Determining the Effectiveness of Multi-User, Hybrid, Human-Computer Assessment Strategies for High-Recall Information Retrieval Systems

by

Solaiappan Alagappan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2022

© Solaiappan Alagappan 2022

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Electronic Discovery (eDiscovery), a use-case of High-Recall Information Retrieval (HRIR), seeks to obtain substantially all and only the relevant documents responsive to a request for production in litigation. Applications of HRIR typically use a human as their oracle to determine the relevance for a large number of documents, which is expensive both in terms of time/effort and cost. HRIR experts suggest that Continuous Active Learning (CAL) systems, the state-of-the-art information retrieval (IR) tools used for eDiscovery have the potential to achieve superior results and achieving them is limited primarily by the fallibility of the accuracy of human relevance assessments.

In this research, we seek to understand the impact of the error rate in human relevance feedback on CAL systems and attempt to address them using six distinct multi-user–based, hybrid, human-computer assessment strategies. In contrast to the widely used single-user-based, hybrid, human-computer assessment strategy, these multi-user strategies re-provision resources to re-reviewing documents that the user may have misjudged, rather than examining more documents, in the pursuit of mitigating human relevance feedback error, while also achieving a high-recall and high-precision review. Within the constraints of a specified review budget, we want to determine which review strategy has the best chance of precisely retrieving more relevant documents.

Our results show that leveraging a multi-user review strategy that "efficiently" uses three reviewers to review documents (CAL QC–Type 1) and a multi-user review strategy that uses the CAL system as one of the users in a three-reviewer approach (CAL QC–Type 2) can enable the end-to-end CAL system to achieve a significantly higher recall and higher precision when compared to that achieved by a single-user-based review strategy while employing the same review budget. This research provides evidence that CAL systems have the potential to better accommodate the needs of the HRIR applications by incorporating multi-user review strategies.

Acknowledgements

I want to give my deepest gratitude to my supervisor, Maura R. Grossman, for making it possible to complete my thesis by guiding me and supporting me in my research. I appreciate the advice given to me by Maura and the freedom to explore my ideas in depth. I have amassed immense knowledge under her supervision over the past two years enabling me to succeed in my future endeavors.

Aside from my supervisor, I would like to thank Gordon V. Cormack and Charles L.A. Clarke for reading and providing feedback on my thesis and allowing me to improve my work. Along with University of Waterloo research scholarship, this research work is supported by the Vector Scholarship in Artificial Intelligence, provided through the Vector Institute.

I want to thank my parents, P. Alagappan and A. Devi, and my brother, A.P. Arjun for providing me with support and guidance throughout my education and endless opportunities to succeed in my career. Finally, I would like to thank my friends in Waterloo who helped me feel at home even in the exceptional circumstances of this time.

Dedication

I dedicate this thesis to my family, and friends whose support, love and presence in the last year has helped me finish this.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Electronic discovery (eDiscovery) is a phase of litigation in which the parties involved in a legal dispute, retrieve and exchange relevant documents from their own document collections to substantiate their positions and disprove their adversary's positions. It involves identifying, preserving, collecting, searching, and producing electronically stored information (ESI) as evidence in lawsuits or investigations. Emails, documents, presentations, databases, voicemail, audio and video files, social media, and websites are some examples of ESI. Legal teams on both sides of a case seek to obtain substantially all relevant ESI with reasonable assessment effort; in other words, they desire to achieve high-recall information retrieval (HRIR), with reasonable precision. Achieving high recall in eDiscovery faces a number of real-time challenges, to name a few:

1. Increasing data volumes
2. Changing data landscapes (i.e. data variety)
3. Limited availability of resources; and
4. Declining budgets

Extracting documents from large-volume and varied sources of ESI and reviewing them to identify relevant documents for production makes eDiscovery one of the most labor-, time- and cost-intensive phase of litigation [23]. Review, a stage in eDiscovery, is where the legal team employs a group of junior or contract attorneys to determine possibly relevant documents in a case. Traditionally, such junior or contract attorneys were required to review each document in the collection, which can often take several minutes per document [38]. The expense of such a manual-review strategy grows linearly with the magnitude of the collection, and therefore, linear review has proven increasingly unsustainable as collections have increased massively [28]. According to a 2012 RAND study, document review alone can contribute up to 73% of the entire cost involved in eDiscovery for high-volume cases [26].

Because document review is typically the most significant and costly element of an eDiscovery effort, it is an area that has received a lot of attention. Recently, supervised machine-learning approaches, referred to as "Technology-Assisted Review" (TAR) have established themselves as the standard eDiscovery technique to handle rapidly increasing document collections, alleviating reviewer effort, and achieving superior results [11]. Technology-Assisted Review is the process of ranking or categorizing a collection of documents using a computer algorithm that harnesses human judgments of one or more subject matter expert(s) (SME(s)) on a smaller set of documents and extrapolating those judgments to the remaining documents in the corpus [9]. A seminal article published in the Richmond Journal of Law and Technology in 2011 strongly established that, with significantly less effort, TAR can (and does) produce a better result than linear manual review [17].

TAR processes typically use one of three protocols to select documents for review: Simple Passive learning (SPL), Simple Active Learning (SAL), or Continuous Active Learning (CAL). The workflows and a comparison of the three protocols will be discussed in Section 2.2. Typically, Continuous Active Learning (CAL) with relevance feedback outperforms all other TAR approaches in terms of overall performance [8]. This relevance feedback, hybrid, human-computer assessment, component entails human involvement to guide the CAL system and label documents as relevant or non-relevant, to achieve a higher recall and precision than the reviewer alone. The Baseline Model Implementation (BMI), an augmented version of the Continuous Active learning algorithm used as the baseline for the 2011 Legal Track and the 2015 and 2016 Total Recall Tracks was shown to be the best "high-recall" IR tool and remains the method to beat [31, 32, 21].

The ultimate goal of an eDiscovery process is to identify substantially all and only the relevant documents in a collection, achieving as close as possible to 100% recall and 100% precision [21]. In reality, the fallibility of human relevance judgements limits the capacity of the CAL system to achieve this goal. Even if it were possible to assess every document in the collection, a certain percentage of the evaluations would be inaccurate, resulting in less than 100% recall and 100% precision. Relevance assessments generated by a trained classifier would also be inaccurate, likewise falling short of 100% recall and 100% precision [13]. Since the concept of "relevance" is subjective and differs for individual reviewers, this fallibility of relevance assessments persists to date [27]. The TREC 2016 Total Recall Track coordinators hypothesized that this ambiguity in human relevance judgements restricts the ability to measure advances beyond what CAL systems (e.g. BMI) have accomplished [32].

Previous work by Ellen M. Voorhees, indicates that considering how similar the reviewers from the same background can be, it was somewhat surprising that even their feedback overlap was usually lesser than 50%; thereby showing evidence of the consistent variability in relevance judgements [34]. Our work seeks to leverage the uniqueness in each reviewer's feedback and mitigate the fallibility in human relevance feedback to the extent possible by studying the effectiveness of different multi-user, hybrid, human-computer assessment strategies for high-recall information-retrieval systems. TAR systems aid in addressing the key challenges of handling large volumes and the variety of legal data. During the course of our work, we aspire to alleviate some of the other pressing challenges in eDiscovery namely, limited availability of resources and declining budgets.

In this research, we study six different multi-user-based, hybrid, human-computer assessment strategies. We compare these six review techniques to a single user-based, hybrid, human-computer assessment approach on a level playing field, i.e., utilizing the same budget for all review strategies. As a result, the goal of this study is to answer the following question:

Can certain multi-user–based, hybrid, human-computer assessments yield higher recall and higher precision than single-user-based, hybrid, human-computer assessments, while employing the same review budget?

The answer is **"Yes"!**
The remainder of this thesis will explain how we know this to be the case and how this improvement in performance can be achieved.

1.1 Thesis Outline

The outline of the rest of this thesis is sketched out as follows.

In Chapter 2, we cover some background and related work. We discuss the TREC Total Recall Track, conducted in 2015 and 2016, the workings of various TAR protocols, the Baseline Model Implementation (BMI), feedback error in technology-assisted review, statistical tools and evaluation measures, and previous work done in line with this research.

In Chapter 3, we describe the dataset leveraged in our study and the method used to model user feedback. We also discuss in detail the design and implementation of our experiment, along with the review budget constraint used to provide a level playing field for all the review strategies.

In Chapter 4, we compare the results obtained from the different review strategies used and discuss the efficiency of the various multi-user review strategies in detail.

Finally, in Chapter 5, we conclude by discussing the results of our study, its limitations, and future efforts that can build on this work.

Chapter 2

Background and Related Work

2.1 TREC Total Recall Track

A Text REtrieval Conference (TREC), organized by the National Institute of Standards and Technology (NIST), is structured into Tracks, or areas of interest, where specialized retrieval tasks are undertaken. The Tracks are designed for a wide range of applications. First, Tracks serve as incubators for new research areas: the first running of a Track predominantly clarifies the identified information retrieval problem, and a Track establishes the required infrastructure (test collections, evaluation techniques, and so on) to enable study of the information retrieval (IR) application of interest. eDiscovery was one of the TREC Total Recall Track's most important applications. The key objective of the Total Recall Track was to evaluate strategies for achieving exceptionally high recall–as near to 100 percent as possible–with a human assessor in the loop, using controlled simulation [31], consistent with eDiscovery's goal of achieving high recall and high precision with reasonable effort.

A Web server hosted by the Track coordinators included the document collection, topic queries, and automated relevance assessments. The participants in the Track were given the task of identifying documents for review, while the Web server functioned as a real-time human-in-the-loop assessor. To meet this requirement, participants had to submit either a fully automated ("automatic") or semi-automated ("manual") process to download the datasets and topics, as well as submit documents for assessment to the Web server. A Baseline Model Implementation (BMI), an augmented version of the continuous active learning algorithm, was made available for participants to assist them in developing their IR tools and to establish a baseline for comparison [30].

The 2015 At-Home collections were comprised of three datasets and a total of 30 topics. The Track coordinators gathered the emails of Jeb Bush and assessed them on ten different topics. From the Dynamic Domain datasets, the Total Recall coordinators developed the "Illicit Goods" and "Local Politics" datasets, each with ten topics. The TREC 2015 Total Recall Track results reveal that several of the participants' approaches obtained very high recall and very high precision across all datasets, meeting the standards set by earlier TREC tasks. The Track coordinators resumed the Total Recall Track in 2016 after observing promising results from the participants in TREC 2015, intending to generate new prospects for future research.

The TREC 2016 Total Recall Track leveraged the same set-up as the previous year and introduced a new dataset called "At-Home4". This dataset extended the document collection of Jeb Bush's emails and presented 34 new topic queries for the participants. Surprisingly, no run in TREC 2015 or TREC 2016, whether manual or automated, was able to achieve greater recall, with lesser effort, than the provided BMI system. To explain this observation, the Track coordinators posited that uncertainty in human relevance assessments restricts the capacity to evaluate advances beyond those achieved by BMI. They concluded by stating that when a majority vote of three assessors is used to establish relevance, rather than a single assessor, recall increases substantially. This observation inspires more research into the impact and potential benefits of deploying multiple users in a hybrid human-computer assessment to limit the uncertainty in relevance assessments [32].

2.2 TAR Protocols

Over the years, the TREC Total Recall Track participants and eDiscovery service providers assisting producing parties in litigation have employed various Technology-Assisted Review (TAR) protocols to achieve high recall information retrieval. This section discusses in detail the different workflows and compares the three major TAR protocols, namely, Simple Passive Learning (SPL), Simple Active Learning (SAL), and Continuous Active Learning (CAL). These three protocols determine how the machine-learning algorithm is used to identify documents for review by the user [8].

SPL protocols begin the training process using randomly selected documents for the user to review. The initial training set is typically referred to as the "seed set," but the term may also be used to refer to the entire training set in an SPL process. SPL involves training the machine learning model until the effectiveness of the training is deemed to be sufficient. SPL protocols typically use ad-hoc sampling methods as the basis for determining when to stop training the algorithm. Using this learning algorithm, the system ranks the documents in decreasing order of relevance; after the training process is complete, subject matter experts (users) review the relevance of the ranked documents [18].

In SAL, the users initially start by tagging the seed-set documents to enable the machine-learning algorithm to learn the classification of what is relevant. The machine-learning algorithm uses "Uncertainty Sampling" to suggest documents for the user to review from those which the algorithm will learn the most [33]. This typically consists of documents that are on the borderline of relevance. A "control set" is used as an "answer key" to determine if "stabilization" of the learning algorithm has been achieved. Stabilization conveys that further training will not improve the effectiveness of the algorithm. SAL protocols use the randomly selected control set to Track the progress of the review. Learning stops based on the accuracy of the algorithm's predictions for the documents in the control set [9].

CAL typically begins with a judgmental rather than a random sample of documents. After the seed document(s) is/are fed into the machine-learning algorithm, the algorithm suggests the next most-likely relevant as-of-yet unreviewed document(s) for the user to review. CAL is similar to a web-search engine, at the outset, providing the documents that are most likely to be of interest first, then those that are less likely to be of interest. But unlike most search engines, CAL continuously refines its understanding of which of the remaining documents are most likely to be of interest based on the user's feedback on the documents already presented. The algorithm stops when the user decides there are few more relevant documents to be found such that the cost of additional review outweighs its benefit [8].

Several time-consuming and complicated steps associated with TAR are absent from CAL, including diligent creation of the seed set, deciding when to stop the training, and identification and assessment of large random control sets, training sets, or validation sets. Additionally, CAL has yielded superior results while requiring significantly less review effort than the other TAR protocols [8]. CAL will produce the best possible results only if the TAR tool utilises a cutting-edge learning algorithm [19]. Support vector machines and logistic regression methods have been shown to be particularly beneficial for TAR when it pertains to supervised machine learning algorithms [9]. The Baseline Model Implementation (BMI), used for the TREC 2015 and 2016 Total Recall Tracks implemented the Continuous Active Learning Protocol and used the state-of-the-art logistic regression model provided by Sofia-ML as the underlying machine-learning model. BMI extended the autonomous version of the CAL protocol through the elimination of topic-specific and dataset-specific tuning parameters [10]. BMI's consistently superior performance, reliability, and autonomy motivated us to leverage this IR tool in our study, thereby making this work easily reproducible for other datasets and future experiments.

