) + logistic regression and has been considered state-of-the-art for the past eight years. This is often referred to as "base model implementation", or abbreviated to BMI.

Since 2016's AutoTar approach, the transformer architecture has advanced virtually all fields within natural language processing, permeating almost every aspect of the field [30]. However, the TAR process has remained surprisingly resistant to these advances. Although a comprehensive analysis of transformer-based improvements is beyond the scope of this literature review, their primary advantage is their ability to provide a nuanced contextual understanding of written texts. This contextual comprehension differs significantly from traditional feature extraction techniques such as TF-IDF. Transformers employ self-attention mechanisms to learn context-dependent text representations, enabling them to capture subtle semantic relationships and long-range dependencies not present with TF-IDF-like approaches, in the TAR process.

Add in about CALBERT

Subsequent research using decoder architectures within CAL attempted to understand why their performance was underwhelming compared to BMI. Goldilocks [31] took a pre-trained BERT model and fine-tuned it first on the unlabelled corpus (0-10 epochs were tested). They then randomly selected a positive example seed, and on each iteration of the AL process 200 documents were sampled using either relevance feedback or uncertainty sampling. During each iteration, after labelling, all labelled documents were used to fine-tune BERT again for 20 epochs, using the previous model from the previous iteration as the starting point. The input was the concatenated title and body text, truncated to 512 tokens. The model was then used to classify a document's relevance. They found that 5-epoch fine-tuning on the unlabelled dataset was "just right", and achieved similar performance to that of TF-IDF & logistic regression within domain dataset (RCV1-V2 corpus), and statistically significantly worse on out-of-domain dataset(Jeb Bush corpus). Although not discussed in the paper, this could be explained through the phenomenon of "catastrophic forgetting", where too much fine-tuning on the task corpus might cause the model to forget useful general knowledge from pertaining [32]. Additionally, they used the model from the previous iteration as a starting point for new classification fine-tuning iterations, which could have further compounded this.

However, the concept of a "goldilocks" epoch for CAL within medical data sets did not seem to hold, with subsequent research demonstrating that a "just right retaining epoch" was not present in the CLEF data set [33].

Issues with Goldilocks:

1. Contrived experimental design.
2. Sensitive to pre-trained model choice.

Background on stopping algorithms

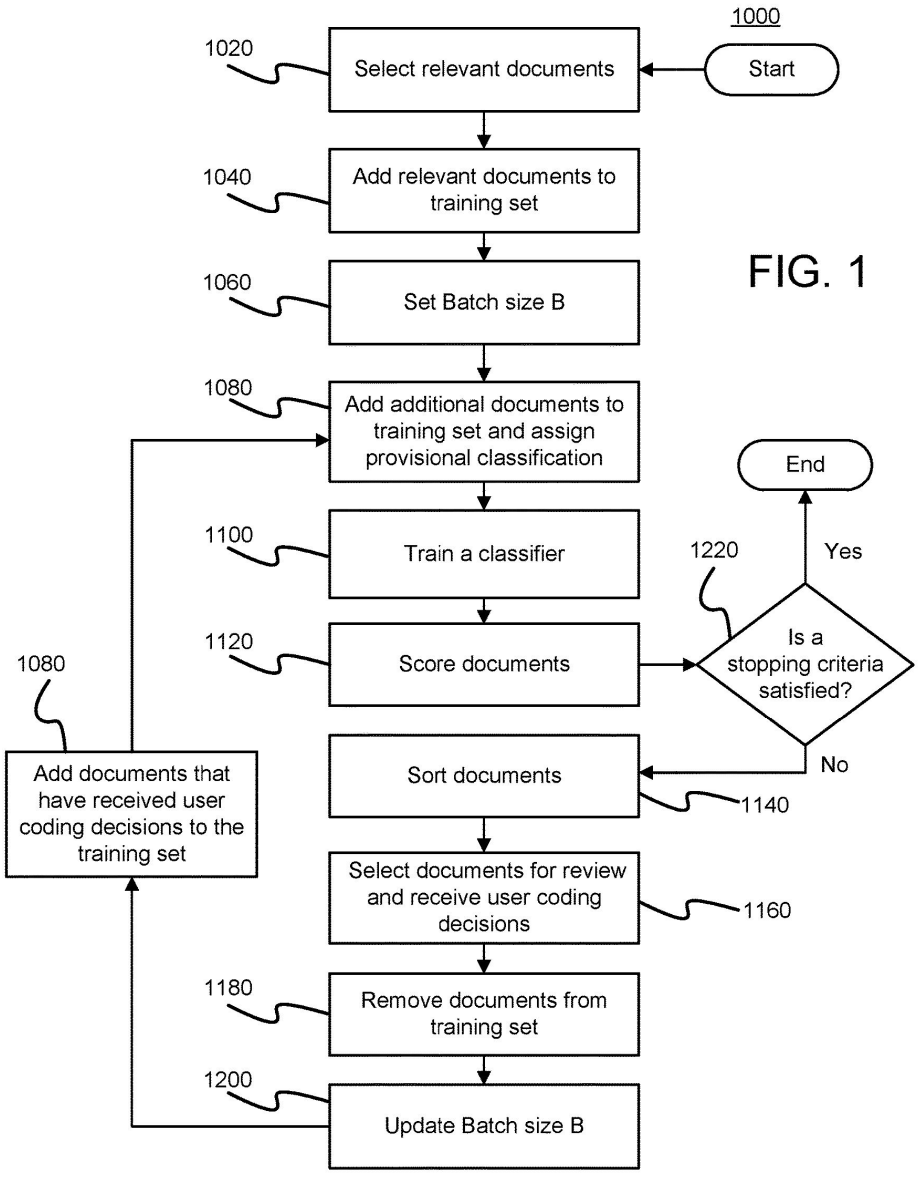


FIG. 1

Figure 4: Auto TAR outline, as documented within the patient filing by Cormack et al. [29]

Dataset	Total SRs	Type(s) of SR	T	TR	TR/T
CLEF 2017	50	DTA	269628	4661	0.017
CLEF 2018	50	DTA	266657	4351	0.016
CLEF 2019	80	DTA	485153	8315	0.017
CLEF 2019	80	Intervention	31644	448	0.014
Synergy	26	Not Applicable	169288	2834	0.017

Table 5: Training Dataset sizes for the TAR datasets

C Datasets

Numerous data sets related to this area have been used in the existing literature.

C.1 CLEF-TAR (2017, 2018, 2019)

CLEF-TAR is a dataset that was released as part of CLEF eTASK 2, and is available on github[1] [34, 35, 36]. Originally designed with document ranking as the primary focus, the information contained within the data set allows for the subprocess simulation of the title and abstract selection of the SR procedure, using published real-world Cochrane SRs. Each year, this data set was incrementally updated and Table 5 outlines the scope of the issue and succinctly highlights the presence of a large imbalance of the TIR class. Diagnostic test accuracy SRs (DTA) summarise a test accuracy, while intervention reviews assess the effectiveness/safety of a treatment, vaccine, device, preventative measure, procedure, or policy. Delineation between the types of SRs is not required for research within this Ph.d.

Of importance, the CLEF dataset did not provide the titles or abstracts for each research found in the Identification Phase, rather relying on the users to download them for experimentation. This is an important oversight of the data set as titles and abstracts can be updated or retracted post-publication, meaning, fair comparison across time might become increasingly challenging. Within this Ph.D. I intend to use this recently collected source 2024 of titles/abstracts that have been collected as part of other work in this area, which has extracted all titles and abstracts for **T** within the CLEF dataset [37][2].

C.2 Synergy Dataset

The Synergy dataset [38], while less frequently used in the literature, offers a more contemporary collection of SRs[3]. This data set comprises 26 SRs that span multiple domains, with a predominant focus on the medical field (20 out of 26 reviews). Reviews included in this data set range from 2002 to 2020, potentially providing more recent information compared to the CLEF data set. The Synergy data set features diverse domains, allowing cross-domain analysis despite its primary focus on medical reviews. It also includes an expanded variable set. In addition to the basic information found in the CLEF dataset, Synergy incorporates authorship details, referenced works, and publication years, all sourced from the OpenAlex API. Due to its more recent compilation and limited use in existing research, this dataset could be used to externally validate pre-trained language models.The inclusion of SRs from nonmedical domains, such as computer science, allows evaluations on the transferability of TAR approaches across different fields. Synergy's TR/T ratio of 0.017 is consistent with the class imbalance observed in the CLEF datasets, making it suitable for comparative studies and model evaluation in the context of title and abstract selection tasks.

[1] https://github.com/CLEF-TAR/tar
[2] https://github.com/ielab/goldilocks-reproduce
[3] https://github.com/asreview/synergy-dataset/tree/master

C.3 TREC Total Recall Track Dataset (2015, 2016)

The TREC Total Recall Track produced data sets specifically designed for high-recall retrieval tasks, similar to those encountered in SRs[4][39, 40]. This data set simulates scenarios where the goal is to find all or nearly all relevant documents in a collection, which aligns closely with the objectives of the title and abstract screening phase in SRs. The data set includes a corpus of documents, topics (which can be seen as analogous to research questions in SRs), and relevance judgments.

C.4 Jeb Bush Emails Dataset

The Jeb Bush Emails dataset is an unconventional choice for TAR research, originally consisting of emails released by former Florida Governor Jeb Bush[5]. This data set is suitable for TAR experiments because of its large size and the presence of both relevant and irrelevant documents. Although not directly related to SRs, it provides a real-world corpus that can be used to simulate document classification tasks inherent in the SR process.