2.3 Baseline Model Implementation

The IR tool employed in this work, BMI [30], is an enhanced version of Cormack and Grossman's CAL approach, known as AutoTAR [10]. Although the BMI system implemented the AutoTAR algorithm, it utilised Sofia-ML as its base classifier, whereas AutoTAR incorporates $\text{SVM}^{\text{light}}$. The principal advantages of using BMI over AutoTAR are that it is published under an open-source license and that it has a run time complexity of $\text{O}(\text{N} \log \text{N})$, where N is the size of the document collection. When applied to the same datasets, studies show that BMI's Sofia-ML obtained a relatively significant improvement over AutoTAR's $\text{SVM}^{\text{light}}$ in terms of retrieval effectiveness [12].

In this study, we use the canonical version of BMI made publicly accessible for TREC Total Recall participants to conduct our experiments because we completely simulate the user feedback during review. The "simulated" user feedback is modelled by inducing errors into the ground-truth relevance assessments released by the TREC Total Recall Track coordinators for the respective datasets. Algorithm 1 provides a comprehensive description of the BMI algorithm [12] and Fig. 2.1 depicts the data flow in the BMI system.

Algorithm 1 BMI Algorithm

1: Find a relevant seed document using ad-hoc search, or construct a synthetic relevant document from the topic description.
2: The initial training set consists of the seed document identified in step 1, labeled "relevant."
3: Set the initial batch size b to 1.
4: Temporarily augment the training set by adding 100 random documents from the collection, temporarily labeled "not relevant."
5: Construct a logistic regression classifier from the training set.
6: Remove the random documents added in step 4.
7: Select the highest-scoring b that have not yet been reviewed.
8: Review the documents, labeling each as "relevant" or "not relevant."
9: Add the documents to the training set.
10: Increase b by $\lceil \frac{b}{10} \rceil$
11: Repeat set 4 through 10 until a sufficient number of relevant documents have been reviewed.

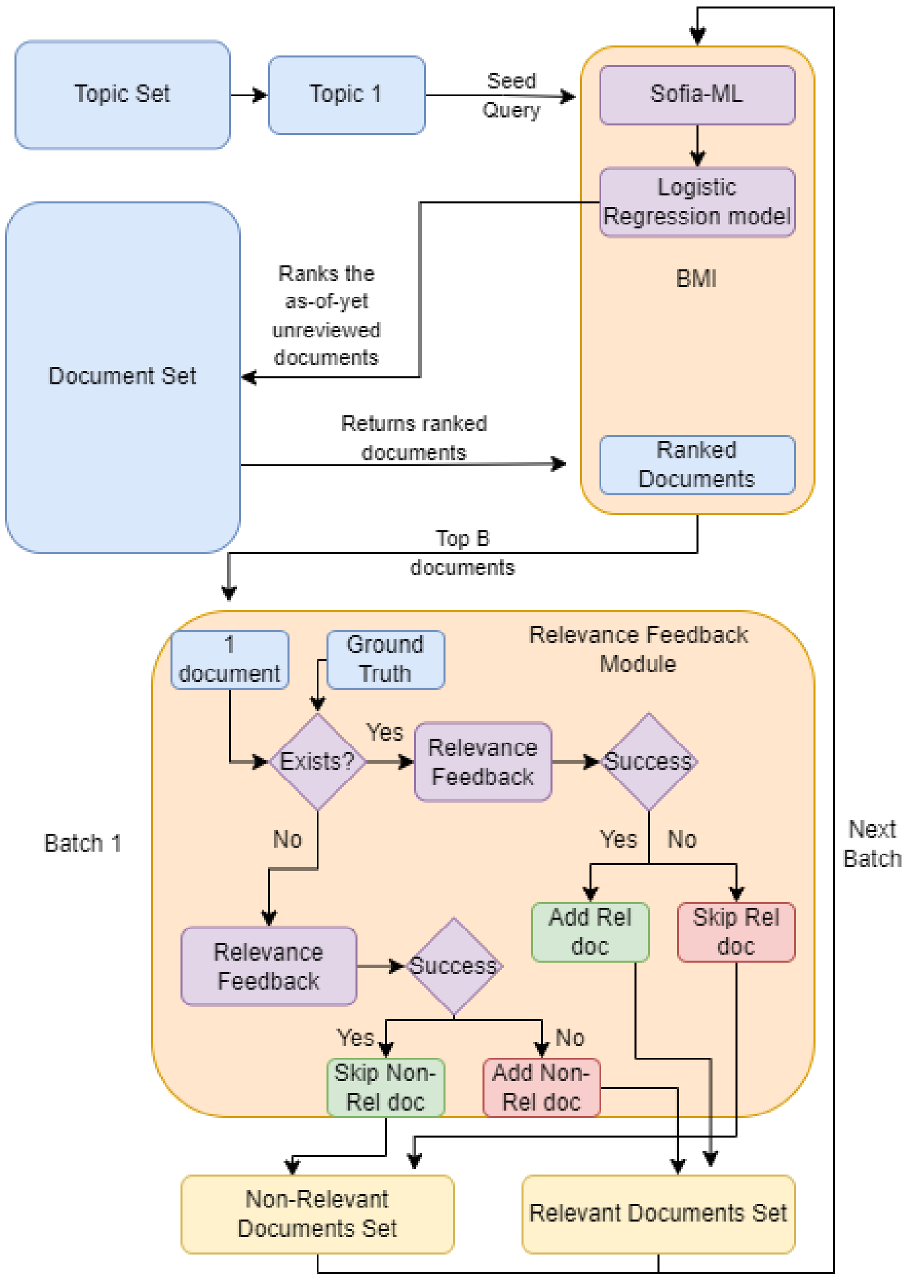


Figure 2.1: Baseline Model Implementation (BMI) Architecture

To make use of the simulated BMI system, we need three individual data components:

- Topic queries, which describe the user's information need(s).
- Document collection, from which we seek to satisfy our information need(s).
- Simulated relevance feedback, used to review the system-presented documents.

The seed document is the initial document/document set used to train the machine-learning algorithm, and; for our experiments, we use the topic query itself as the seed document. After the seed document is fed into the machine-learning algorithm, the algorithm suggests the next-most-likely relevant as-of-yet unreviewed document(s) for the user to review. BMI presents the most-likely documents for review with exponentially increasing batch sizes (b), starting with size 1. The relevance feedback module for batch 1 has been incorporated. The "simulated reviewers" provide relevance feedback to that batch of documents and the machine learning algorithm learns from the relevance feedback to provide the next batch of most-likely relevant documents.

Usually, the algorithm stops when the user decides there are few more relevant documents to be found, such that the cost of additional review outweighs its likely benefit. In our experiments, we formulate the review budget (B) to stop the automated relevance feedback mechanism and to provide a level playing field for evaluating our hybrid human-computer strategies; as a result, the "simulated reviewers" can only review until the review budget is exhausted. The review budget (B) is determined from the number of relevant documents (R) in the document collection for the specified query topic. A detailed study of the review budget will be discussed in Section 3.3.1. Finally, when the review budget for one topic is spent, the BMI system picks the next topic from the topic collection as the next query and repeats the process.

2.4 Feedback Error in Technology-Assisted Review

"Garbage in, garbage out," the premise that flawed inputs yield flawed outputs, is one of the foundational principles of computer science [15]. No reviewer is perfect; in other words, the ground-truth/gold-standard set by one person may not be exactly the same as another person's ground-truth/gold-standard set. Cormack et al. discovered that even when the same user reviews the same topics at different times, there is not a perfect overlap in relevance assessments [13]. Because the concept of relevance can vary significantly between same/different reviewers; this provides compelling evidence that relevance feedback errors will continue to exist.

When users provide their relevance feedback during the review process, their feedback can be classified into four categories: True Positives (TPs), True Negatives (TNs), False Negatives (FNs), and False Positives (FPs). True Positive documents are ground-truth relevant documents that have been marked as "relevant" by the user. A True Negative document is a ground-truth non-relevant document that has been identified as "not-relevant" by the user. A False Negative document is one in which the user marks a ground-truth relevant document as "not relevant," while a False Positive document is one in which the user tags a ground-truth non-relevant document as "relevant." TP and TN are correct classifications, while FN and FP represent human relevance feedback errors with respect to the ground truth.

Figure 2.1 illustrates the reviewer's correct classifications in green and the incorrect classifications in red. TP is represented by "Add Rel doc," TN by "Skip Non-Rel doc," FN by "Skip Rel doc," and finally FP by "Add Non-Rel doc". The two types of feedback errors, FNs and FPs, can impede the TAR system from achieving high-recall. When FNs are introduced into the feedback system, those ground-truth relevant documents are lost in the process. The system will miss learning the relevant parameters from those documents, and can be expected to miss other, similar documents to those relevant documents. As a result, the recall of the TAR operation may be reduced. Conversely, when FPs are induced into the feedback system, those ground-truth non-relevant documents are included as relevant documents and the system will incorrectly learn the non-relevant parameters as relevant; and can be expected to identify other, documents similar to those non-relevant documents. Therefore, introducing FPs into the system can decrease the precision of the TAR operation.

In sum, these assessment errors can negatively impact the performance measures discussed in Section 2.5.1 and all three types of retrieval techniques discussed in Section 2.6.

2.5 Statistical Tools

2.5.1 Performance Measures

As outlined in the introduction, legal teams conduct eDiscovery to find substantially all relevant documents pertaining to their search topic, with reasonable effort. A missed relevant document in eDiscovery can result in potential legal sanctions, thus attaining a high recall for the search topics is critical. As a result, recall, which is the fraction of all relevant documents found by an assessor, is a valuable baseline performance measure to consider:

$$recall = \frac{|U_{rel} \cap R|}{|R|} \tag{2.1}$$

where $\mathrm{U_{rel}}$ is the set of documents assessed to be relevant by the assessor, and R is the true set of relevant documents. $\mathrm{S_{rel}}$, in contrast to $\mathrm{U_{rel}}$, denotes the documents that the system believes to be relevant to a search topic. We will review these distinctions in depth in Section 2.6, since this distinction leads to different recall measures depending on whose recall we are calculating. We chose to measure "recall" rather than the absolute number of relevant documents found by the assessor because different search queries (topics) have different number of relevant documents, and recall normalizes this variability.

eDiscovery review tasks frequently involve two passes of relevance judgments, the first pass by a junior lawyer or contract attorney who is qualified to broadly identify relevant information (i.e., a reviewer that is less expensive), then by a more senior lawyer (i.e., one that is more expensive), who typically reviews only the documents marked as "potentially relevant" during the first pass to make final decisions on the document's relevance, confidentiality and/or privilege [38]. An article "The Truth About Doc Review" written by a junior associate, Mary Kate Sheridan, states that the first pass is usually done by an army of junior associates spending several weeks looking at the universe of documents and the second pass is typically performed by senior associates. In these instances, each non-relevant document that makes it beyond the first pass wastes more of the expensive attorney's time; thus, our second metric addresses precision, which is the proportion of relevant documents recognized by the assessor that is truly correct:

$$precision = \frac{|U_{rel} \cap R|}{|U_{rel}|} \tag{2.2}$$

Precision is a measure of error caused by False Positives (FPs), whereas recall is the measure of error caused by False Negatives (FNs). Precision and recall are two different ways of measuring a model's predictive capability. To perform well, an IR system should have minimal FPs and FNs, thereby maximizing both recall and precision. In practice, the trade-off between precision and recall makes it difficult to maximize both at the same time. Increasing one metric will inevitably result in a decrease in the other. As a result, it is useful to have a metric that incorporates both parameters, and the F1 score, a suitable metric that computes the harmonic mean of both precision and recall, is used in this study.

$$F_1 = \frac{2 \times recall \times precision}{recall + precision} \tag{2.3}$$

2.5.2 Statistical Testing

Statistical hypothesis testing is widely used to examine whether the mean performance of one group outperforms that of another. We perform the Paired Two-Sample T-Test for this set of experiments because the two population groups we compare (topic-wise relevance) are independent of one another and are approximately normally distributed. In this study, we seek to find if there is a significant statistical difference in favor of a multi-user review strategy compared to a single-user review strategy. Without having formed prior hypotheses, the two-tailed test is appropriate for determining whether the performance of a multi-user review strategy is consistent for all datasets when compared to that of the single-user review strategy. It yields the same conclusion as the one-tailed test but with lower confidence [5].

To perform the T-Test computation, we first calculate $\mathrm{S_p}$, the pooled estimate of common standard deviation, using the following equation:

$$S_p = \sqrt{\frac{(m_1 - 1)s_1^2 + (m_2 - 1)s_2^2}{m_1 + m2 - 2}} \tag{2.4}$$

The T-Test computes the confidence interval (T) of the difference in means between two populations using the test statistic shown below:

$$T = \frac{m_1 - m_2}{S_p\sqrt{n_1^{-1} + n_2^{-1}}} \tag{2.5}$$

where

- m_i: mean of population i
- n_i: size of population i
- s_i: sample standard deviation of population i

Then, using the value T from the topic-wise T distribution with significance level to α and degree of freedom ν, we can construct an 1- α confidence interval

$$\left[(m_1 - m_2) \pm tS_p\sqrt{n_1^{-1} + n_2^{-2}}\right] \tag{2.6}$$

which estimates where the true difference between the means may lie. If zero exists in this interval, the null hypothesis (i.e., that the two review strategies produce the same results) cannot be rejected, and there is no statistical significance between the means of the two populations. This study has the following null hypothesis:

$$H_0 : m1 = m2 \tag{2.7}$$

and the following alternate hypotheses:

$$H_1 : m1 \neq m2 \tag{2.8}$$

where m1 corresponds to the mean of the single-user review strategy and m2 corresponds to the mean of a multi-user review strategy.

2.6 Evaluation Measures/Retrieval Types

TREC, CLEF, and NTCIR are notable examples of research programs that use the Cranfield evaluation paradigm. Cranfield researchers utilize test collections to evaluate the relative effectiveness of different retrieval techniques [36]. The two most common types of evaluation are system evaluation and user-based evaluation. System evaluation measures how accurately the *system* ranks documents, whereas user-based evaluation assesses how satisfied *users* are with the IR system [35]. Extending the core Cranfield evaluation approaches, Cormack and Grossman presented another retrieval type measuring the end-to-end retrieval effectiveness of an interactive IR process [13]. In this section, we will discuss in detail these three important retrieval evaluation measures used in real-time applications.