C.5 RCV1-v2 Dataset

The RCV1-v2 (Reuters Corpus Volume 1, Version 2) is a large, manually categorised newswire data set [41] that was published by Reuters between August 20, 1996, and August 19, 1997[6]. The dataset features 804,414 documents with multi-label classification across 103 topic categories, organised in a hierarchy. The documents are provided in XML format with rich metadata and the content is primarily English news stories covering a wide range of topics. Although not originally designed for SRs, RCV1-v2 has been used in various text classification and information retrieval tasks. In the context of SRs and TAR, the use of the RCV1-v2 data set lies in the simulation of approaches on a large-scale dataset to test the scalability and efficiency of screening algorithms and to evaluate any potential transferability of the approaches.

RCV1-v2 dataset is adapted for use in AL by denoting all documents as T, T_R as all documents having a specific label, and those without it, as by treating the entire corpus as T, T_{IR}, we can approximate the binary classification challenge of title and abstract Screening within SRs.

D Evaluation Metrics

Evaluation metrics for SR TAR process can be categorised between assessing how well a classifier minimised the relevant documents excluded by the classifier with a set work budget (i.e. effectiveness) or the reduction in the reviewer's workload by excluding the maximum number of irrelevant documents while maintaining recall (efficiency). The majority of the research produced within this will focus on improving the effectiveness of AL models within the medical TAR domain. The author chooses not to optimise the computational efficiency between approaches, rather to improve the final result achieved. This is for numerous reasons; however, the main two are that as computer processing increases, these practical limitation concerns become less and improvement in effectiveness will have a greater impact on SR usefulness than maximising efficiency. The author aims to report the time taken to run the algorithms, time complexity and the hardware that ran upon, so that comparison to time taken by humans undertaken can occur.

[4]https://trec.nist.gov/data/total-recall/
[5]https://ab21www.s3.amazonaws.com/JebBushEmails-Text.7z
[6]https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/datasets/_rcv1.py

D.1 Recall@k

In the context of SRs, achieving high recall is more critical than high precision. Recall represents the proportion of relevant documents correctly identified among all truly relevant documents [42]. This focus on recall may seem counterintuitive, but it is crucial for two reasons. First, each missed document could potentially contain significant information for the SR. Second, the initial screening is followed by a more precise full-text review (as outlined in the PRISMA workflow, Figure 1), where precision is emphasised. Although maximising recall is important, it is not practical to aim for 100% due to diminishing returns. As recall approaches higher levels, the computational cost of screening additional documents increases substantially, often yielding minimal benefit. To balance effectiveness and efficiency, researchers of TAR for SR commonly use and consider recall @ 95% as useful (that is, k = 0.95). This measure indicates the recovery achieved when 95% relevant documents are recovered, striking a pragmatic balance between comprehensive coverage and resource use. A higher recall@k is considered a more effective approach. Recall is calculated via:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

Recall@95% is calculated via:

$$\text{recall@95\%} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Where recall is calculated once an AL classifier achieves 95% TP.

A small point on nomenclature: Historically, recall has been referred to in the medical literature as sensitivity. These are two different terms for the same metric, and the use of either term depends on the domain. Additionally, in the legal domain, there might be references to recall@75% which is not as useful for the medical domain. Legal domains often prioritise based on cost-effectiveness, time-constraints and proportionability, which medical reviews require as close to absolute information as possible [43].

D.2 R-Precision

This effectiveness metric determines, given the T_R, what proportion of documents returned by the approach within the total number of relevant documents were actually relevant [44]. The best score for R-precision is 1 (i.e., all relevant documents were returned in the top T_R position). It allows for an adaptive cutoff for T_R, which adapts to the SR, and also considers precision. Note that this evaluation metric can only be used when the T_R for a query is known. It is calculated via:

$$\text{R-Precision} = \frac{\text{Relevant Documents in top } T_R}{T_R} \tag{3}$$

D.3 Work Saved Over Sampling (WSS@k)

WSS@k is an efficiency metric that would be valuable to report on to enable other researchers in the field to compare their approaches to mine and, if appropriate, improve upon. Again, k (recall) is typically set to 0.95. This metric evaluates the work saved over random sampling, with a higher WSS@k being more efficient and is calculated via:

$$\text{WSS} = \frac{\text{TN} + \text{FN}}{\text{T} - (1 - \text{Recall})} \tag{4}$$

This can also be expressed as:

$$\mathrm{WSS} = \frac{\mathrm{TN} + \mathrm{FN}}{\mathrm{T} - 1 + \frac{\mathrm{TP}}{\mathrm{TP+FN}}} \tag{5}$$

E Leveraging Citation Networks for Medical TAR

Systematic reviews utilise research evidence to provide clinical practice recommendations. The communication of medical research follows standardised formatting conventions and primarily occurs through peer-reviewed publication [45]. When authors compose research papers, they must reference related works to substantiate their claims and situate their findings within the existing body of knowledge. These citations follow standardised formatting guidelines and are documented in the paper's reference section. This rigorous documentation of citations enables analysis of the relationships between research papers, operating under the assumption that studies that cite or are cited by a research article are relevant to that research.

E.1 Relation analysis improves CAL TAR performance

Recent advances in medical CAL TAR have indirectly demonstrated the benefit of relationship analysis for citations. The current leading encoder model, $BioLinkBERT_{base}$ achieved state-of-the-art performance on the CLEF dataset in a CAL setting by leveraging citations networks between research papers [46, 37].

The $LinkBERT$ approach was to view a pertaining corpus as a graph of documents, with each document being a vertex and hyperlinks forming edges between documents. These related documents were then placed within the same context window. The approach differs from traditional $BERT$ architectures, which randomly allocate documents to context windows without considering their relationships. While this might appear similar to curriculum learning approachings, $LinkBERT$ is distinct in that it does not organise context windows by difficulty level.

$BioLinkBERT$, a domain-specific adaption of $LinkBERT$, was developed specifically for biomedical aplications and pretrained exclusively on PubMed articles, using citation relationships to estimate document relationships [7]. The model trianing process incorporated standard masked language modelling and next-sentence prediction techniques. Analysis of both the base model (100M parameters) and large model (340M parameters) against $PubMedBERT$ across multiple benchmarks: BLURB[47], MedQA-USMLE[48], and MMLU-professional medicine[49]. The results demonstrated $BioLinkBERT_{large}$'s superior performance across all evaluated benchmarks, notably achieving a 3.2% improvement over PubMedBERT in the BLURB score.

Current research on document relationship-based encoders in the CAL process has not definitively established that document relations are the primary driver of performance improvements. Furthermore, the assumption that larger models consistently yield better results is not always the case. The author replicated the previously reported Goldilock Reproduce study, where $BioLinkBERT_{base}$ formed the classifier, except changing the model to the $BioBERT_{large}$ variant[8] as a classifier model, yet only achieved higher performance in R-Precision in 7 of 12 datasets/policy combinations. The empirical results, detailed in Table 7, show peak R-precision values of 0.847 for the relevancy selection policy (at FPT epoch 2) and 0.832 for uncertainty selection (at FPT epoch 1).Statistical analysis using the Friedman test revealed significant differences between Further Pre-Training (FPT) epochs in only 4 of 12 datasets when examined individually. More importantly, when analyzing all datasets collectively, no statistically significant differences emerged in R-precision values across FPT epochs for either relevancy selection or uncertainty selection policies. This findiing challenges the previously documented "Goldilocks problem" observed in non-medical domains. Specifically demonstrating that FPT does not yield statistically significant improvemnts in R-Precision.

This replication study has generated valuable insights for this PhD investigation. A significant finding indicates that seeking an optimal pretraining epoch within the CLEF dataset is unlikely to be productive for future research endeavors. The experimental design revealed several methodological considerations, particularly regarding the implementation of hyperparameters without robust empirical justification. These include the selection of a batch size of

[7] https://huggingface.co/michiyasunaga/BioLinkBERT-base
[8] https://huggingface.co/michiyasunaga/BioLinkBERT-large

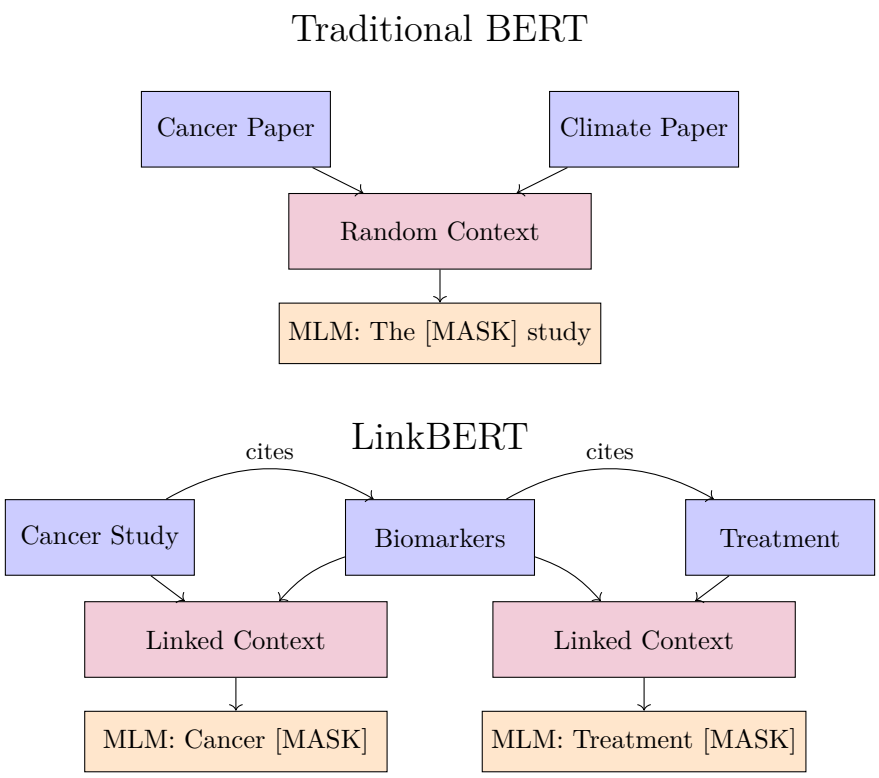


Figure 5: Comparison of document processing in traditional BERT versus LinkBERT. Traditional BERT (top) randomly groups documents into context windows, while LinkBERT (bottom) uses citation relationships to create meaningful document groupings for pretraining. The citation-based grouping ensures that semantically related documents are processed together during masked language modeling tasks.