2.6.1 System Retrieval

System retrieval is the set of all the "next-most-likely" documents presented by the IR system (BMI) to the user seeking relevance feedback. It is imperative to note that a document is regarded to be considered relevant in the calculation of System Recall when it is presented to the user, regardless of the user's final relevance judgment. For the set of documents retrieved by IR system S_{ret}, the System Recall ($Recall_{sys}$) and System Precision ($Precision_{sys}$) are defined as below:

$$Recall_{sys} = \frac{S_{rel}}{R} \tag{2.9}$$

$$Precision_{sys} = \frac{S_{rel}}{S_{ret}} \tag{2.10}$$

where S_{rel} is the set of relevant documents retrieved by the IR system, and R is the set of relevant documents in the ground-truth set.

2.6.2 User Retrieval

User retrieval is the set of all documents marked as relevant by the user, following the IR system's effort. The User Recall ($Recall_{user}$) and User Precision ($Precision_{user}$) are calculated using the following equations:

$$Recall_{user} = \frac{U_{rel}}{S_{rel}} \tag{2.11}$$

$$Precision_{user} = \frac{U_{rel}}{U_{ret}} \tag{2.12}$$

where U_{rel} is the set of relevant documents marked as relevant by the reviewer, and U_{ret} is the set of documents that the user marked as relevant irrespective of whether they are considered relevant in the ground-truth set.

2.6.3 End-to-End Retrieval

End-to-End retrieval is the set of all documents that are presented by the IR system for review and also marked as relevant by the user. The End-to-End Recall ($Recall_{e2e}$) and End-to-End Precision ($Precision_{e2e}$) are calculated using the following equations:

$$Recall_{e2e} = \frac{E_{rel}}{R} \tag{2.13}$$

$$Precision_{e2e} = \frac{E_{rel}}{E_{ret}} \tag{2.14}$$

where E_{rel} is the set of truly relevant documents retrieved by the IR system and marked relevant by the reviewer (identical to U_{rel}), R is the total number of relevant documents in the ground truth, and E_{ret} is the set of documents retrieved by the IR system and marked relevant by the reviewer, regardless of their true relevance (identical to U_{ret}).

The end-to-end retrieval effort is determined by how successfully the system retrieves relevant documents and how well the users correctly code the relevance of the documents. Therefore, End-to-end Recall is numerically equal to the product of System Recall and User Recall, whereas End-to-End Precision is equivalent to user precision, as the document sets E_{rel} and E_{ret} are identical to U_{rel} and U_{ret} respectively. Extending these observations, we describe equations 2.15 and 2.16, which are used to validate our retrieval efforts.

$$Recall_{e2e} = Recall_{sys} * Recall_{user} \tag{2.15}$$

$$Precision_{e2e} = Precision_{user} \tag{2.16}$$

Reiterating on the e-Discovery goal, we must identify substantially all relevant documents pertaining to the search topic with reasonable effort. As a result, we focus mainly on analysing the TAR system's end-to-end retrieval performance, as it seeks to achieve high recall without compromising much on the precision. System retrieval and user retrieval performance are just intermediate results. Furthermore, when determining the superiority of a given TAR technique, the primary metric to consider is the *end-to-end recall* (since end-to-end precision only needs to be reasonable).

2.7 Related Work

The concept of "relevance" is ambiguous, and various assessors–or even the same assessor at different times–may make conflicting relevance assessments for the same document, regardless of their knowledge and skill, or the detail with which relevance is defined [1]. In an experiment conducted by Voorhees, when TREC AdHoc documents (from TREC-4 and TREC-6 datasets [37]) were reviewed by two independent assessors, substantial levels of assessor disagreement were reported. Voorhees' work at TREC on assessor agreement has sparked numerous subsequent investigations into task-specific agreement [34]. To investigate the effect of the assessor errors in IR evaluation, Carterette and Soboroff conducted a TAR-based simulation experiment and discovered that overly-conservative assessors (those who find fewer documents relevant) have a lower impact on retrieval effectiveness rating than the liberal ones [7]. Webber investigates assessor agreement levels on various datasets and provides results indicating the significant variations in assessor reliability [39].

With significantly less effort, TAR can (and does) produce better retrieval results than review by assessors alone without the aid of TAR tools [17]. Continuous Active Learning (CAL), a TAR protocol with relevance feedback, outperforms all other TAR approaches and manual approaches in terms of overall performance [8]. Although CAL tools can efficiently assist users by providing the next "most-likely" relevant document for review, if the relevance feedback provided in response to the system-presented document is erroneous, the task of achieving high-recall across all three retrieval types becomes challenging.

Several strategies have been examined in studies seeking to reduce the two types of human relevance feedback errors during TAR (i.e., false negatives and false positives). To name a few, Brodley and Friedl propose strategies for detecting outliers by automatically identifying mislabeled training data using ensemble classifiers [3]. Similarly, Ramakrishnan et al. employ a Bayesian network to detect outliers in textual data [29]. Such strategies, however, are ineffective if the assessor is consistently erroneous. As a consequence, research efforts are presently focused on providing the system with the most authoritative/reliable relevance feedback in order to eliminate false negative and false positive errors to the maximum extent possible.

As alluded to in Section 2.5.1, engaging an experienced attorney to re-review the documents tagged as "potentially relevant" by a junior or contract attorney is used to reduce relevance feedback errors and this involves providing reliable feedback by the more experienced attorney [38]. But this approach reduces only false positives, and if a relevant document has been missed by the junior or contract attorney, it is lost in the universe of documents forever. To supplement this approach, the experienced attorney may randomly sample the documents tagged as "potentially non-relevant" by the junior or contract attorney, but when the prevalence of relevant documents is low to begin with, and even lower following the review, this method is unlikely to be terribly effective or efficient.

The TREC Legal Track coordinators employed a two-pass review approach to obtain the ground-truth assessments for the test collections. Initially, they engaged first-pass assessors , who were equipped with detailed assessment guidelines, to provide the first-pass relevance assessments. The first-pass assessors included both professional document reviewers and individual volunteers [20]. Then, this batch of first-pass relevance feedback for documents was provided to the Track participants and they were asked to review those assessments and to reach out to a Topic Authority (TA)–SMEs with respect to those query topics–for final adjudication where the participants disagreed with or challenged the first-pass relevance assessment [21]. If the TA codes the responsiveness of the documents different from the first-pass assessments, then the TA's feedback was used to improve the quality of the ground-truth and at the same time help participants achieve higher recall and precision, provided their challenge was sustained. In an attempt to enhance the quality of the ground-truth, the Legal Track coordinators in effect adopt a Majority-Vote-of-Three review strategy, where the three participants are: the first-pass assessment, participants' relevance feedback and the Topic Authority's relevance feedback.

Observing promising results from the participants in the TREC Total Recall Track 2015, the TREC Track was conducted in the year 2016 as well. In 2016, although the participants were provided with the IR system (BMI) used by the Track coordinators, none of the participant submission was able to achieve greater recall, with less effort, than the baseline BMI system. The Track coordinators explained this observation, in part, by stating that when a Majority-Vote-of-Three assessors review strategy, used for the baseline, is used to establish relevance rather than the single-assessor review strategy used by most participants, recall increases substantially [32]. Although the Majority-Vote-of-Three review strategy's results are impressive, it is an expensive review strategy, as three high-expertise level reviewers are asked to spend time on the same set of documents, so this strategy would be unlikely to be employed in practice.

In 2017, Grossman and Cormack published a research paper proposing that, to build—as well as evaluate—TAR systems with near-perfect recall and precision, it is imperative to model human assessment as an indirect indicator of the amorphous property known as "relevance" [13]. In other words, to considerably reduce relevance feedback errors, relevance judgments must be modelled and provided to the system, comparably to that of a perfect assessor. This principle is consistent with the fundamental strategy employed in the TREC Legal and Total Recall Tracks for ground-truth construction. Therefore, Grossman et al. strive to enhance the "quality" of the relevance feedback by modelling the relevance judgments.

Quality is a measure of the extent to which a TAR method can find as much relevant information as possible with reasonable effort [11]. Reviewers must provide relevance feedback roughly comparable to that of the Ideal User for relevance judgements to be of high quality; in other words, the relevance feedback should contain little to no false negative or false positive errors. Quality-Control ("QC") techniques use one or more supplemental assessments for some or all of the documents in an effort to minimize the impacts of erroneous relevance assessments [2]. The Majority-Vote-of-Three review strategy is a paradigm of a TAR approach with quality control. The Topic Authority (TA), being the adjudicator, provides quality control in the form of supplement relevance assessments when the "first pass" judgement and the participant's relevance feedback disagree.

Cormack et. al.'s recent study extends this "quality control" approach by providing a subset of the documents for further adjudication, either by the user or another assessor, rather than using expensive SMEs, to reduce the fallibility of the user's original evaluations [13]. With positive findings, they conclude that when greater recall and precision are considered necessary, additional resources should be spent re-reviewing documents that the user may have misjudged rather than reviewing the ranked list to extreme depths or sampling low-ranked documents. In this work, we propose two novel quality control review strategies, CAL with Quality Control–Type 1 and CAL with Quality Control–Type 2, inspired from Cormack et. al.'s work [13], and compare it with the existing review strategies to determine whether our proposed quality control review strategies are effective.

Chapter 3

Study Design

3.1 Corpus and Topics

Our goal of achieving high recall closely aligns with the goal of the TREC Total Recall Track, therefore we leverage some of the datasets used in the Track across the years 2015 and 2016. The Track coordinators made the datasets publicly available to participants through the Total Recall Track's web server. For this research, we leverage the At-Home series of datasets that was released in both years.

Three datasets and 30 topics were drawn from the 2015 At-Home collections. The Track coordinators collected and analyzed the Jeb Bush emails for ten topics to formulate the At-Home1 dataset. The "Illicit Goods" (At-Home2) and "Local Politics" (At-Home3) datasets were derived datasets used for another TREC Track, Dynamic Domain, and evaluated by the Total Recall Track coordinators. The Track coordinators constructed the At-Home4 dataset in 2016 using the same Jeb Bush email collection as the At-Home1 dataset, but with 34 new topics. To obtain the ground-truth for the new topics, the Track coordinators re-evaluated the documents using Cormack and Mojdeh's CAL technique [14] to discover as many relevant documents for each topic as feasible, with reasonable effort.

At-Home1 Dataset: This document collection included 290,099 redacted emails spanning Jeb Bush's eight-year term as the Governor of Florida. The Track coordinators chose 10 major issues associated with his governorship as themes for At-Home1 test collection: "school and preschool funding," "judicial selection," "capital punishment," "manatee protection," "new medical schools," "affirmative action," "Terri Schiavo," "tort reform," "Manatee County," and "Scarlet Letter Law."

At-Home2 Dataset: This refers to the "Illicit Goods" dataset collected for the TREC 2015 Dynamic Domain Track [40]. It is comprised of 465,147 documents extracted from Blackhat World and Hack Forum. This dataset includes the following ten topics: "paying for Amazon book reviews," "CAPTCHA services," "Facebook accounts," "surely Bitcoins can be used," "Paypal accounts," "using TOR for anonymous browsing," "rootkits," "Web scraping," "article spinner spinning," and "offshore Web sites."

At-Home3 Dataset: This refers to the "Local Politics" dataset collected for the TREC 2015 Dynamic Domain Track [40]. It is comprised of 902,434 articles aggregated from news networks in the northwest United States and southwestern Canada. This dataset includes the following ten topics: "Pickton murders," "Pacific Gateway," "traffic enforcement cameras," "rooster chicken turkey nuisance," "Occupy Vancouver," "Rob McKenna gubernatorial candidate," "Rob Ford Cut the Waist," "Kingston Mills lock murder," "fracking," and "Paul and Cathy Lee Martin".

At-Home4 Dataset: This dataset uses the same Jeb Bush email document collection comprising of 290,099 emails as used in the At-Home1 Dataset, but it consists of 34 distinct topic queries developed for the 2016 Total Recall Track. The topics span from local concerns like "Traffic Cameras" and "New Stadiums" to global agendas like "War Preparations" and "Space Programs". Figure A.1 provides a detailed description of the 34 topics.

3.2 Modelling User Feedback

To conduct our experiments, we require human assessors in the loop to review the documents presented by the BMI system. We desire to obtain relevance feedback from a realistic (non-perfect) reviewer with reasonable error percentages, which is both convenient to employ and feasible within the budget of a TAR operation. According to Cormack and Grossman's research on navigating imprecision in relevance assessments, to build, as well as evaluate, TAR systems that approach 100% recall and precision, it is necessary to model human assessment as an indirect indicator of the amorphous property known as "relevance" [13].