25, the decision to fine-tune for 20 epochs, and the termination criterion of 501 labeled documents. These parameter choices, while functional, may impose limitations on potential improvements to encoder CAL process performance within the experimental framework.

The significance of these limitations becomes particularly evident when considering that observed R-Precision values approach the theoretical maximum of 1.0, with some instances achieving values as high as 0.945. In the context of the Goldilocks reproduce paper, datasets showing lower performance metrics, such as the CLEF 2019 dataset (with R-precision values of 0.82 for relevancy and 0.791 for uncertainty), present additional analytical challenges. The utilization of Large Language Models (LLMs) introduces complexity in interpreting the underlying causes of reduced performance in these cases.

While exploring larger, more sophisticated models presents a potential avenue for improvement, this approach faces practical constraints. Given the limitations of High-Performance Computing resources and PhD time constraints, pursuing research dependent on the development and availability of superior LLMs may not be the most pragmatic direction.

A crucial observation emerged from this research regarding the relationship between early document classification and overall performance. In iterations where strong performance was ultimately achieved at iteration 20, a notably higher number of relevant documents were classified earlier in the CAL process. This finding aligns with theoretical expectations: a larger corpus of correctly classified documents early in the process provides a more robust foundation for subsequent classification decisions. This insight carries significant implications for the next phase of this PhD research, suggesting that enhancing document availability in the early stages of the active learning process could

substantially improve overall performance outcomes.

While *BioLinkBERT* represents a sophisticated approach that combines citation networks with contextual language understanding, this integration presents both advantages and limitations. The model's ability to capture complex semantic relationships between documents is valuable, but the contextual processing introduces potential inefficiencies. During pretraining, when linked documents are placed in the same context, the model must process all content within those documents—including sections that may be tangential or unrelated to the citing paper's specific reference. This contextual noise could potentially dilute the precision of the more direct relationships that citations inherently represent. In contrast, pure citation links directly capture intentional scholarly connections made by domain experts, providing a cleaner signal without the additional complexity of processing potentially irrelevant contextual information.

A fundamental question emerges from this research: Is contextual understanding of references truly necessary for effective CAL? Several factors suggest that citation networks alone might be sufficient and potentially superior. First, citations themselves represent a form of knowledge distillation, where domain experts have already identified meaningful relationships between documents. Second, analysing reference networks is computationally more efficient than processing full textual contexts. Third, citation network models tend to be more stable when updated, compared to contextual models. Fourth, the contextualization of citation networks may actually introduce noise into what would otherwise be clear citation signals.

E.2 Direct citation network mining within medicine research

Performant, simple, and robust approaches to citation network mining already exist within medicial research. Let G be a citation graph where:

- D_i represents a research article of interest as a vertex in G
- D_{ip} represents the set of articles referenced by D_i
- D_{if} represents the set of articles that reference D_i
- Both sets are subsets of G: $D_{ip}, D_{if} \subset G$
- $D_{ip} \cap D_{if} = \emptyset$, so searching both sets will provide different relevant articles

Relevancy is defined as a function $R : D \rightarrow [0, 1]$, where:

- 0 denotes no relevance
- 1 denotes maximum relevance
- For any set of documents D_{set}, relevancy is defined as $R(D_{set}) = R(d)|d \in D_{set}$

Two primary citation network mining approaches are defined:

- Backward citation searching (BCS): examining all articles in D_{ip}[50, 51]
- Forward citation searching (FCS): examining all articles in D_{if}*[9]

[9]FCS involves using a citation index to identify studies that cite a source study. A citation index is a database of scholarly articles and their citations, such as PubMed, Google Scholar, Scopus or OpenAlex

Collection	Dataset size	Model	R-Precision (↑)		Friedman (p)	
			Rel.	Unc.	Rel.	Unc.
Clef 2019 dta test	8	BiolinkBert-Base-ep0	**0.909**	**0.857**	—	
		BiolinkBert-Large-ep0	0.897	0.803		
		BiolinkBert-Large-ep1	0.827	0.832		
		BiolinkBert-Large-ep2	0.812	0.774	0.914	0.632
		BiolinkBert-Large-ep5	0.841	0.814		
		BiolinkBert-Large-ep10	0.881	0.846		
Clef 2017 test	30	BiolinkBert-Base-ep0	0.812	0.794	—	
		BiolinkBert-Large-ep0	0.828	0.797		
		BiolinkBert-Large-ep1	0.826	**0.827**		
		BiolinkBert-Large-ep2	**0.858**	0.804	**<0.05**	**<0.05**
		BiolinkBert-Large-ep5	0.827	0.777		
		BiolinkBert-Large-ep10	0.799	0.757		
Clef 2017 train	20	BiolinkBert-Base-ep0	**0.838**	0.761	—	
		BiolinkBert-Large-ep0	0.778	0.765		
		BiolinkBert-Large-ep1	0.808	0.789		
		BiolinkBert-Large-ep2	0.767	0.701	**<0.05**	0.28
		BiolinkBert-Large-ep5	0.816	0.786		
		BiolinkBert-Large-ep10	0.827	**0.796**		
Clef 2018 test	30	BiolinkBert-Base-ep0	0.794	0.780	—	
		BiolinkBert-Large-ep0	0.789	0.774		
		BiolinkBert-Large-ep1	**0.812**	0.790		
		BiolinkBert-Large-ep2	0.797	**0.791**	0.52	0.50
		BiolinkBert-Large-ep5	0.763	0.773		
		BiolinkBert-Large-ep10	0.763	0.769		
Clef 2019 DTA int. train	20	BiolinkBert-Base-ep0	0.939	0.923	—	
		BiolinkBert-Large-ep0	0.939	0.902		
		BiolinkBert-Large-ep1	0.941	0.935		
		BiolinkBert-Large-ep2	0.948	0.921	0.78	0.50
		BiolinkBert-Large-ep5	0.952	0.945		
		BiolinkBert-Large-ep10	**0.945**	**0.947**		
Clef 2019 DTA int. test	20	BiolinkBert-Base-ep0	**0.934**	**0.900**	—	
		BiolinkBert-Large-ep0	0.899	0.856		
		BiolinkBert-Large-ep1	0.904	0.840		
		BiolinkBert-Large-ep2	0.909	0.878	0.87	**<0.05**
		BiolinkBert-Large-ep5	0.882	0.835		
		BiolinkBert-Large-ep10	0.865	0.841		

Table 6: Performance comparison across different collections and models

Table 7: Average R-precision of each FPT epoch for CLEF dataset

Policy	ep0	ep1	ep2	ep5	ep10
Uncertainty	0.813	0.832	0.813	0.815	0.814
Relevancy	0.840	0.845	0.847	0.842	0.835

Backward and forward citation searching (BCS and FCS) are both straightforward and effective approaches that inherently respect the chronological relationships between research articles, as papers can only cite previously published work. The significance of these methods is demonstrated by their recommended use in Cochrane systematic reviews, particularly during the identification phase. A study of Cochrane reviews conducted between November 2016 and January 2017 found that 87% reported using BCS, while 9% utilized FCS [52].. The Cochrane Handbook explicitly mandates the use of BCS (criterion C30) in the search stage, though it makes no mention of FCS [53]. However, neither the use of BCS nor FCS is addressed in the Handbook's guidelines for the screening phase.

The application of Backward and Forward Citation Searching (BCS and FCS) within an active learning process represents an understudied area of research (see Figure 6 for search strategy details). To establish the novelty of this augmentation, several key distinctions must be clarified. While this PhD research focuses on the title and abstract screening phase of systematic review generation, BCS and FCS have traditionally been confined to the identification phase (as illustrated in Figure 1). Conventionally, title and abstract screening serves to reduce the workload for the more resource-intensive full-text screening phase. However, from a computational perspective, restricting the screening process to titles and abstracts is unnecessary, as the computational cost remains manageable when including full texts.