Thus, in this work, we "simulate" human reviewers and this simulation accommodates a broad spectrum of reviewer expertise by designing user relevance feedback according to their recall and precision rates. According to the literature, TAR methods that use relevance feedback can achieve far more than the 65 percent recall and 65 percent precision claimed by Voorhees as the "realistic upper bound on retrieval performance...because that is the level at which humans agree with one another" [34]. As a corollary, we simulate

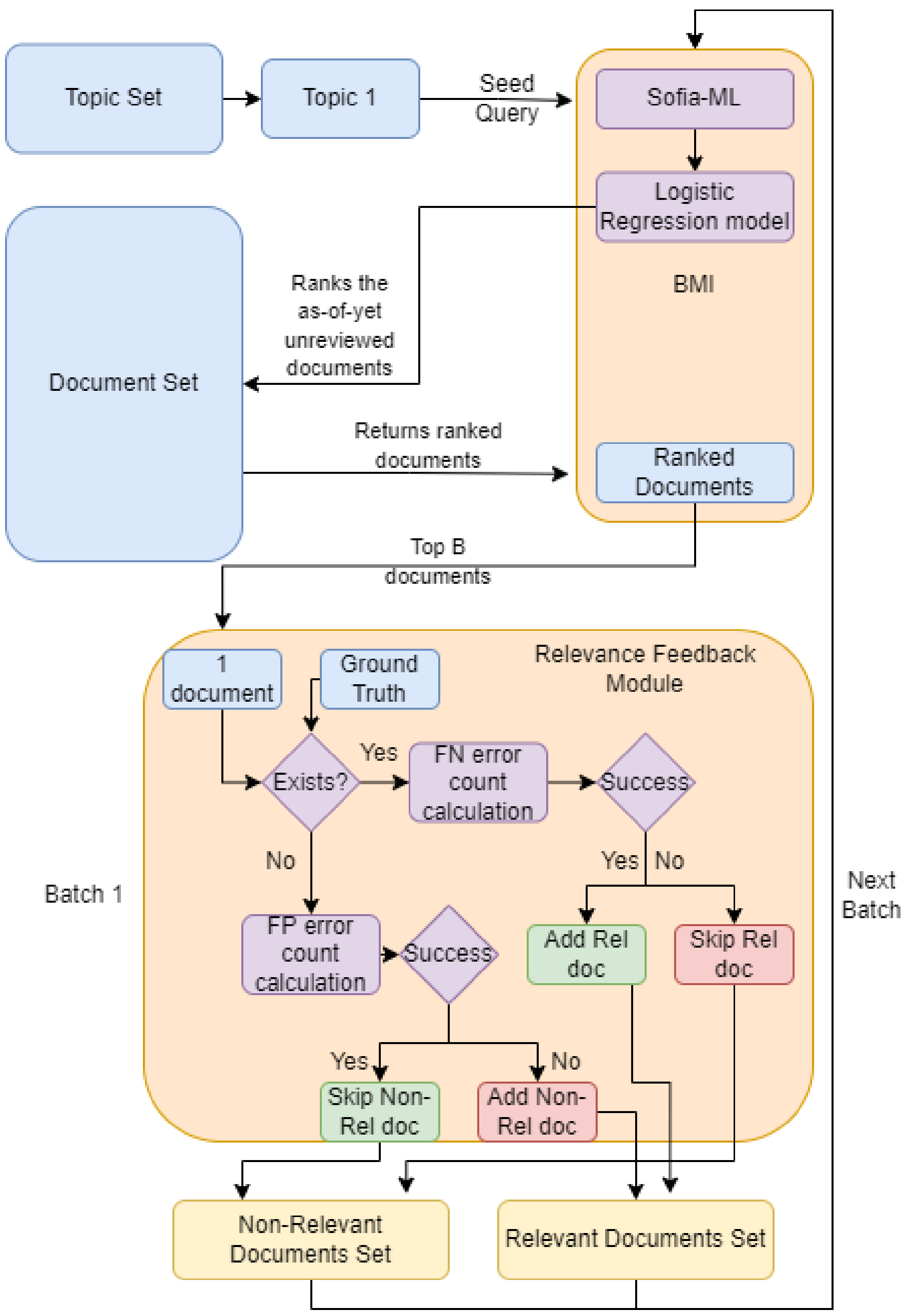


Figure 3.1: Error-Induced Baseline Model Implementation Architecture

reviewers with recall and precision rates as low as 60% and as high as 100%. We can construct 25 distinct simulated reviewers with 10% increments ranging from 60% to 100% across both recall and precision rates, which encompasses practically every expertise level within the human error range reported in Voorhees study.

Ground-truth, the document set which is the "true" relevance for each topic query to the degree that can be determined, is obtained from the publicly released ground-truth relevance assessments (referred to as "qrels") in the TREC 2015 and 2016 Total Recall Track. This document set corresponds to the "gold standard," in other words the relevance feedback of the perfect user with 100% user recall rate and 100% precision rate. Given a specific, non-ideal simulated user, we formulate the user's relevance feedback by inducing the two forms of feedback errors, false negatives and false positives, on the ground–truth depending on the user's recall and precision rate. For example, a user with 70% recall and 80% precision will recognize only 70% of the system-presented relevant documents as relevant, and only 80% of the documents tagged as relevant by the reviewer will be truly relevant.

To design a simulated BMI system with user's relevance feedback errors, we introduce two new components namely, a false negative (FN) error-count calculation and a false positive (FP) error-count calculation. Figure 3.1 depicts the BMI system architecture incorporating the relevance feedback error simulation; when compared to the original BMI architecture in Figure 2.1, we can observe that the relevance feedback module alone has been modified. We independently calculate how many true positives (TPs), true negatives (TNs), false negatives (FNs), and false positives (FPs) are to be induced for each batch of system-presented documents based on the user's recall and precision rate. The false negative errors and false positive errors are induced into the feedback system randomly but the number of FN and FP errors to be induced into the feedback system is fixed based on the simulated reviewer's user recall and user precision. The formulas used to calculate the number of TPs and FNs to be induced in relevance feedback are represented by equations 3.1 - 3.4. After determining the number of TPs to be induced in a particular batch, we apply equations 3.5 - 3.9 to calculate the number of FPs and TNs to be generated in the batch.

Let us consider the following notations,

- P[b] is the number of ground-truth positives in batch b.
- P[a...b] is the number of ground-truth positives in batches a through b.
- N[b] is the number of ground-truth negatives in batch b.

- N[a...b] is the number of ground-truth negatives in batches a through b.
- TP[b] is the number of true positives in batch b.
- TP[a...b] is the number of true positives in batches a through b.
- FP[b] is the number of false positives in batch b.
- FP[a...b] is the number of false positives in batches a through b.
- UR is the User Recall rate.
- UP is the User Precision rate.

Calculating true positives and false negatives to be introduced in a batch 'b':

$$TP = UR * P \tag{3.1}$$

$$TP[1...b] = UR * P[1...b] \tag{3.2}$$

$$TP[b] = TP[1...b] - TP[1...(b-1)] \tag{3.3}$$

$$FN[b] = P[b] - TP[b] \tag{3.4}$$

Calculating false positives and true negatives to be introduced in a batch 'b':

$$UP = \frac{TP}{TP + FP} \tag{3.5}$$

$$UP = \frac{TP[1...b]}{TP[1...b] + FP[1...b]} \tag{3.6}$$

$$FP[1...b] = \frac{TP[1...b]}{UP} * (1 - UP) \tag{3.7}$$

$$FP[B] = FP[1...b] - FP[1...(b-1)] \tag{3.8}$$

$$TN[b] = N[b] - FP[b] \tag{3.9}$$

For all the 25 distinct users in our study, we separately run the simulated BMI system by inducing the pertinent relevance feedback errors in each batch and judge every document in the corpus. The relevance feedback accumulated in these 25 runs forms the relevance feedback for the 25 simulated users. For our experimentation, we chose simulated users with the same level of expertise, since it was simpler than meticulously selecting a user pool with a range of experience levels. Hence the review strategies in this study will only involve users with the same level of expertise. In the case of a 60User pool, for instance, all reviewers, regardless of the number of users involved in the review strategy, all users have a 60% user recall and 60% user precision, but this does not imply that they all review in the same way; rather, it just refers to the probability that they will precisely identify a relevant document. When implementing the TAR strategies, we use various combinations of these simulated users to provide relevance feedback. With the dataset, BMI system, and users of the BMI system described, we will now turn to the experimental setup that employs the various review strategies.

3.3 Experimental Set-Up

3.3.1 Review Budget

HRIR systems aspire to achieve the highest possible recall rate, for a reasonable amount of effort, in order to strike the right balance between thoroughness and cost [16]. Thoroughness corresponds to the retrieval of all relevant documents for a particular topic from the corpus. Cost, in the context of eDiscovery, refers to the remuneration allocated to reviewers for providing relevance feedback to the BMI system-presented documents. The cost of TAR is inversely proportional to the degree of thoroughness. Therefore, it is essential to identify when to stop the TAR strategy to avoid overspending the review budget, but at the same time achieve a reasonably high recall.

Several heuristic stopping criteria for one-phase and two-phase TAR reviews have been published previously [6, 11, 12, 24, 31, 32], with the "knee method" proving to be highly reliable and efficient even when the collection contains scant relevant documents. In this study, each review strategy requires different combinations of reviewers, therefore using the state-of-the-art "knee method" as a stopping criterion would entail allocating more review budget for certain review strategies while being unfair to other review strategies. Instead, in our experiments, we use the review budget (B) itself as a stopping criterion to ensure that each review strategy operates within the same allocated budget. The review budget (B) is determined from the number of relevant documents (R) in the document collection for the specified query topic, for example, B=R,2R,3R. The review budget was chosen in this manner because achieving the goal of 100% recall necessitates reviewing at least R documents, and we accommodate multiples of R as potential review budgets because both the users and the system are not perfectly precise, thereby requiring to review more than R documents.

The review budget specifies the permissible number of documents to be reviewed. Our principal review budget is B=3R, as it can provide a reasonable number of document reviews for all review strategies to achieve near 100% recall. Since we do not know in advance the exact number of relevant documents per query topic, using a review budget in terms of R may not be a practical strategy, but in our study, it is used as an example to show, how to accommodate a reasonable number of documents reviews to achieve high recall. Most importantly, it is an artificial constraint to better control the comparison of the review strategies. In other words, incorporating a predefined review budget like 3R allows us to compare various hybrid human-computer assessment strategies on a level playing field, because they are all limited to performing only 3R document reviews.

The use of the review budget as a stopping criterion also contributes to addressing one of eDiscovery's most critical challenges: declining budgets. With the same level of resources (i.e., review budget) allocated, we can determine whether a multi-user-based review strategy is able to attain better performance than single-user-based review strategy in terms of recall, precision, and F1 score. The quality control offered by certain multi-user-based review strategies comes at the expense of the review budget (i.e., reduced depth in user review) and therefore, these users may not even come across some of the documents that the single-user-based CAL user may have reviewed. With these considerations, we will examine, in the next section, if one or more of six multi-user review strategies are nonetheless able to outperform the single-user review strategy.

3.3.2 Review Strategies

Single-User CAL

Single-User CAL (Continuous Active Learning) is a commonly leveraged review strategy for eDiscovery, where a single user is engaged in providing relevance feedback for a budget of B documents. After the seed document(s) is/are fed into the machine-learning algorithm, the algorithm suggests the next most-likely relevant as-of-yet unreviewed document(s) for the user to review. The user reviews the documents suggested by the IR system and provides their relevance feedback for those documents.

A user review budget (i.e., limit) of 'B' is placed on the relevance feedback task as discussed in Section 3.3.1. One review budget unit corresponds to the activity of reviewing one document. In this review strategy, the user reviews each of the B documents only once, and this review strategy is used as the baseline for our study. To quantify the retrieval effort of this review strategy, all the B documents presented by the system for user review

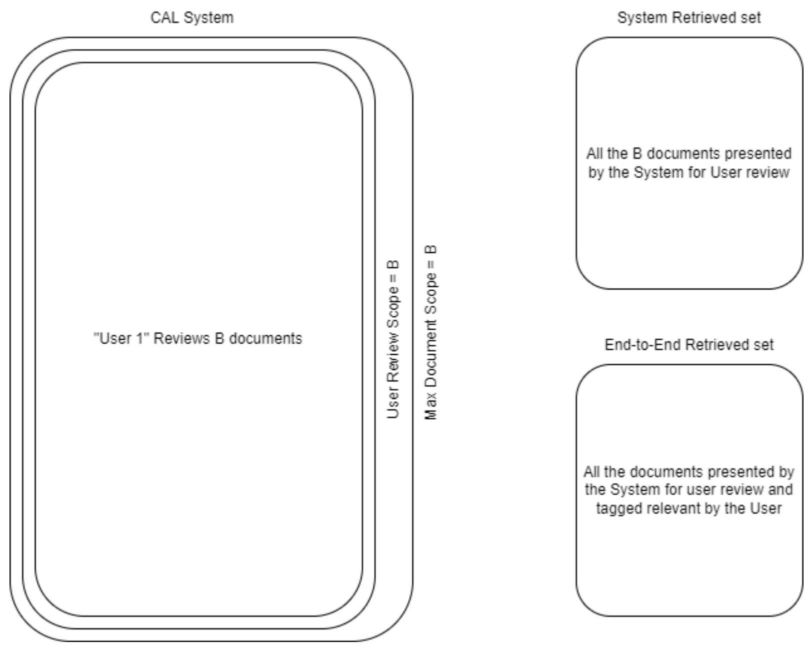


Figure 3.2: Single-User CAL Review Strategy

are considered the system-retrieved set and the documents presented by the system and tagged 'relevant' by the user are considered to be the end-to-end-retrieved set. Figure 3.2 depicts the structure of the Single-User CAL review strategy.

Separate CAL

In this review strategy, we leverage the review efforts of two users and the budget B is split equally between the two users. Each of the two users is given a separate CAL system and the same query topic for which they are supposed to find relevant documents. Both users review B/2 documents independently, starting with the same seed document, and based upon their feedback, the machine learning algorithm is trained separately. Since different users have different user recall and user precision rates, the sequence of documents from the document collection presented by the IR system will also vary for both the user's CAL runs.

Separate CAL review strategy users will fetch different document subsets from the document collection and because the feedback provided by one user does not influence the other, there is a possibility for the users to encounter new potentially relevant documents. To quantify the retrieval effort of this strategy, we combine all the documents presented

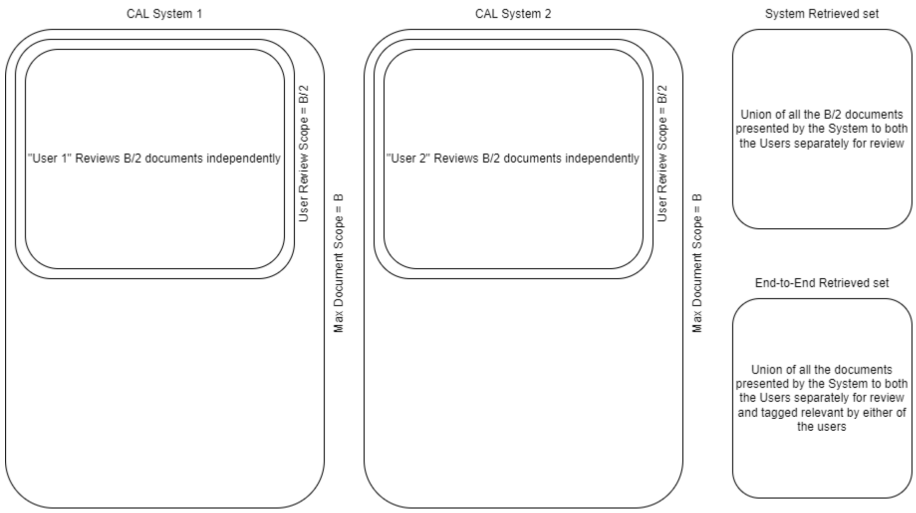


Figure 3.3: Separate CAL Review Strategy

separately by both IR systems to obtain the system-retrieved set; and we combine all the documents that were presented to the reviewers and tagged relevant by them separately to obtain the end-to-end-retrieved set. Figure 3.3 depicts the structure of the Separate CAL review strategy.

Lock-Step CAL–Type 1

In Separate CAL, the extra relevant documents found by one user are not being used by the CAL system being trained by the second user. Therefore, to understand the impact of combined-user relevance feedback we present the Lock-Step CAL–Type 1 review strategy. Like Separate CAL, the Lock-Step CAL–Type 1 review strategy also leverages the review effort of two users and the budget B is split equally between the two users. But in this review strategy, we engage both users to review the same B/2 documents together on a single CAL system. This review technique enables the single CAL system to be trained by both reviewers and potentially to capture more relevant documents as this strategy considers a document to be relevant even if only one of the two users tags it relevant.

We hypothesize that any additional relevant documents found by either of the users will help in training the system and therefore, more relevant documents can be presented by the system for the user's review in subsequent batches. To quantify the retrieval effort of this strategy, the B/2 documents presented to both users by the system are included in the system-retrieved set and the documents presented by the system and tagged relevant

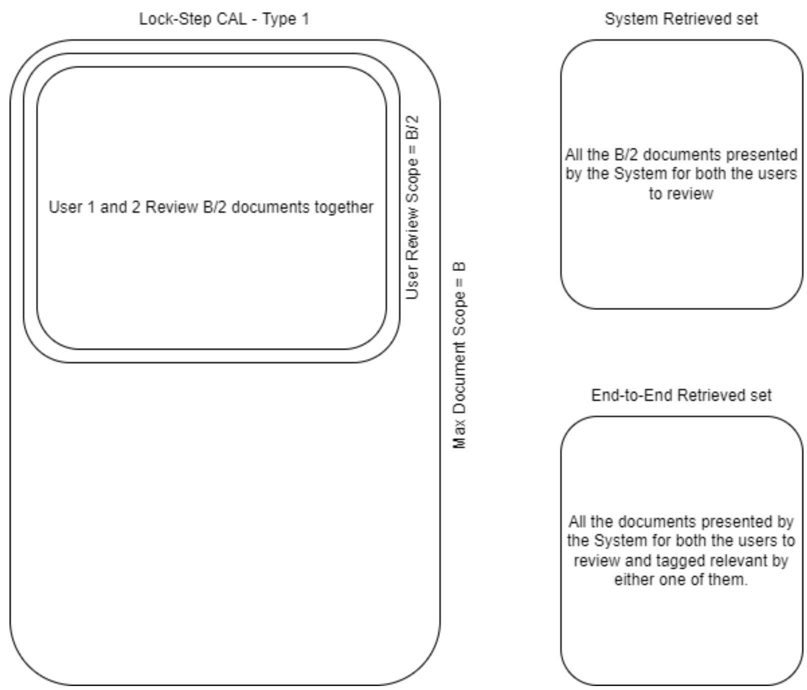


Figure 3.4: Lock-Step CAL–Type 1 Review Strategy

by either one or both users constitute the end-to-end-retrieved set. Figure 3.4 depicts the structure of the Lock-Step CAL–Type 1 review strategy.

Lock-Step CAL–Type 2

The previous technique, Lock-Step CAL–Type 1, accepts a document as relevant if either of the users tags it as relevant; as a result, both users' false positives are combined in the tagged-relevant set. The precision of a review strategy may be significantly impacted when the number of false positives in the system and end-to-end-retrieved sets is high. We aim to overcome this potential low-precision condition with respect to the end-to-end retrieval using Lock-Step CAL–Type 2. With this review strategy, we enable the system to learn both true positive documents and false positive documents tagged by both users, similar to Lock-Step CAL–Type 1's system retrieval, because those documents have been given a higher ranking by the machine-learning model and could be valuable in system training. However, when it comes to end-to-end retrieval, we only include documents in the end-to-end-retrieved set if "User 1" (the user who is likely to be a better reviewer because they have a higher user recall/user precision rate) marks them as relevant.

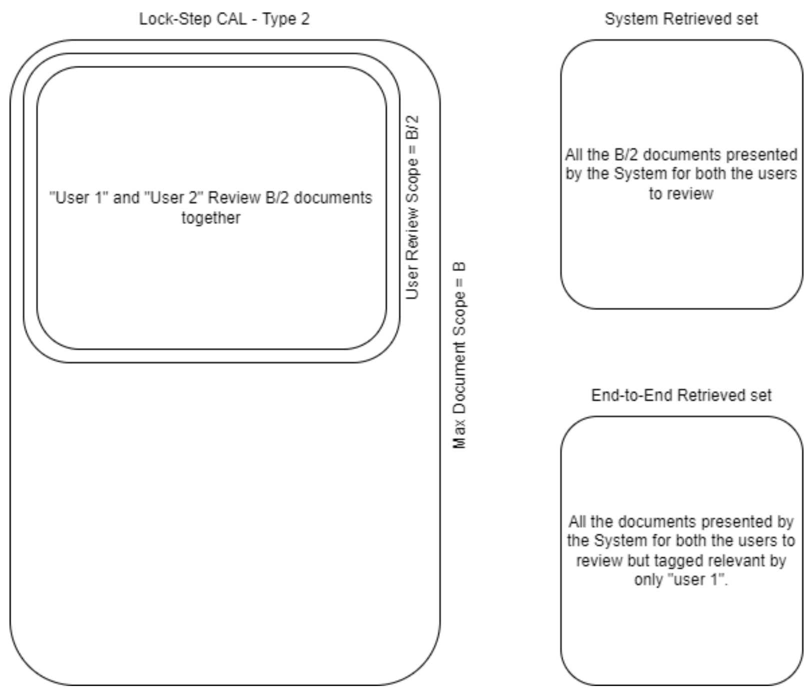


Figure 3.5: Lock-Step CAL–Type 2 Review Strategy

In this way, we can give more weight to the more effective user's judgments and presumably, reduce the number of false positives in the end-to-end-retrieved set. To quantify the retrieval effort of this strategy, the B/2 documents presented to both users by the system are included in the system-retrieved set and the documents presented by the system and tagged relevant only by "User 1" (the comparatively more effective user) constitute the end-to-end-retrieved set. Figure 3.5 depicts the structure of the Lock-Step CAL–Type 2 review strategy.

Majority-Vote-of-Three

In the previous strategy, Lock-Step CAL–Type 2, we can only achieve an end-to-end precision equal to the user precision of the most effective user, and even with such a sophisticated review strategy, the end-to-end retrieval effort may not be able to extract all relevant documents presented by the system. In order to increase the possibility of maximizing the number of relevant documents presented by the system, we consider a review strategy that leverages three users and takes the majority vote of their relevance feedback to classify the relevance of the documents presented to them by the system.This review strategy is constructed from the majority-vote review technique previously deployed by TREC Legal and Total Recall Track coordinators to obtain the ground-truth for the datasets [21, 31, 32].

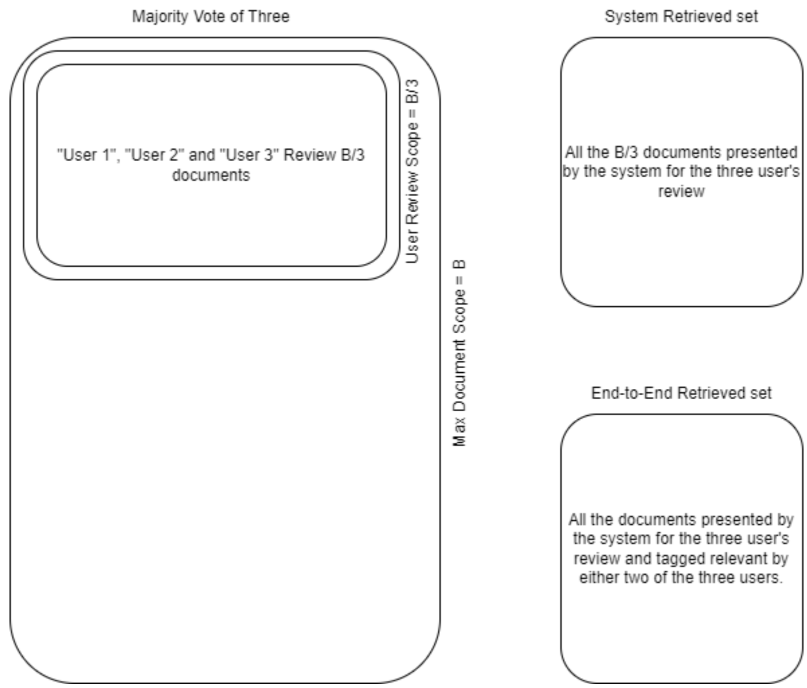


Figure 3.6: Majority-Vote-of-Three Review Strategy

In the TREC 2016 Total Recall Track, among all the review strategies used on the BMI system, the Majority-Vote-of-Three review technique used in the baseline was highly effective at reducing relevance feedback errors and hence, could achieve high recall [32]. Consider the BMI system presenting a relevant document where User 1 tags the document as relevant and User 2 marks the document as non-relevant. User 3 will cast the majority vote on the relevance feedback, marking the document as relevant, indicating that this review was successful in identifying a relevant document. In this example, we see that when User 2 potentially attempted to introduce a false-negative error into the system, the majority vote helped to prevent the error from being introduced (assuming the majority vote to be the correct response). Similarly, if a user incorrectly marks a non-relevant document as relevant, a majority vote can assist in eliminating false positives as well.

Since all three users review each document presented by the system, the budget B is being split equally among the three users; in such a way that each of the three reviewers will get to review only the same B/3 documents. It is important to note that the presumably high-quality relevance feedback provided by this review strategy comes at the cost of the review depth (which is reduced in scope to B/3). To quantify the retrieval effort of this strategy, the B/3 documents presented to the three reviewers are included in the system-retrieved set, and the documents presented by the system and tagged relevant by either two of the three reviewers constitute the end-to-end-retrieved set. Figure 3.6 depicts the structure of the Majority-Vote-of-Three review strategy.

CAL with Quality Control–Type 1

From the Majority-Vote-of-Three review strategy, we can understand that obtaining relevance feedback from three reviewers on each document can help prevent both types of feedback errors. In addition, we can also infer that a majority vote on the relevance feedback is only required when any two users disagree with each other, otherwise, if they agree on the relevance feedback, there is already a majority vote for that relevance feedback. Thus, the third user's relevance feedback has a role to play only when the two (primary) users disagree with each other on the relevance feedback.

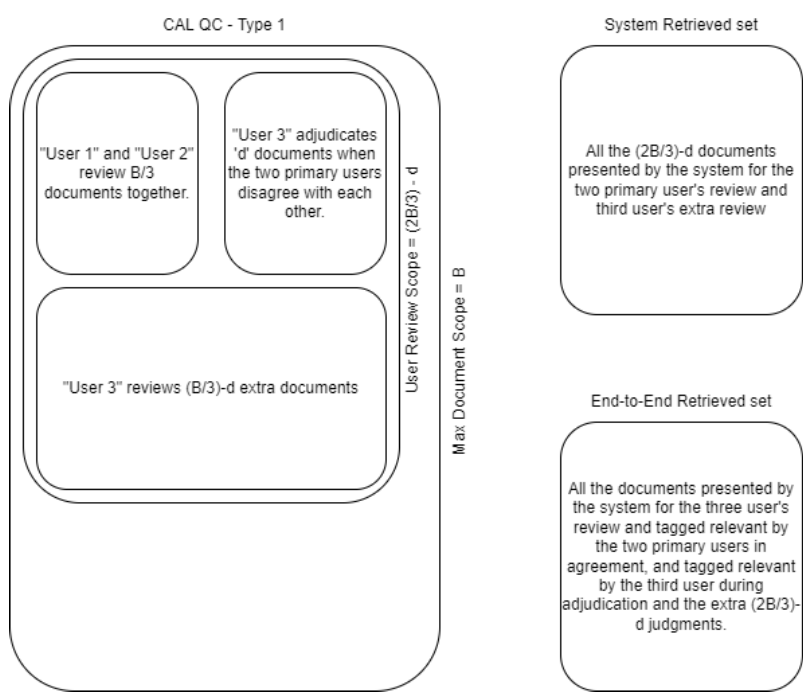


Figure 3.7: CAL with Quality Control–Type 1 Review Strategy

Similar to Majority-Vote-of-Three, the CAL QC–Type 1 review strategy also splits the review budget equally amongst the three users; each user has an opportunity to review B/3 documents. In this review strategy, we initially employ the two primary users to review B/3 documents and only in case of disagreement over the relevance of a document does the third user provide relevance feedback yielding the majority vote. Once the third user has provided quality control on the documents as to which the primary users have disagreed, the third user will have an additional review budget which can be used to review more documents, thereby increasing the user review scope while maintaining the Majority-Vote-of-Three review strategy's effectiveness.

Let us imagine that the two primary users review B/3 documents and disagree on 'd' documents. The third user would provide majority-vote assistance only on these d documents. Once the review scope of B/3 documents is complete, the third user would have an extra review budget of '(B/3)-d', which they can use to review additional documents using what should already be a well-trained CAL system. Through this review technique, the review scope is increased as compared to the Majority-Vote-of-Three review strategy. Figure 3.7 depicts the structure of the CAL QC–Type 1 review strategy. From our simulation experiments, we found that if the third user has a higher user recall rate and a higher user precision rate than the two primary users, the retrieval effectiveness of this strategy increases significantly. This is observed because the efficient third user can effectively review the additional '(B/3)-d' documents. However, in our Results and Discussion Section, we consider all the participating users in this review strategy to have the same user quality (i.e., user recall/precision rates) to establish a level playing field with other review strategies.