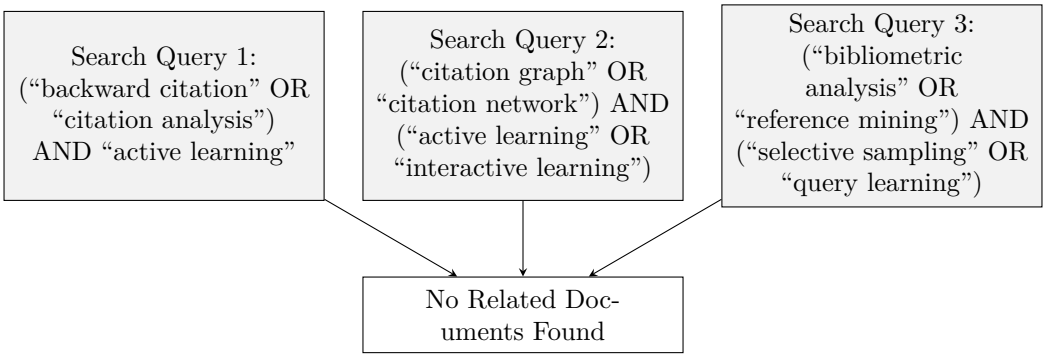
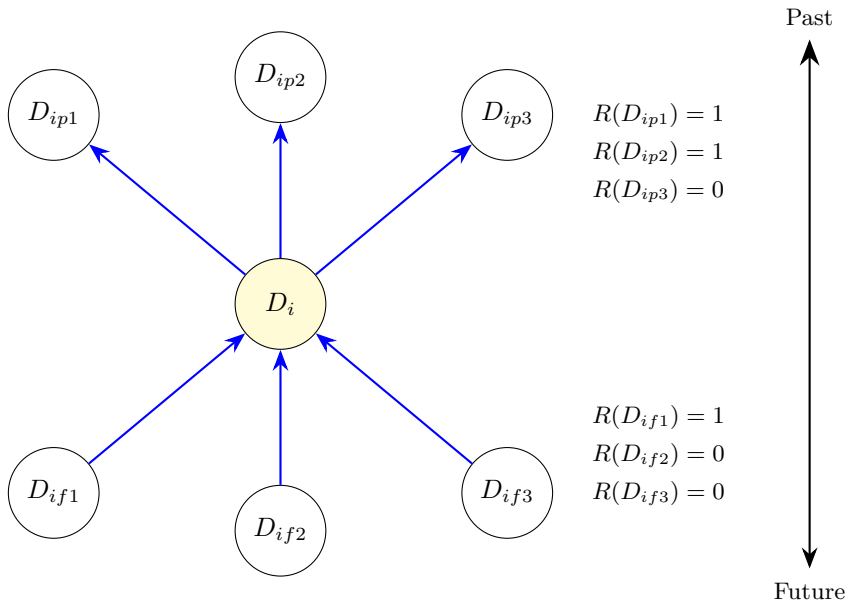


Figure 6: Results from literature search on citation index arxiv and pubmed demonstrating absence of related works, ran on 13th November 2024



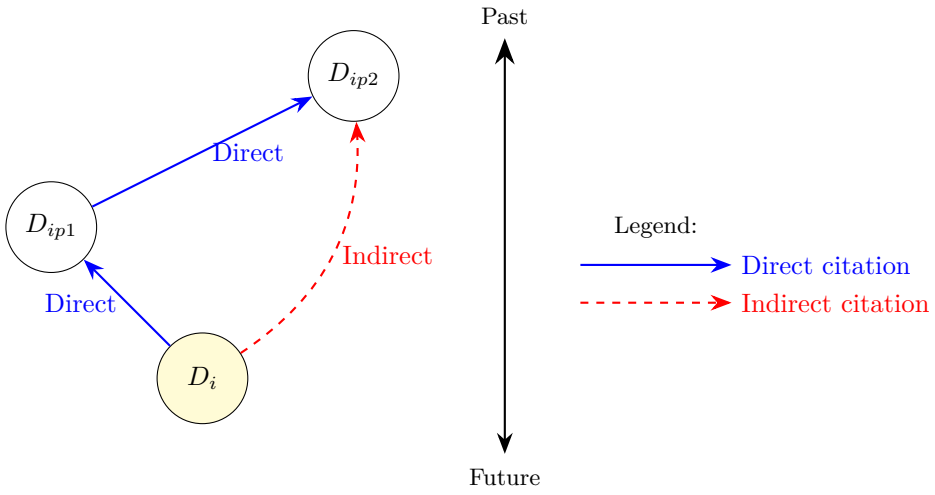
These citation networks are rich in relevant documents, much more so than that of the document collection, which is demonstrated by the author comparing precision of pools using BCS and FCS against that of the entire document collection in Figure **??**. The logical, and simple augmentation of the encoder CAL approach would be to exhaust both BCS and FCS networks of a seed document prior to initiating the encoder CAL process.

The theoretical benefits of citation network mining are that it can be used to augment the CAL process in ways that overcome some of the limitations of this process. Firstly, CAL requires labelled data to train a classifier model, which is assumed to perform better with more data points. Encoder CAL approaches suffer disproportionately to that of feature-based CAL approaches due to their need for larger amounts of training data to effectively learn meaningful representations. This is because encoder models like BERT need to learn complex contextual relationships between words and concepts, whereas feature-based models can rely on simpler statistical patterns. When working with limited labeled data in the early stages of screening, encoder models may struggle to generalise well, potentially leading to suboptimal performance in identifying relevant documents. As discussed in the Encoder CAL process, often a single sample seed document is used during the first epoch for fine-tuning. A better approach would be to exhaust the citation network of that seed document first for labelling, before using revealed relevant documents to fine-tune the model, potentially resulting in a more performant model at the earlier stages of screening with less oracle cost.

E.3 Extending current citation network mining approaches

BCS and FCS citation network mining faces a significant limitation in its inability to identify indirect citation relationships. An indirect citation occurs when research papers are connected through intermediate references, forming a chain of citations rather than a direct reference. For instance, when document D_i cites document D_{ip1}, which in turn cites document D_{ip2}, a relationship exists between D_i and D_{ip2} despite the absence of a direct citation. This relationship represents an indirect citation, which is shown in Figure E.3. This causes issues if D_{ip1} is not included in the document pool, as D_i and D_{ip2} will no longer have an edge.

This constraint makes it unsuitable as a complete solution for document relationship discovery for the encoder CAL process. However, researchers have proposed several modifications to the citation network mining process to address this limitation:



- **Matching isolated nodes based on similarity metric of their embeddings**: If N is all the documents in the total pool, and $N_{isolated}$ is the set of documents that are not cited by any other document in N, then

for each document $D_{ip} \in N_{isolated}$, find the document $D_i \in N$ with the highest similarity metric (i.e. cosine similarity) to D_{ip}. Add a artificial edge between D_i and D_{ip}.

- **Matching isolated compoments on similarity metric of their embeddings**: When analyzing document clusters, some small groups of documents (called isolated components) may be disconnected from the main cluster. These isolated components have fewer connections to other documents, which can reduce classification accuracy. To fix this:
  - Identify isolated components $C_{isolated}$ that have fewer or equal nodes than the main cluster
  - For each node in these isolated components
  - Calculate a similarity metric (i.e. cosine similarity) to nodes in larger clusters C_i
  - Connect it to the most similar large cluster by adding a artificial edge

This constraint however doesn't make it unsuitable as a partial improvement to the encoder CAL process for identifying relevant documents based on the initial seed document. Even without considering indirect citations, assessing the citation network of the seed document is potentially more relevant than that of the entire document collection. In table 8 it is unequivacle that the precision of relevant documents within pools using BCS, FCS and both together against that of the entire document collection is much higher.

Make this data!

Table 8: Precision of relevant documents within pools using BCS, FCS and both together against that of the entire document collection

E.4 Research Question 1

As outlined above, current approaches to encoder-based medical CAL rarely or indirectly leverage BCS/FCS as an initial expansion to relevant documents. Yet, BCS and FCS are known to yield high-precision citation pools, which could jumpstart the learning process. Therefore, the first research question is: *To what extent can leveraging BCS and FCS before initiating an encoder-based CAL pipeline improve precision in identifying relevant medical research articles?*

The proposed methodology would be to use the CLEF dataset, for which document relations could be extracted from forming a citation network using the opensource OpenAlex API[10]. A variety of seeds of known relevant documents would be used to form the initial BCS/FCS citation pools, which would be used initially within the encoder-based CAL process. The aim would be to have citation pools that have varying sizes (denoted by the number of nodes within the citation pool), so that the performance of different citation pool sizes can be compared. This could be achieved by creating a citation pool for every known relevant document using the OpenAlex API, which can be parralised across multiple threads. From the list of citation pool sizes, seeds would be selected from the lower, middle and upper ranges of the list.

After exhausting the citation network of the seed documents, the process would continue with the standard encoder-based CAL process, up to a maximum number of iterations.

Datasets: CLEF Seeds: A selection of seeds denoting known relevant documents.

Backward and forward citation pool construction

Experimental design

[10] https://openalex.org/docs/api

Citation Augmented considerations Baseline considerations Evaluation metrics Ablation studies BCS - only vs FCS only vs BCS+FCS expansions Varying seed sizes (1 vs 5 vs 10 known relevant documents) Analysis plan: Check to see how quickly each approach achieves a given level of prevision Computational costs - measure how much computational resources are required to achieve a given level of precision Citation network density - Correlate the final performance with the desnsity / size of the BCS/fcs network to understand if a bigger BCS/FCS network leads to better performance.

E.5 Graph Neural Networks

A research paper is a rich source of information, and contains multiple features that can be used to represent that document, however in the title and abstract screening phase, it is limited to only using the title and abstract features. As the previous research area aims to demonstrate, utilising other features could improve the precision of the encoder CAL process, so, logically utilising more features could improve the precision of the encoder CAL process further.