CAL with Quality Control–Type 2

Finally, we propose the CAL with Quality Control–Type 2 review strategy to further expand the scope of user review while maintaining the quality control offered by the Majority-Vote-of-Three and CAL with Quality Control–Type 1 review strategies. We can deduct from the previous techniques that employing a larger number of reviewers to provide relevance feedback results in each reviewer receiving a small review budget, which limits review depth and thereby prevents users from reviewing more potentially relevant documents. Therefore, the primary goal of this review strategy is to reduce the number of reviewers employed in TAR without compromising on effectiveness. Throughout our previous review techniques, we have taken for granted the role of the machine-learning system, which provides the most-likely relevant documents. In this review strategy, we seek to harness this silent participant in the process, during the relevance feedback phase, to provide quality control, instead of third user.

In this final review strategy, we consider the machine-learning system itself as one of the users and use only two other human (in this study, simulated human) users. The BMI system ranks the document collection in decreasing order of relevance, indicating that the initial documents in the ranking list are more likely to be relevant than the documents at the bottom of the ranking list. We extend this understanding to our review strategy, with a budget constraint, and formulate a technique to provide relevance feedback to the system. In this review review strategy, budget B is split between the three users in the following manner: User 1 is allocated 2B/3 review budget, User 2 is allocated the remaining “B/3” review budget and User 3, the machine-learning system, does not require a review budget.

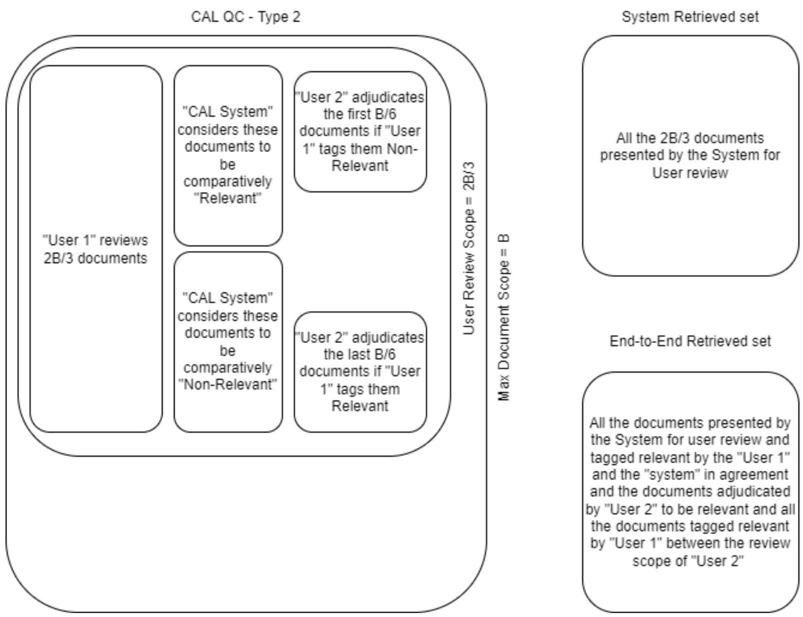


Figure 3.8: CAL with Quality Control–Type 2 Review Strategy

Since the machine-learning system ranks and provides documents for review in the decreasing likelihood of relevance, the system (here, User 3) predicts the initial set of documents presented to the user(s) to be more likely to be relevant than the subsequent documents. Therefore, for a review budget of B, the first B/3 (half of 2B/3) documents reviewed by User 1 are considered comparatively more likely to be relevant by the machine-learning system and the remaining B/3 (the other half of 2B/3) documents reviewed by User 1 are considered comparatively less likely to be relevant by the system. If User 1 tags a document presented from the first half of the system ranking documents as relevant, then User 1 and the system are presumed to be agreeing with each other, and together, they establish a majority irrespective of User 2’s feedback. If User 1 tags a document presented from the first half of the system ranking as not relevant, then we assume there is a disagreement between User 1 and the system as the system predicts that document to be comparatively more likely to be relevant.

To resolve this disagreement, User 2 provides relevance feedback on the document, serving as the majority vote. This process helps to avoid the introduction of false negatives into the system. User 2 adjudicates the first B/6 documents where User 1 and the system "disagree" with each other. Similarly, the system's ranking for the bottom half of the 2B/3 documents is considered comparatively more likely to be non-relevant. If User 1 tags a document from that set as relevant, there is a presumed disagreement between User 1 and the system, and User 2 casts a majority vote on those relevance judgements. This method helps to prevent false positives from entering the system. The remaining B/6 review budget is spent by User 2 on adjudicating the last B/6 documents where User 1 and the system "disagree."

As a result, User 2 resolves conflicts between User 1 and the machine-learning system, thereby endeavoring to extend the positive impacts of the majority vote offered by the Majority-Vote-of-Three and CAL QC–Type 1 review strategies, while also increasing the user review scope to 2B/3 from 2B/3-d, the maximum scope available for our multi-user, hybrid, human-computer assessment strategies. Figure 3.8 depicts the details of the CAL QC–Type 2 review strategy. From our simulation experiments, we found that if User 1 has a higher user recall rate and a higher user precision rate than User 2, the retrieval effectiveness of this strategy increases significantly. This is observed because, the efficient User 1 can effectively review all the 2B/3 documents, thereby carefully allocating the quality control budget of B/3 to User 2. However, in our Results and Discussion Section, we consider all the participating users in this review strategy to have the same quality (i.e., user recall and precision rates) to establish a level playing field with other review strategies.

Chapter 4

Results and Discussion

In this chapter, we compare the performance of the six multi-user, hybrid, human-computer review strategies under study against the single-user, hybrid, human-computer review strategy. The primary goal of eDiscovery, regardless of review strategy, is to retrieve substantially all but only the relevant documents from the corpus, with reasonable effort; thus, we use the recall metric to calculate the effectiveness of the review strategy in retrieving substantially all relevant documents and the precision metric to determine whether the review strategy has retrieved only the relevant documents. We calculate the corresponding F1 score to assess the overall performance of the review strategy because both recall and precision are important in determining if a TAR technique is suitable, although in legal practice, they may not be equally important, as the F1 score assumes.

We have conducted an elaborate set of hybrid, human-computer review experiments for various budget criteria, such as R, 2R, and 3R (where R corresponds to the ground-truth number of relevant documents for each query topic) and different simulated-user variations. These choices for the review budget were considered because we needed to review at least R documents to capture all the relevant documents for a particular topic. Appendix B contains a full breakdown of each review strategy's effectiveness for all the above-mentioned review budgets and user combinations. In this chapter, we discuss only the results employing the following setup: Budget B equals 3R document reviews, and all users, regardless of the number of users, should be fungible and of the same review quality (i.e., same user recall and precision rate); to provide a level playing field for comparing the effectiveness of the various multi-user review strategies against the Single-User CAL review technique across all datasets. The applicability of the results and conclusions reached in this section extend to other review budgets and user combinations as well, but for the sake of brevity, we will not discuss them all.

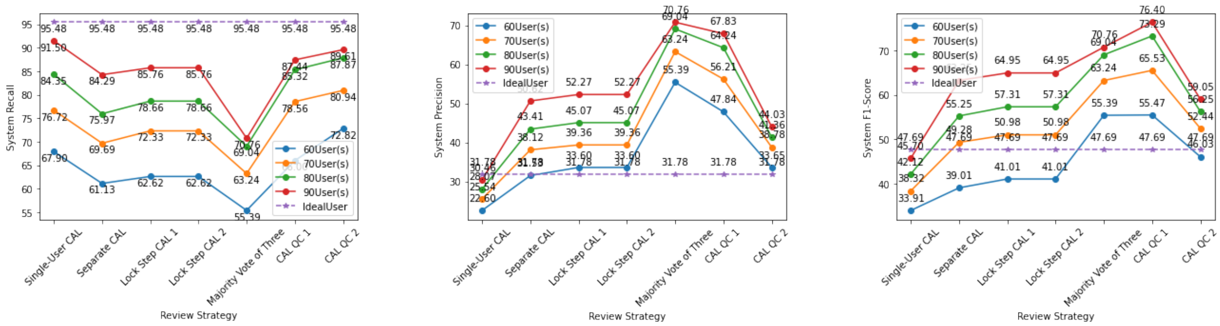
4.1 Results

A hybrid, human-computer TAR system has two independent participants: the system and the reviewer(s), who work together to identify relevant documents from the corpus. As mentioned in Section 2.6, when they come together to perform eDiscovery, we can classify their effort under three different retrieval categories: System Retrieval, User Retrieval, and End-to-End Retrieval. We only get two groups of unique resultant sets during an eDiscovery procedure using TAR: one provided by the system and obtained during system retrieval, and the other generated when the documents are presented by the system and tagged "relevant" by the user, and thus obtained during end-to-end retrieval. As a result, in this section, we independently investigate the effectiveness of the proposed multi-user review strategies against Single-User CAL review strategy under system and end-to-end retrieval efforts.

In our set-up, the resultant document set for user retrieval is identical to the resultant document set for end-to-end retrieval. When the system presents the documents for review, user retrieval evaluates the efficiency of the user/user pool in retrieving the relevant documents. In section 4.2, we will also discuss user retrieval efforts in order to provide a comprehensive justification for the effectiveness of a review strategy. To provide relevance feedback, we consider the following five user/user-pool expertise levels: 60% user recall and precision(60User), 70% user recall and precision(70User), 80% user recall and precision(80User), 90% user recall and precision(90User), and 100% user recall and precision (the Ideal User, included for reference purposes only). Finally, in section 4.1.3, we perform the Two Sample T-Test to verify if a proposed multi-user review strategy is superior to the Single-User CAL review strategy.

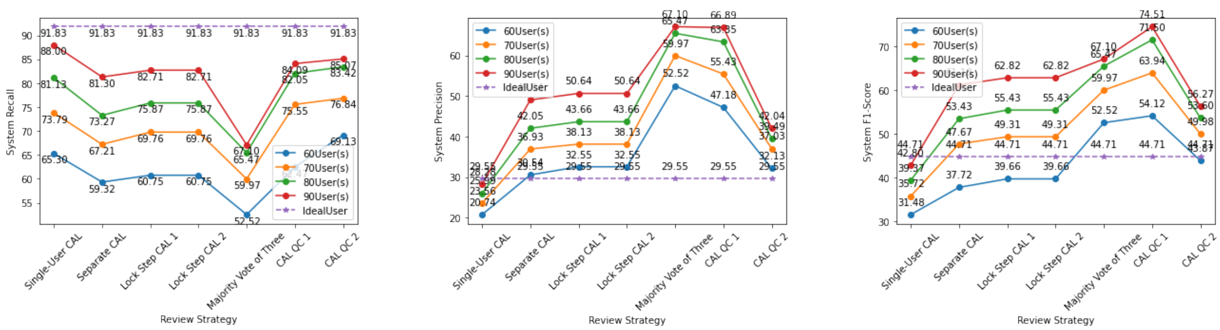
4.1.1 System-Retrieval Results

When employing different review strategies, we seek to understand how well the CAL system alone has performed in obtaining the relevant documents, within the available review budget of B equal to 3R document reviews. We will analyze and compare the performance of the CAL system when implementing each of the review strategies with respect to the At-Home1 Dataset; the other three datasets show similar results as can be observed in Graphs 4.2 - 4.4.



(a) System Recall vs Review Strategy (b) System Precision vs Review Strategy (c) System F1 Score vs Review Strategy

Figure 4.1: At-Home1 Dataset: System-Retrieval Results

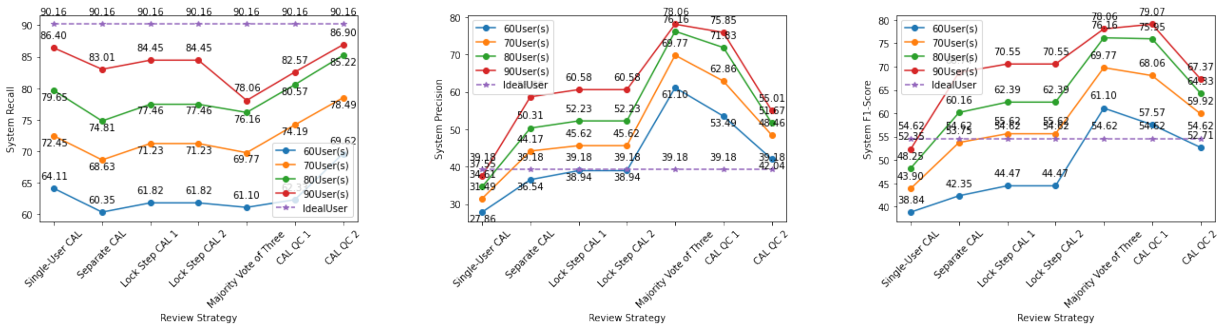


(a) System Recall vs Review Strategy (b) System Precision vs Review Strategy (c) System F1 Score vs Review Strategy

Figure 4.2: At-Home2 Dataset: System-Retrieval Results

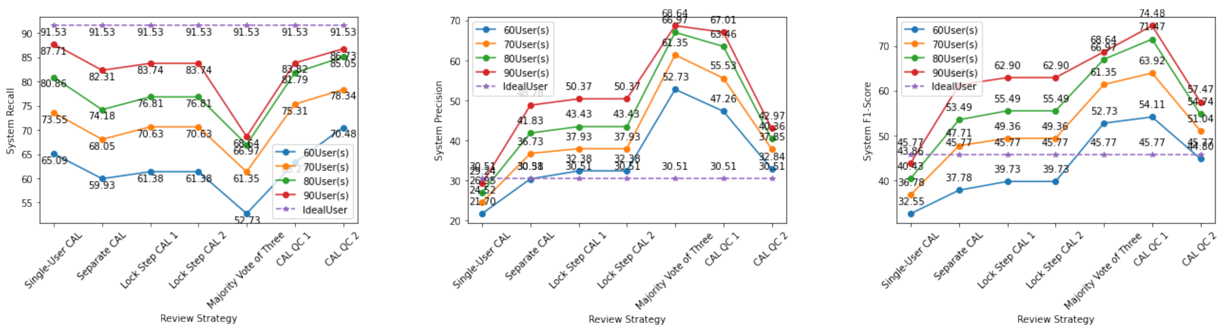
System Recall Effectiveness

For the allocated review budget of B equal to 3R documents, we seek to determine which review strategy has enabled the CAL system to retrieve more relevant documents. Initially, we leverage the Ideal User with 100% user recall and 100% user precision to provide relevance feedback in the Single-User CAL review strategy, showing the maximum possible system recall that can be achieved for the given review budget. With reference to Graph 4.1a, we infer that an Ideal User reviewing 3R documents for each of the ten topics in At-Home1 Dataset will be able to achieve a system recall of 95.48%. This is an ideal scenario, which is unlikely to occur in practice as reviewers are not perfect, and a certain amount of false negative feedback error will inevitably exist. Addressing this is one of key reasons to formulate multi-user review strategies. A realistic Single-User CAL system, perhaps 80User's CAL system will be able to retrieve only 84.35% of the relevant documents, thereby failing to retrieve 11.13% of the relevant documents.



(a) System Recall vs Review Strategy (b) System Precision vs Review Strategy (c) System F1 Score vs Review Strategy

Figure 4.3: At-Home3 Dataset: System-Retrieval Results



(a) System Recall vs Review Strategy (b) System Precision vs Review Strategy (c) System F1 Score vs Review Strategy

Figure 4.4: At-Home4 Dataset: System-Retrieval Results

To analyze the system recall effectiveness of the multi-user review strategies, let us consider 80User(s) providing relevance feedback to the CAL system. The Separate CAL review strategy seeks to increase the retrieval of relevant documents by engaging two users reviewing documents using separate CAL systems, but this approach manages to retrieve only 75.97% of the relevant documents because the user-review scope (i.e. number of documents reviewed) is halved and the relevance feedback provided by one user working alone is not helpful for the other user. Lock-Step CAL–Type 1 and Type 2 engage two reviewers to provide relevance feedback together, on the same CAL system, so with respect to system recall both strategies have similar performances and retrieve 78.66% of the relevant documents. Although there is an increase of 2.69% in system recall in relation to Separate CAL, Lock-step CAL strategies fall short of Single-User CAL review strategy by 5.69%. To increase the chance of identifying the system-presented relevant documents correctly and facilitating better system training, we study the Majority-Vote-of-Three strategy. But we end up retrieving only 69.04% of the relevant documents because the user-review scope is further reduced from B/2 to B/3 and the relevant documents only within that small scope can be retrieved by the CAL system.

To increase the user review scope and at the same time allocate the review budget to provide quality control on relevance feedback, we presented two multi-user review strategies: CAL with Quality Control–Type 1 and CAL with Quality Control–Type 2. We observe that CAL QC–Type 1 retrieved 85.32% of the relevant documents and CAL QC–Type 2 retrieved 87.87% of the relevant documents, showing that CAL QC–Type 1 and Type 2 review strategies can achieve a greater system recall than the Single-User CAL review strategy. An interesting observation is that the more prone the user is to making errors, the better is the system performance boost when using the CAL QC strategies. Another observation is that, users with higher user recall and precision rates, like 90User(s), will retrieve slightly fewer relevant documents while using CAL QC strategies than the Single-User CAL review strategy because a portion of the review budget is wasted on quality control which is not required for a highly effective reviewer. But, finding a reviewer with consistent 90% and above user recall and precision rate would presumably be unusual and therefore, for most scenarios, the CAL QC review strategies appears to be an effective alternative to the Single-User CAL review strategy in achieving high system recall.

System Precision Effectiveness

For the allocated review budget of B equal to 3R documents, we seek to determine which review strategy has enabled the CAL system to retrieve relevant documents more precisely. Initially, we leverage the Ideal User with 100% user recall and 100% user precision to provide relevance feedback in the Single-User CAL review strategy to show the Ideal User's corresponding system precision which can be achieved for the given review budget. With reference to Graph 4.1b, we infer that an Ideal User reviewing 3R documents for each of the ten topics in At-Home1 Dataset will be able to achieve a system precision of 31.78%. This is an ideal scenario, which is unlikely to occur in practice as reviewers are not perfect and a certain amount of false positive feedback error will inevitably exist. Addressing this is one of key reasons to formulate multi-user review strategies. A realistic Single-User CAL system, e.g., 80User's CAL system will be able to achieve only 28.07% system precision, showing a slight drop in system precision of 3.71% from the ideal condition.

We observe that the drop in system precision when using a 80User in Single-User CAL is not that substantial and thereby might question whether there is a need for multi-user review strategies to improve on the system precision. The answer is "yes." Setting aside the fact that every minute spent reviewing a non-relevant document is wasted time, so any improvement in precision reduces cost. Because multi-user review strategies spend a portion of the review budget on the initial batches, they therefore help to prevent the system from retrieving more documents having a lower confidence score, which, in turn, increases system precision. This is the reason why we observe that the multi-user review strategies, irrespective of the user's performance level, can achieve a system precision even greater than the Ideal User in Graph 4.1b. The Majority-Vote-of-Three review strategy achieves the highest system precision of 69.04%, followed by the CAL QC–Type 1 review strategy, achieving 64.24% system precision, because these strategies allocate all or almost all of their review budget in reviewing the first "R" (B/3) system-presented documents attempting to avoid the introduction of false positive errors to a great extent.

With reference to Table B.1, we can observe that an 80Users system precision using the Majority-Vote-of-Three review strategy is only 2.05% short of the Ideal User's system precision when R documents are reviewed, demonstrating the Majority-Vote-of-Three and CAL QC–Type 1 review strategies' near perfect system precision. The CAL QC–Type 2 review strategy's system precision drops to 41.36% because of the increased user review scope and because it considers the system itself as a precise reviewer equal to a human reviewer, which it is not. With these findings we can confirm that multi-user review strategies appear to be consistent in achieving superior system precision to the Single-User review strategy.

System F1 score Effectiveness

For the allocated review budget of B equal to 3R documents, we seek to determine which review strategy has enabled the CAL system to achieve high recall while also maintaining high precision. With reference to Graph 4.1c, we infer that the Ideal User reviewing 3R documents for each of the ten topics in At-Home1 Dataset will be able to achieve a system F1 score of 47.69%. A realistic single-user CAL system, like 80User's CAL system, will be able to achieve only a 42.12% system F1 score. We can observe that when imperfect user(s) (i.e., 65% and above user recall and precision), as in Voorhees' study [34]) provide relevance feedback using any of the multi-user review strategies, the CAL system is able to achieve an F1 score greater than even that of the Ideal User using a Single-User review strategy.

The highest system F1 score of 73.29% is achieved using the CAL QC–Type 1 review strategy because it has the best balance between both system recall and system precision. It is interesting to note that the CAL QC–Type 2 review strategy has one of the lowest system F1 scores across all strategies despite having the highest system recall scores for a multi-user review strategy. This is due to two key factors: first, the broad user review scope of 2B/3 documents and second, it considers the system to be as precise as a human reviewer, which it is not. These two factors are responsible for CAL QC–Type 2 review strategy's lower system precision, thereby affecting its resultant system F1 score. We cannot discard the CAL QC–Type 2 review strategy just yet, by analyzing only the system retrieval results, since there is human feedback involved in the TAR process and we need to consider the end-to-end retrieval effort to determine a review strategy's true efficacy.

4.1.2 End-to-end-Retrieval Results

When employing different review strategies, we need to understand how well the CAL system and the reviewers have performed together in finding the relevant documents, within the available review budget of B equal to 3R documents. We will analyze and compare the performance of the end-to-end system when implementing each of the six review strategies with respect to the At-Home1 Dataset. We observe similar findings for our other three datasets as shown in Graphs 4.6a - 4.8c.

(a) End-to-End recall vs Review Strategy

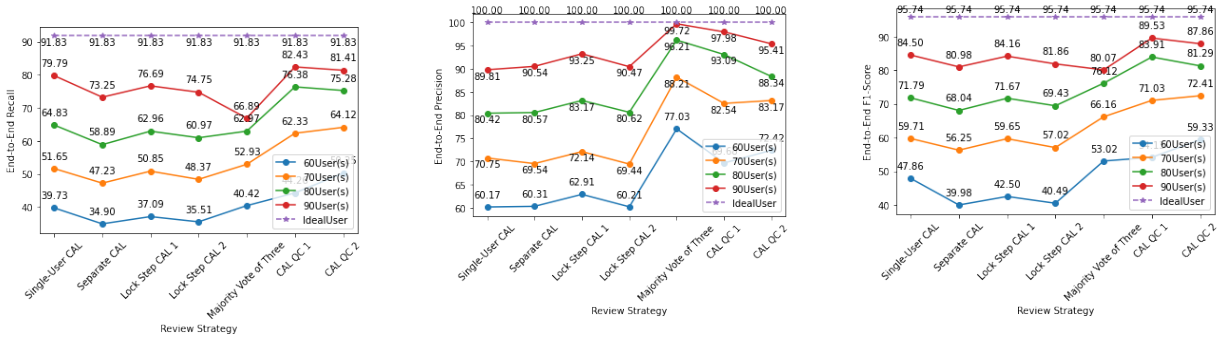
(b) End-to-End Precision vs Review Strategy

(c) End-to-End F1 Score vs Review Strategy

Figure 4.5: At-Home1 Dataset: End-to-end-Retrieval Results

End-to-End Recall Effectiveness

For the allocated review budget of B equal to 3R documents, we seek to determine which review strategy has enabled the end-to-end system to retrieve more relevant documents. Initially, we leverage the Ideal User with 100% user recall and 100% user precision to provide relevance feedback in the Single-User review strategy, showing the maximum possible end-to-end recall that can be achieved for the given review budget. With reference to Graph 4.5a, we infer that an Ideal User reviewing 3R documents for each of the ten topics in At-Home1 dataset will be able to achieve an end-to-end recall of 95.48%. The Ideal User

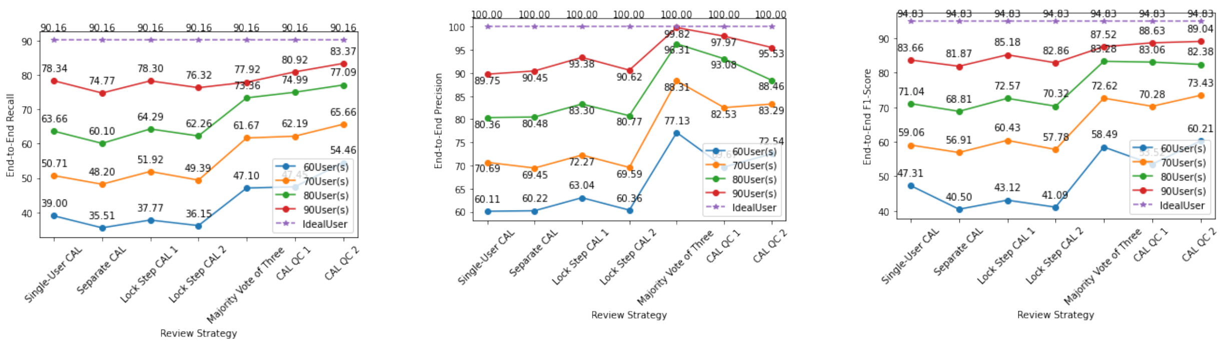


(a) End-to-End recall vs Review Strategy

(b) End-to-End Precision vs Review Strategy

(c) End-to-End F1 Score vs Review Strategy

Figure 4.6: At-Home2 Dataset: End-to-end-Retrieval Results



(a) End-to-End recall vs Review Strategy

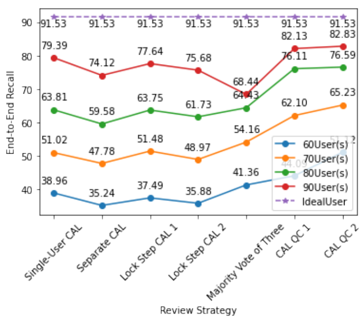
(b) End-to-End Precision vs Review Strategy

(c) End-to-End F1 Score vs Review Strategy

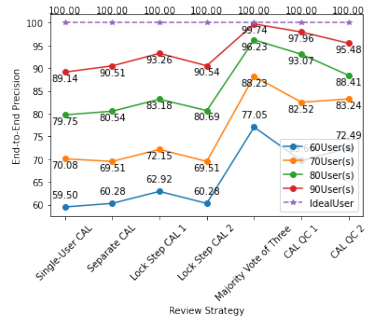
Figure 4.7: At-Home3 Dataset: End-to-end-Retrieval Results

correctly codes all the system-presented relevant documents, which is why we observe the end-to-end recall to be same as the corresponding system recall. This is an ideal scenario, which is unlikely to occur in practice as reviewers are not perfect and a certain amount of false negative feedback error will inevitably occur. A realistic single-user end-to-end system, like 80User's end-to-end system recall, will be able to retrieve only 67.42% of the relevant documents, thereby failing to retrieve 16.93% of the relevant documents presented by the system for review.

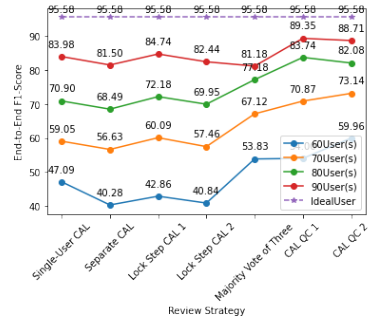
When an imperfect user provides relevance feedback, we observe that there is a two-fold loss in finding relevant documents. First, the system receives inferior relevance feedback from the imperfect user, and therefore is trained poorly and presents fewer relevant documents to the user for review. Second, the imperfect reviewer will inadvertently fail to correctly code all the relevant documents from the fewer relevant documents presented by the poorly trained system. This is the reason why we often observe the end-to-end recall



(a) End-to-End recall vs Review Strategy



(b) End-to-End Precision vs Review Strategy



(c) End-to-End F1 Score vs Review Strategy

Figure 4.8: At-Home4 Dataset: End-to-end-Retrieval Results

to be lower than the system recall. The Separate CAL review strategy retrieved 61.11% of the relevant documents, and the Lock-Step CAL–Type 1 and Type 2 review strategies retrieved 65.28% and 63.22% of the relevant documents respectively, which are both slightly less than the Single-User review strategy's end-to-end recall. The Majority-Vote-of-Three review strategy increases the end-to-end recall to 66.45% which is slightly closer to that of the Single-User review strategy.