Previous work by this author has demonstated that features about authors, primary topic and publication date all impact classification accuracy. link this to the retraction watch paper

In-keeping with the above research theme, graph neural networks offer a natural extension to considering additional document features, and still being able to utilise the structural information about relationships between documents.

E.6 LLMs and citation network mining

The motivation for using LLMs and graph networks is to combine the structurality of graph citation networks with the ability of the LLMs to comprehend the semantic meaning of documents. As outlined, citation network graph analysis occurs above the document level through utilisation of extracted features about documents. LLMs are a natural replacement for extraction of features, as they possess the ability to understand semantic meaning of documents. The ultimate goal to use LLMs and graph networks is to complment and enhance isuses with the other.

Research has been conducted into the use of LLMs within the graph neural networks, and has developed a robust taxonomy for categorising the use of LLMs within graph networks [**llm4g**].

The first application of LLms within graph networks is to use LLM as an enchancer. Typically graph neural networks encode text into nodes using simple bag-of-words, skip-gram or TF-IDF. LLMs are able to encode text into nodes using more complex features, such as semantic meaning, which can be used within the graph neural networks. This can be further subdivided into explanation based and embedding based enchancers.

Explanation based enhancers query an LLM using prompting to capture higher level features about documents, which is used to enrich node representations prior to processing with a graph neural network, with the process being abstracted in in Figure 7. The approach used by https://arxiv.org/pdf/2305.19523 was to prompt GPT 3-5 with the abstract and text of a document along with a questions about that document using a zero-shot approach. The LLM reponse then forms features which are amended to the original node representations. Issues with this approach is that this requires domain specific knowledge, as features which are deemed important (and hence prompt used) are dependent on the domain of the research. It was performant on the pubmed domain, scoring greater node classification accuracy using this approach (0.9618 ± 0.0053) than utilising an LLM alone on pubmed data (0.9494 ± 0.0046).

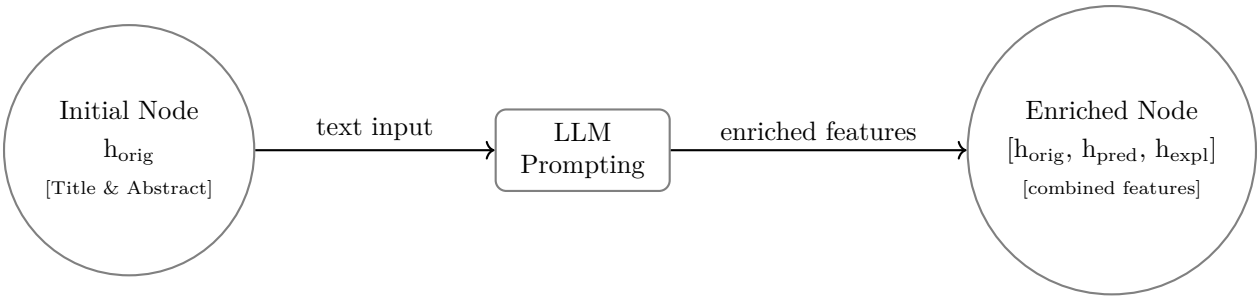


Figure 7: Node feature enrichment process using LLM and LM

E.7 Research Question 2

Proposal: Utilising more features in the encoder CAL process can improve the precision of the encoder CAL process.

F Notes on Graph Neural Networks

A node is represented by a feature matrix, which contains information about the document. This **Node feature matrix**, X, which has the dimensions of m (the number of nodes) and n (the number of features). $X \in \mathbb{R}^{m \times n}$. X does not have to be a square matrix, and does not encode any information about the structure of the graph.

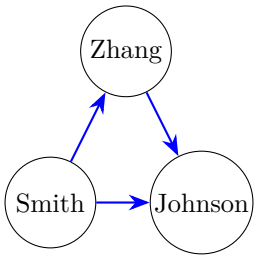
Consider 3 research papers as nodes, with features: [Author, Title Length, Abstract Length, Citation Count]

$X = \begin{bmatrix} \text{"Smith"} & 82 & 500 & 45 \\ \text{"Johnson"} & 95 & 475 & 23 \\ \text{"Zhang"} & 67 & 612 & 89 \end{bmatrix}$ Where $X \in \mathbb{R}^{3 \times 4}$ represents:

3 papers (rows) 4 features per paper (columns) Mixed data types (categorical and numerical)

Structural information is encoded in the **adjacency matrix**, A, which has the dimensions of m (the number of nodes) and m (the number of nodes). $A \in \mathbb{R}^{m \times m}$. A encodes information about the structure of the graph, and is used to determine relationships between nodes. Conventionally the source nodes are the rows, and the destination nodes the columns of the matrix. 1 indicates an edge between the source node u and destination node v. Note that there is a choice to make here, with the diagonal of the matrix being 0 or 1. This choice is based on wheter you consider the source node to be connected to itself. In cases where the representation of the node is dependent on itself and adjacent nodes, the diagonal should be set to 1. In the scenario of citation networks, the diagonal should be set to 1, as a paper is likely to reference and build upon its own findings throughout. By setting the diagonal to 0, it is akin to attempting to predict the representation of the node base only on its adjacent nodes, which is not the case in citation networks. If an adjaceny metrix is symmetric around it's diagonal, then the graph is undirected, otherwise it is directed (i.e. U is connected to V and V is conencted to U). In citation networks, this is not the case, as because paper A cites paper B, it does not mean the reverse is true.

Consider the same 3 research papers, with the following adjacency matrix: $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ Which represents the following graph:



With both X and A defined, we can numerically represent the graph. The node feature matrix X is the initial/input node features, with our goal for learning on graphs to learn node embeddings $H \in \mathbb{R}^{N \times D}$ where D is a chosen hidden dimension size.

G Message Passing Neural Networks

We need an approach that can work with the graph structure, which has variable number of nodes and edge conenctions between nodes. Historically with the CNN architecture, the input size was fixed, and the network was able to learn spatial invariance through the use of convolutional filters that were invariant to the location of the feature in the input. With graph structured data, the number of nodes and connections between nodes can vary for each graph, and spatial invariance is not invariant to the location of the feature in the graph.

Message Passing Neural Networks (MPNNs) are a type of graph neural network that can learn spatial invariance through the use of message passing between nodes. The basic idea of MPNNs is to iteratively update node representations by passing messages between connected nodes. This process is repeated for a fixed number of iterations, or until convergence.

The process is defined as follows:

- Message: every node decides how to send information to neighboring nodes it is conencted to by edges
- Aggregate: nodes recieve messages from all their neighbors, who also passed messages and decides how to combine the information from all of its neighbors.
- Update: each node decies how to combine neighbourhood information with its own information and updates it embedding for the next timestep.

By doing this we have nodes pass each other information and disseminate information around the graph, allowing the network to learn spatial invariance. This can be repeated for a fixed number of iterations (K), with the larger the value of K, the more the more diffuse the information around the graph becomes.

Each section of the MPNN process in more detail:

G.1 Message

The source node U will pass a message m_{uv} to the destination node V. The message depends on the GNN architecture with the easiest example message being passed being U node's feature h_u vector to V.

G.2 Aggregate

The destination node V will recieve messages from all its neighbouring nodes, and needs to decide how to combine the information from all of its neighbours. This is typically done using a sum, average or max pooling of the messages from all neighbouring nodes. It is important that the aggregation function has to be a permutation invariant function, as the order of the messages should not affect the output.

This gives us a combined neighbourhood node embedding, denoted as $h_{N(V)}$, where $N(V)$ is the set of all neighbouring nodes to V, meaning all nodes connected to V by an edge.

$h_{N(v)}^{k+1} = AGGREGATE(h_u^k, \forall u \in N(v))$

G.3 Update

Each node updates its own embedding based on the combined neighbourhood embedding and its own embedding from the previous timestep.

$h_v^{k+1} = \sigma(W \cdot CONCAT(h_v^k, h_{N(v)}^{k+1}))$

Search criteria for Graph Neural Networks and Active Learning ("graph neural network" OR GNN) AND ("active learning" OR "interactive learning") AND (document OR citation OR literature) AND ("relevance feedback" OR "document classification") AND ("semi-supervised" OR "partially labeled") Database-specific versions:

arXiv: search within cs.LG, cs.IR, cs.CL categories PubMed: add "systematic review" OR "literature review" terms

III TIMELINE

This is a non-binding timeline for this PhD. It is accepted that this will likely change and only represents estimates given the current available information, the timeline is available in Gantt format in Figure 8

In creating this time-line for the PhD several assumptions were made:

- **Holiday Periods:** Two weeks of holiday have been accounted for during the summer and a two-week period over Christmas and New Year. This timeline does not account for additional holidays, which will be determined later.
- Front-Loaded Research: The research plan is front-loaded, with significant emphasis placed on the early stages of the project. This approach allows additional literature searching, coding, and experimental setup to be completed in advance, providing a solid foundation for later stages of research, the research questions of which are currently flexible. This strategy also ensures that any necessary adaptations can be made based on early findings, reducing the risk of major delays later in the project.
- **Research Flexibility:** Significant breaks are scheduled between work on research questions to allow for adaptation or extension if research questions need to be adjusted. In addition, there is a dedicated period for each of the three research themes before beginning coding or experimental setups. This time is intended for further literature review, acknowledging the rapidly evolving nature of this field. Advancement of the literature are expected before the start of each research question period.
- **Undetermined Research Questions:** Two research questions have deliberately been left open. Depending on the findings of the first three research questions, new research opportunities or developments in the field may emerge that require investigation.
- **Publication goals:** The goal is to produce at least one publishable piece of work for each research question. The preferential publication venue is SIGIR, recognised as the highest-rated publication venue for this sub-domain, aside from broader higher impact journals such as ACL. This goal is deliberately ambitious, as until the research is complete, it is impossible to determine the merits of its findings.
- **Conference Scheduling:** SIGIR abstract submissions and conferences occur around the same time each year, which has been considered in the planning. Note that many RQs are concluded prior to the expected submission dates.

A Potential Threats

A risk matrix of all potential threats is outlined in Table 9.

- **Research Delays:** Unforeseen challenges in research or coding could lead to delays. These could arise from the complexity of the research questions, issues with experimental setups, or unexpected results that require additional analysis. This is mitigated through periods at the end of research questions which can be utilised, if necessary.
- **Technological advancement:** The fast-moving nature of the field could result in the emergence of new technologies or methodologies that could make parts of the planned research less relevant or require a significant change of focus. This is mitigated through periods allowing additional literature review before coding on that research question.

- **Publication risks:** There is always the risk that the research might not yield results suitable for publication or that the publication process itself might be more time-consuming than anticipated, especially if revisions are required. The nature of the publication system is that there will be extended gaps between the completion of the research and its publication, making modification of the research to match a reviewer's expectations difficult.
- **Personal and Health Factors:** Extended periods of high-intensity work can lead to burnout or health problems, potentially affecting the planned schedule. The Gantt chart does not and cannot fully account for extended absences due to illness or other personal factors.
- **Open ended research questions:** It could be that at the time of reaching the open-ended research questions, a research area has not been identified. The author believes that this is unlikely, given that this field is limited, and other potential research questions have been dropped to ensure this flexibility.
- **Competing time constraints from PGDip:** Due to the concurrent requirements of the PGDip with this PHd, there will at times be constraints placed on my available time. So far, I believe, I have demonstrated good time keeping skills, and believe this schedule to be reasonable given the other competing time constraints. If there is a clash, PH.d. will take priority over PGDip.

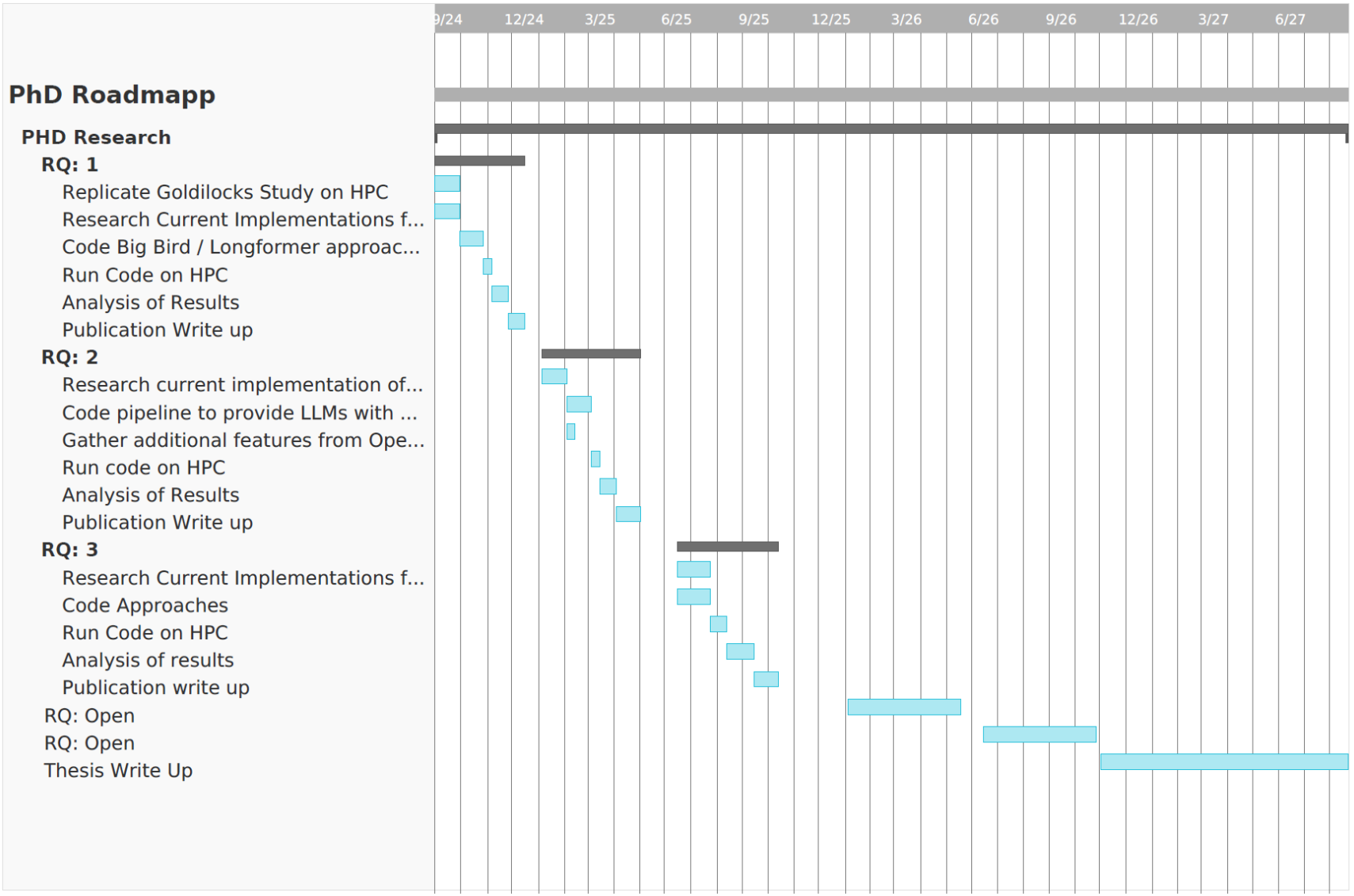


Figure 8: Gantt Chart for OVerview of Timeline for PHd

Threat	Likelihood (L)	Impact (I)	Risk	Response
Research Delays	3	4	12	Reduce the number of research questions to mitigate delays.
Technological Advancement	3	3	9	Increase time allocated for additional literature review to stay updated with advancements.
Publication Risks	3	3	9	If results are not deemed valid during the analysis phase, proceed to the next research question without spending time on write-up.
Personal and Health Factors	2	4	8	Prioritise health by taking regular breaks and managing workload effectively to avoid burnout.
Open-ended Research Questions	2	3	6	Continuously monitor new publications within the domain to ensure relevant and timely research questions.
Competing Time Constraints (PGDip)	3	3	9	Minimize involvement in PGDip activities where possible to focus on PhD work.

Table 9: Risk Scoring Matrix for threats to PhD. Responses are provided for medium impact risks.

IV ETHICS

This research will comply with the research ethics norms and rules outlined by the University of Sheffield. Most of the data sets used by this research (CLEF, Synergy, and RCV1-v2 data sets) would fall into the category of publicly available anonymised data/published media; hence, ethical approval is not required.

However, the Jeb Bush email dataset contains unanonymised emails. Although it is publically available and used for research, an application for its use will be sought through the university's ethical approval process in due time.

V PROFESSIONAL DEVELOPMENT PLAN

The training provided as part of the PhD with integrated PGDip replaces the standard Doctoral Development programme, and further activities outside of this are not required to be undertaken.[11]. The PGDip component has multiple strands associated with it, such as participation in group work, a requirement for technical training, responsible research and innovation, attending discipline-specific lectures, involvement with journal clubs, outreach work from the department itself, and completion of the FCE6100 (Research Ethics and Integrity training) as part of CDT's COM61003. As evidence of active involvement with professional development, the author submits his logs from his first year of engagement with the process, which were submitted in June 2024 - see Appendix **??** and **??**. These are due yearly (i.e. every June) as the programme progresses. Due to the large size of this, this is not included in this body of work but rather is included as additional PDFs for the readers.

[11] https://sites.google.com/sheffield.ac.uk/slt-cdt-handbook/the-centre

VI AUTHORS SUPPORTING WORKS

The author is actively involved in 3 other projects with publishable outcomes.

A Predicting Retracted Research

In my study, "Predicting Retracted Research," I, along with my supervisor Mark Stevenson, explored the challenge of identifying flawed scientific publications before their dissemination - see Appendix **??**. We developed a novel data set by combining information from the Retraction Watch database and the OpenAlex API. This dataset includes metadata, abstracts, and citation metrics for 16,224 articles (8,112 retracted and 8,112 non-retracted) published between 2000 and 2020. Various machine learning models are used to predict retracted articles, with a gradient booster model achieving the highest precision at 0.691. An ablation study highlighted the critical role of the abstract in classification accuracy, recall, and F1 score, whereas the First Author's Country was pivotal for precision in feature-based classifiers. The research demonstrates the feasibility of using machine learning to aid peer review by highlighting potentially problematic research, although further refinement is necessary for practical implementation. The data set and code are publicly available to encourage further research in this area. This work has been written and is intended to be submitted to Informetrics[12]for publication. This fits the overall PhD research theme of assessing research data and ultimately showing that the evidence being assessed, even despite being published, can be flawed. Currently, no further follow-up is intended for this work.

Threat to Ph.D.: Low. Most of the work has been completed for this project.

B The stopping problem

Within the creation of SRs, the overall goal of TAR is to get to as near perfect total recall as possible, or when you have exhausted a resource such as a human reviewer. However, other stopping strategies could be more materially useful, to fulfil an information need, such as stopping when you have returned enough information to make a decision. In this joint research, between the author, Mark Stevenson and Anthony Hughes, we use information provided from cochrane reviews and create algorithms that stop when the acquisition of knowledge (positively included studies) meets a criterion and then evaluate how close these stopping algorithms got to the final outcome.

Threat to Ph.d.: Medium, this work is ongoing and has undergone many revisions. However, it is likely to result in high-quality publishable research.

C CPET analysis and deep neural networks

A cardiopulmonary exercise test (CPET) is performed before certain anaesthetic procedures, the outcome of which is used to determine the suitability of the patient for this procedure. Current approaches use summarised data to generate decision models, whose data are derived from summary values provided by the machine. The machine also records " breath-by-breath" data measurements, which, while they are the basis for the summary values, are not used by these models. This research project attempts to determine whether the use of deep neural networks, with these " 'breath-by-breath"' data, is superior to that of the traditional summary-model approach. This research was devised by an NHS researcher.

Threat to Ph.d.: Low. I am providing coding assistance to this project, and will not be involved in analysis or extensively involved in research write-up, outside of the technical side. This is also likely to result in publishable research, and will likely be published in medical domain venues, promoting interdisciplinary work.

[12]https://www.sciencedirect.com/journal/journal-of-informetrics

References

[1] Peter Kranke. "Evidence-based practice: how to perform and use systematic reviews for clinical decision-making". In: *European Journal of Anaesthesiology* 27.9 (Sept. 2010), pp. 763–772. ISSN: 1365-2346. DOI: 10.1097/EJA.0b013e32833a560a (cit. on p. 6).

[2] *Oxford Centre for Evidence-Based Medicine: Levels of Evidence (March 2009)*. Type: Web Page. URL: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009 (visited on 07/29/2024) (cit. on p. 6).

[3] Jennifer A. Swanson, DeLaine Schmitz, and Kevin C. Chung. "How to Practice Evidence-Based Medicine". In: *Plastic and reconstructive surgery* 126.1 (July 2010). tex.pmcid: PMC4389891, pp. 286–294. ISSN: 0032-1052. DOI: 10.1097/PRS.0b013e3181dc54ee. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4389891/ (visited on 07/29/2024) (cit. on p. 6).

[4] Gordon H. Guyatt et al. "GRADE: an emerging consensus on rating quality of evidence and strength of recommendations". In: *BMJ (Clinical research ed.)* 336.7650 (Apr. 2008). tex.copyright: © BMJ Publishing Group Ltd 2008, pp. 924–926. ISSN: 0959-8138, 1756-1833. DOI: 10.1136/bmj.39489.470347.AD. URL: https://www.bmj.com/content/336/7650/924 (visited on 07/29/2024) (cit. on p. 6).

[5] "Cochrane Handbook for Systematic Reviews of Interventions". In: (cit. on p. 6).

[6] Asghar Ghasemi et al. "Scientific Publishing in Biomedicine: A Brief History of Scientific Journals". In: *International Journal of Endocrinology and Metabolism* 21.1 (Dec. 2022). tex.pmcid: PMC10024814, e131812. ISSN: 1726-913X. DOI: 10.5812/ijem-131812. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10024814/ (visited on 07/29/2024) (cit. on pp. 6, 8).

[7] Jeremy Howick. "Front Matter". In: *The Philosophy of Evidence-Based Medicine*. John Wiley & Sons, Ltd, 2011, pp. i–xiv. ISBN: 978-1-4443-4267-3. DOI: 10.1002/9781444342673.fmatter. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444342673.fmatter (visited on 07/29/2024) (cit. on p. 6).

[8] Gehad Mohamed Tawfik et al. "A step by step guide for conducting a systematic review and meta-analysis with simulation data". In: *Tropical Medicine and Health* 47.1 (Aug. 2019), p. 46. ISSN: 1349-4147. DOI: 10.1186/s41182-019-0165-6. URL: https://doi.org/10.1186/s41182-019-0165-6 (visited on 07/29/2024) (cit. on pp. 7, 8).

[9] Ian Shemilt et al. "Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews". In: *Systematic Reviews* 5.1 (Aug. 2016). tex.pmcid: PMC4989498, p. 140. ISSN: 2046-4053. DOI: 10.1186/s13643-016-0315-4 (cit. on pp. 7, 8).

[10] Melita J. Giummarra, Georgina Lau, and Belinda J. Gabbe. "Evaluation of text mining to reduce screening workload for injury-focused systematic reviews". In: *Injury Prevention* 26.1 (Feb. 2020). tex.copyright: © Author(s) (or their employer(s)) 2020. No commercial re-use. See rights and permissions. Published by BMJ., pp. 55–60. ISSN: 1353-8047, 1475-5785. DOI: 10.1136/injuryprev-2019-043247. URL: https://injuryprevention.bmj.com/content/26/1/55 (visited on 07/29/2024) (cit. on p. 7).

[11] Miguel Marques Antunes et al. "Preoperative statin therapy for adults undergoing cardiac surgery - Marques Antunes, M - 2024 — Cochrane Library". In: (). ISSN: 1465-1858. URL: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD008493.pub5/full (visited on 07/29/2024) (cit. on p. 8).

[12] B. Nussbaumer-Streit et al. “Resource use during systematic review production varies widely: a scoping review”. In: *Journal of Clinical Epidemiology* 139 (Nov. 2021), pp. 287–296. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2021.05.019. URL: https://www.sciencedirect.com/science/article/pii/S0895435621001712 (visited on 07/29/2024) (cit. on p. 8).

[13] Justin S. Smith et al. “Less is more: Sampling chemical space with active learning”. In: *The Journal of Chemical Physics* 148.24 (May 2018), p. 241733. ISSN: 0021-9606. DOI: 10.1063/1.5023802. URL: https://doi.org/10.1063/1.5023802 (visited on 07/30/2024) (cit. on p. 9).

[14] Steven C. H. Hoi et al. “Batch mode active learning and its application to medical image classification”. In: *Proceedings of the 23rd international conference on Machine learning*. ICML ’06. New York, NY, USA: Association for Computing Machinery, June 2006, pp. 417–424. ISBN: 978-1-59593-383-6. DOI: 10.1145/1143844.1143897. URL: https://doi.org/10.1145/1143844.1143897 (visited on 07/30/2024) (cit. on p. 9).

[15] Maura R. Grossman and Gordon V. Cormack. “Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient than Exhaustive Manual Review Annual Survey”. In: *Richmond Journal of Law and Technology* 17.3 (2010), [i]–48. URL: https://heinonline.org/HOL/P?h=hein.journals/jolt17&i=471 (visited on 07/31/2024) (cit. on p. 9).

[16] Dana Angluin. “Queries and Concept Learning”. In: *Machine Learning* 2.4 (Apr. 1988), pp. 319–342. ISSN: 1573-0565. DOI: 10.1023/A:1022821128753. URL: https://doi.org/10.1023/A:1022821128753 (visited on 07/31/2024) (cit. on p. 9).

[17] Opeoluwa Akinseloyin, Xiaorui Jiang, and Vasile Palade. *A Novel Question-Answering Framework for Automated Abstract Screening Using Large Language Models*. tex.copyright: © 2024, Posted by Cold Spring Harbor Laboratory. This pre-print is available under a Creative Commons License (Attribution-NonCommercial-NoDerivs 4.0 International), CC BY-NC-ND 4.0, as described at http://creativecommons.org/licenses/by-nc-nd/4.0/. June 2024. DOI: 10.1101/2023.12.17.23300102. URL: https://www.medrxiv.org/content/10.1101/2023.12.17.23300102v3 (visited on 06/27/2024) (cit. on p. 9).

[18] David D. Lewis and William A. Gale. “A Sequential Algorithm for Training Text Classifiers”. In: *SIGIR ’94*. Ed. by Bruce W. Croft and C. J. van Rijsbergen. London: Springer, 1994, pp. 3–12. ISBN: 978-1-4471-2099-5. DOI: 10.1007/978-1-4471-2099-5_1 (cit. on p. 9).

[19] Pengzhen Ren et al. *A Survey of Deep Active Learning*. Dec. 2021. DOI: 10.48550/arXiv.2009.00236. URL: http://arxiv.org/abs/2009.00236 (visited on 07/31/2024) (cit. on p. 10).

[20] Ron Artstein and Massimo Poesio. “Survey Article: Inter-Coder Agreement for Computational Linguistics”. In: *Computational Linguistics* 34.4 (2008), pp. 555–596. DOI: 10.1162/coli.07-034-R2. URL: https://aclanthology.org/J08-4004 (visited on 07/31/2024) (cit. on p. 10).

[21] Tim Pearce, Alexandra Brintrup, and Jun Zhu. *Understanding Softmax Confidence and Uncertainty*. June 2021. DOI: 10.48550/arXiv.2106.04972. URL: http://arxiv.org/abs/2106.04972 (visited on 07/31/2024) (cit. on p. 10).

[22] Dan Wang and Yi Shang. “A new active labeling method for deep learning”. In: *2014 International Joint Conference on Neural Networks (IJCNN)*. July 2014, pp. 112–119. DOI: 10.1109/IJCNN.2014.6889457. URL: https://ieeexplore.ieee.org/document/6889457/?arnumber=6889457 (visited on 07/31/2024) (cit. on p. 10).

[23] Chuan Guo et al. *On Calibration of Modern Neural Networks*. Aug. 2017. DOI: 10.48550/arXiv.1706.04599. URL: http://arxiv.org/abs/1706.04599 (visited on 07/31/2024) (cit. on p. 10).

[24] Burr Settles. *Active Learning Literature Survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences, 2009. URL: https://minds.wisconsin.edu/handle/1793/60660 (visited on 07/31/2024) (cit. on p. 10).

[25] Gordon V. Cormack and Maura R. Grossman. *Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review*. Apr. 2015. DOI: 10.48550/arXiv.1504.06868. URL: http://arxiv.org/abs/1504.06868 (visited on 06/27/2024) (cit. on pp. 10, 12).

[26] A.M. Cohen et al. "Reducing Workload in Systematic Review Preparation Using Automated Citation Classification". In: *Journal of the American Medical Informatics Association : JAMIA* 13.2 (2006). tex.pmcid: PMC1447545, pp. 206–219. ISSN: 1067-5027. DOI: 10.1197/jamia.M1929. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1447545/ (visited on 08/01/2024) (cit. on p. 11).

[27] Gerbrich Ferdinands et al. "Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average Time to Discover relevant records". In: *Systematic Reviews* 12 (June 2023). DOI: 10.1186/s13643-023-02257-7 (cit. on p. 11).

[28] Gordon V. Cormack and Maura R. Grossman. "Evaluation of machine-learning protocols for technology-assisted review in electronic discovery". In: *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. Gold Coast Queensland Australia: ACM, July 2014, pp. 153–162. ISBN: 978-1-4503-2257-7. DOI: 10.1145/2600428.2609601. URL: https://dl.acm.org/doi/10.1145/2600428.2609601 (visited on 07/31/2024) (cit. on p. 11).

[29] Gordon V. Cormack and Maura R. Grossman. "Systems and methods for conducting a highly autonomous technology-assisted review classification". U.S. pat. 20160371261A1. Individual. Dec. 2016. URL: https://patents.google.com/patent/US20160371261A1/en (visited on 08/04/2024) (cit. on p. 13).

[30] Ashish Vaswani et al. *Attention Is All You Need*. Aug. 2023. DOI: 10.48550/arXiv.1706.03762. URL: http://arxiv.org/abs/1706.03762 (visited on 08/04/2024) (cit. on p. 12).

[31] Eugene Yang et al. *Goldilocks: Just-Right Tuning of BERT for Technology-Assisted Review*. Jan. 2022. DOI: 10.48550/arXiv.2105.01044. URL: http://arxiv.org/abs/2105.01044 (visited on 06/27/2024) (cit. on p. 12).

[32] Y. Xu et al. *Forget Me Not: Reducing Catastrophic Forgetting for Domain Adaptation in Reading Comprehension*. Nov. 2020. DOI: 10.48550/arXiv.1911.00202. URL: http://arxiv.org/abs/1911.00202 (visited on 08/04/2024) (cit. on p. 12).

[33] *ielab/goldilocks-reproduce*. June 2024. URL: https://github.com/ielab/goldilocks-reproduce (visited on 07/31/2024) (cit. on p. 12).

[34] Evangelos Kanoulas et al. "CLEF 2017 technologically assisted reviews in empirical medicine overview". In: *CEUR Workshop Proceedings* 1866 (Sept. 2017), pp. 1–29. ISSN: 1613-0073. URL: http://ceur-ws.org/Vol-1866/ (visited on 07/31/2024) (cit. on p. 14).

[35] Evangelos Kanoulas et al. "CLEF 2018 technologically assisted reviews in empirical medicine overview: 19th Working Notes of CLEF Conference and Labs of the Evaluation Forum, CLEF 2018". In: *CEUR Workshop Proceedings* 2125 (July 2018). ISSN: 1613-0073. URL: http://www.scopus.com/inward/record.url?scp=85051077484&partnerID=8YFLogxK (visited on 07/31/2024) (cit. on p. 14).

[36] Evangelos Kanoulas et al. "CLEF 2019 technology assisted reviews in empirical medicine overview". In: *CEUR Workshop Proceedings* 2380 (Sept. 2019). tex.copyright: cc_by. ISSN: 1613-0073. URL: https://strathprints.strath.ac.uk/71253/ (visited on 07/31/2024) (cit. on p. 14).

[37] Xinyu Mao, Bevan Koopman, and Guido Zuccon. "A Reproducibility Study of Goldilocks: Just-Right Tuning of BERT for TAR". In: *Advances in Information Retrieval*. Ed. by Nazli Goharian et al. Vol. 14611. Cham: Springer Nature Switzerland, 2024, pp. 132–146. ISBN: 978-3-031-56065-1 978-3-031-56066-8. DOI: 10.1007/978-3-031-56066-8_13. URL: https://link.springer.com/10.1007/978-3-031-56066-8_13 (visited on 07/31/2024) (cit. on pp. 14, 18).

[38] Jonathan De Bruin et al. *SYNERGY - Open machine learning dataset on study selection in systematic reviews*. In collab. with Jonathan De Bruin and Rens Van De Schoot. 2023. DOI: 10.34894/HE6NAQ. URL: https://dataverse.nl/citation?persistentId=doi:10.34894/HE6NAQ (visited on 07/31/2024) (cit. on p. 14).

[39] Adam Roegiest et al. "TREC 2015 Total Recall Track Overview". In: *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*. Ed. by Ellen M. Voorhees and Angela Ellis. Vol. 500-319. NIST Special Publication. National Institute of Standards and Technology (NIST), 2015. URL: https://trec.nist.gov/pubs/trec24/papers/Overview-TR.pdf (visited on 07/31/2024) (cit. on p. 15).

[40] Maura R. Grossman, G. Cormack, and Adam Roegiest. "TREC 2016 Total Recall Track Overview". In: 2016. URL: https://www.semanticscholar.org/paper/TREC-2016-Total-Recall-Track-Overview-Grossman-Cormack/126240dedd75626fd736f0485d06f1f516517e54 (visited on 07/31/2024) (cit. on p. 15).

[41] David D. Lewis et al. "RCV1: A New Benchmark Collection for Text Categorization Research". In: *Journal of Machine Learning Research* 5 (Apr 2004), pp. 361–397. ISSN: ISSN 1533-7928. URL: https://www.jmlr.org/papers/v5/lewis04a.html (visited on 07/31/2024) (cit. on p. 15).

[42] Alison O'Mara-Eves et al. "Using text mining for study identification in systematic reviews: a systematic review of current approaches". In: *Systematic Reviews* 4.1 (Jan. 2015), p. 5. ISSN: 2046-4053. DOI: 10.1186/2046-4053-4-5. URL: https://doi.org/10.1186/2046-4053-4-5 (visited on 08/02/2024) (cit. on p. 16).

[43] Guy Tsafnat et al. "Systematic review automation technologies". In: *Systematic Reviews* 3.1 (July 2014), p. 74. ISSN: 2046-4053. DOI: 10.1186/2046-4053-3-74. URL: https://doi.org/10.1186/2046-4053-3-74 (visited on 08/12/2024) (cit. on p. 16).

[44] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. 1st ed. Cambridge University Press, July 2008. ISBN: 978-0-521-86571-5 978-0-511-80907-1. DOI: 10.1017/CBO9780511809071. URL: https://www.cambridge.org/core/product/identifier/9780511809071/type/book (visited on 08/03/2024) (cit. on p. 16).

[45] *BMC Medical Research Methodology*. BioMed Central. URL: https://bmcmedresmethodol.biomedcentral.com/submission-guidelines/preparing-your-manuscript/research-article (visited on 11/13/2024) (cit. on p. 18).

[46] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. *LinkBERT: Pretraining language models with document links*. 2022. arXiv: 2203.15827[cs.CL]. URL: https://arxiv.org/abs/2203.15827 (cit. on p. 18).

[47] Yu Gu et al. *Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing*. version: 6. Sept. 16, 2021. arXiv: 2007.15779. URL: http://arxiv.org/abs/2007.15779 (visited on 11/13/2024) (cit. on p. 18).

[48] Di Jin et al. *What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams*. version: 1. Sept. 28, 2020. arXiv: 2009.13081. URL: http://arxiv.org/abs/2009.13081 (visited on 11/13/2024) (cit. on p. 18).

[49] Dan Hendrycks et al. *Measuring Massive Multitask Language Understanding*. Jan. 12, 2021. arXiv: 2009.03300. URL: http://arxiv.org/abs/2009.03300 (visited on 11/13/2024) (cit. on p. 18).

[50] Carol Lefebvre et al. "Cochrane handbook for systematic reviews of interventions". In: *Oxfordshire, UK: The Cochrane Collaboration* (2011) (cit. on p. 20).

[51] Jo Akers, R Aguiar-Ibáñez, and A Baba-Akbari. "Systematic reviews: CRD's guidance for undertaking reviews in health care". In: *University of York* (2009) (cit. on p. 20).

[52] Simon Briscoe, Alison Bethel, and Morwenna Rogers. "Conduct and reporting of citation searching in Cochrane systematic reviews: A cross-sectional study". In: *Research Synthesis Methods* 11.2 (July 4, 2019), p. 169. DOI: 10.1002/jrsm.1355. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7079050/ (visited on 11/13/2024) (cit. on p. 22).

[53] *MECIR Manual — Cochrane Community*. URL: https://community.cochrane.org/mecir-manual (visited on 11/13/2024) (cit. on p. 22).