The CAL QC–Type 1 and Type 2 review strategies achieve a higher end-to-end recall when compared to Single-User review strategy, and the other multi-user review strategies. For example, 80User(s) achieve an end-to-end recall of 67.42% when reviewing B documents using Single-User review strategy, but retrieve 79% of the relevant documents just by reviewing 2B/3 (for CAL QC–Type 1, slightly less than 2B/3) documents using either of the CAL QC–Type 1 or 2 techniques. An interesting observation with respect to the Majority-Vote-of-Three review strategy and the two CAL QC review strategies is that they are highly effective if the user-pool is of a lower quality. Unlike the Majority-Vote-of-Three review strategy, CAL QC techniques show improved end-to-end recall results even for expert reviewers, e.g., 90User(s), when compared to that of the Single-User review strategy.

End-to-End Precision Effectiveness

For the allocated review budget of B equal to 3R documents, we seek to determine which review strategy has enabled the end-to-end system to retrieve relevant documents more precisely. Initially, we leverage the Ideal User with 100% user recall and 100% user precision to provide relevance feedback in the Single-User review strategy, showing the maximum possible end-to-end precision that can be achieved for the given review budget. With reference to Graph 4.5b, we infer that an Ideal User reviewing 3R documents for each of the ten topics in At-Home1 Dataset will be able to achieve an end-to-end precision of 100%. This is an ideal scenario which is unlikely to occur in parctice as reviewers are not perfect and a certain amount of false positive feedback errors will inevitably occur. Addressing this issue is one of the key reasons to formulate multi-user review strategies. A realistic single-user CAL system like 80User's end-to-end system will be able to achieve only 80.63% system precision, showing a drop in system precision of approximately 20% from the ideal condition.

Separate CAL and Lock-Step CAL–Type 2 also provide a similar end-to-end precision to that of the Single-User review strategy, i.e., approximately 80%. However, Lock-Step CAL–Type 1 combines the effort of the two users and manages to increase the end-to-end precision to 83.12%. The Majority-Vote-of-Three review strategy achieves the maximum end-to-end precision surpassing that of more efficient users like 90User using Single-User review strategy and approach the end-to-end precision of the Ideal User, at 96.26%. CAL QC–Type 1 and CAL QC–Type 2 techniques show a slight drop in the end-to-end precision from the Majority-Vote-of-Three review strategy to 93.06% and 88.53%, respectively, but consistently achieve an end-to-end precision greater than a Single-User review strategy. Overall, multi-user review strategies like Lock-Step–Type 1, Majority-Vote-of-Three, CAL QC–Type 1 and Type 2 achieve higher end-to-end precision than the Single-User review strategy, consistently for all user expertise levels.

End-to-End F1 score Effectiveness

For the allocated review budget of B equal to 3R documents, we seek to determine which review strategy has enabled the end-to-end system to achieve high recall while also maintaining high precision. With reference to Graph 4.5c, we infer that an Ideal User reviewing 3R documents for each of the ten topics in At-Home1 Dataset will be able to achieve an end-to-end F1 score of 97.69%. A realistic single-user end-to-end system, like 80User's end-to-end system will be able to achieve only a 73.44% end-to-end F1 score. The initial three review strategies: Separate CAL, Lock-Step CAL–Type 1 and Type 2 show slightly inferior end-to-end F1 score results when compared to the baseline Single-User review strategy. The Majority-Vote-of-Three review approach achieved comparable end-to-end recall to the single user review strategy, however, the considerable difference in the end-to-end precision gives the former method a strong F1 score.

The CAL QC strategies Type 1 and Type 2, having achieved high end-to-end recall and high end-to-end precision consistently across all user expertise levels, prove to be the superior review strategies when compared to the Single-user CAL as well as the other previously discussed multi-user review strategies. An interesting observation is that with, a higher-expertise review pool, CAL QC–Type 1 performs slightly better than CAL QC–Type 2, but CAL QC–Type 2 slightly outperforms CAL QC–Type 1 when the reviewers belong to a lower-expertise review pool. For example, the 90User pool achieves a 91.39% F1 score using CAL QC–Type 1, but achieves only 90.31% when using CAL QC–Type 2. When using a 70User review pool, a 72.57% F1 score is achieved using CAL QC–Type 1, but a 74.54% F1 score is achieved using CAL QC–Type 2. As a result, these two techniques can be implemented for different user groups to maximize effectiveness if the level of expertise of the reviewers is known in advance.

4.1.3 Paired Two-Sample T-Test Results Comparing End-to-End Recall

Legal parties in eDiscovery aim to get almost all relevant documents with reasonable effort; as a result, we focus on determining the statistical significance of the end-to-end recall values rather than the precision and F1 score values, since a reasonable value is acceptable. Using a Paired Two-Sample T-Test, we seek to determine whether the improved end-to-end recall results observed in Section 4.1.2, for the CAL QC–Type 1 and Type 2 review strategies are statistically significant compared to the Single-User CAL review strategy at a 95% confidence level. The At-Home4 dataset was used for the development of multi-user review strategies, and therefore, we do not use the results obtained from this dataset for the T-Test to avoid bias. The review strategy set-up is applied without modification or tuning for the following datasets: At-Home1, At-Home2, and At-Home3. As a result, while applying these multi-user review strategies, we perform a T-Test on the topic-wise recall percentages derived from these three datasets.

In Tables B.19, B.22, B.25, B.28, we describe the T-Test calculation for the end-to-end recall performance of 60User(s) (i.e., the users with the lowest observed expertise level), 70User(s), 80User(s), and 90User(s) (the most efficient user pool considered in our study) respectively. When 60User(s), 70User(s), and 80User(s) leverage CAL QC–Type 1 and CAL QC–Type 2 review strategies, their corresponding p-values are negligible (almost zero). Hence, both CAL QC–Type 1 and CAL QC–Type 2 review strategies show statistically significant improvements in end-to-end recall, at 95% confidence level over the Single-User CAL review strategy for low to medium expertise-level users (i.e., 60User(s) - 80User(s)).

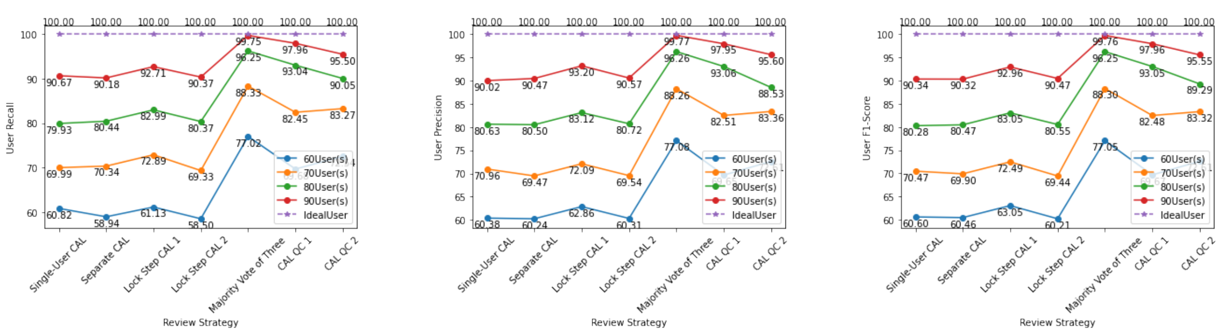
When high expertise-level users, say 90User(s), leverage CAL QC–Type 1 and CAL

QC–Type 2 review strategies, they achieve p-values of 0.14211 and 0.08549 respectively. According to the obtained p-values for the 90User(s) pool, the end-to-end recall results show improvement, but not enough improvement to achieve significance at the 95% confidence level. We posit that this is observed because high expertise-level reviewers make significantly less relevant feedback errors and thus these QC strategies have less capacity for improvement in recall through quality control. The mean recall of the other multi-user review strategies like Separate CAL, Lock-Step CAL–Type 1, Lock-Step CAL–Type 2 and Majority-Vote-of-Three is substantially lower than that of the Single-User CAL review strategy. Therefore, the T-Test results help demonstrate that they are inferior review strategies compared to the Single-User review strategy.

To summarize, only our proposed quality control review strategies, CAL QC–Type 1 and CAL QC–Type 2, show statistically significant superior end-to-end recall when compared to the Single-User CAL review approach.

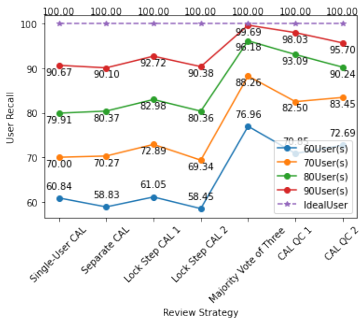
4.2 Discussion

According to our statistical testing, two of the proposed multi-user QC review techniques show statistical significance in terms of superior recall when compared to the Single-User CAL review approach. Therefore, we can state that our two CAL QC review strategies are potentially reasonable alternatives as compared to the baseline, Single-User CAL review strategy. In this section, we will explain the trends observed in the system and end-to-end results by discussing the user retrieval performances shown for each of the multi-user review strategies with respect to the At-Home1 Dataset, 4.9a - 4.9c. We will also discuss the advantages and limitations associated with each of the multi-user review strategies.

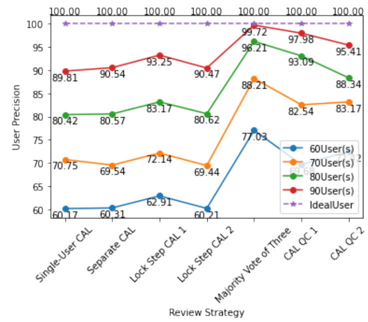


(a) User Recall vs Review Strategy (b) User Precision vs Review Strategy (c) User F1 Score vs Review Strategy

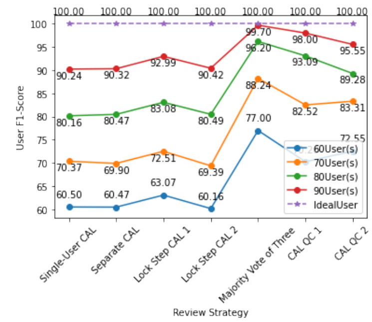
Figure 4.9: At-Home1 Dataset: User-Retrieval Results



(a) User Recall vs Review Strategy

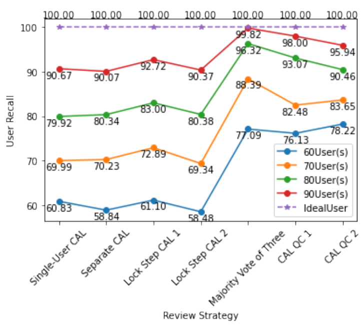


(b) User Precision vs Review Strategy

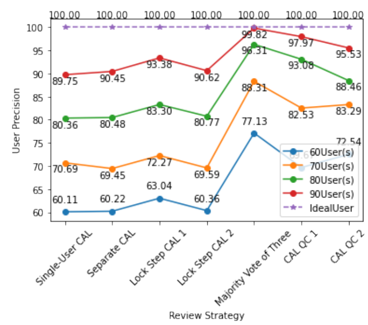


(c) User F1 Score vs Review Strategy

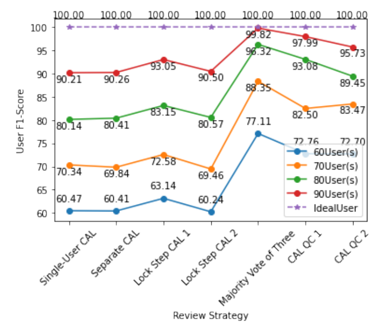
Figure 4.10: At-Home2 Dataset: User-Retrieval Results



(a) User Recall vs Review Strategy



(b) User Precision vs Review Strategy

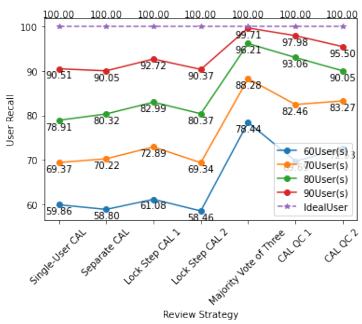


(c) User F1 Score vs Review Strategy

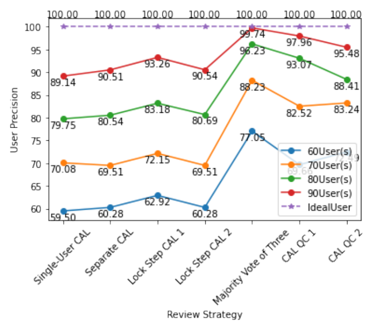
Figure 4.11: At-Home3 Dataset: User-Retrieval Results

4.2.1 Separate CAL

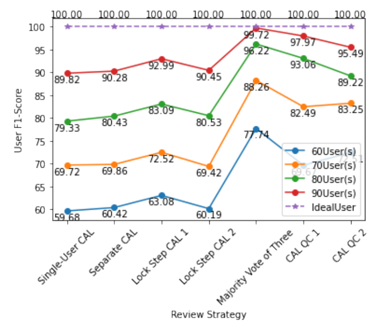
This review strategy splits the review budget B among two reviewers and enables them to each review B/2 documents on two separate CAL systems. This strategy attempts to avoid missing relevant documents when presented by the system; in other words, it seeks to overcome false negative errors common to the Single-User review strategy. With the Separate CAL review strategy, we observe that the resultant user recall and user precision are the same as the defined user's recall and precision; for example, 80User(s) will retrieve 80% of the system-presented relevant documents correctly and 80% of the documents tagged as relevant by the reviewer will be relevant. As a consequence, using two 80Users does not help to raise their resultant recall and precision rates; nevertheless, when their retrieval sets are combined, we can see a modest rise in the total number of relevant documents retrieved.



(a) User Recall vs Review Strategy



(b) User Precision vs Review Strategy



(c) User F1 Score vs Review Strategy

Figure 4.12: At-Home4 Dataset: User-Retrieval Results

Advantages:

1. Since two CAL systems are trained separately, the systems will potentially be able to identify different documents from differently trained systems. Since the relevance feedback provided by one user does not influence the other, there is a possibility for each of the users to encounter new "potentially relevant" documents not seen by the other.

Limitations:

1. Both users start from the same starting point, with the same topic query, so they should end up reviewing very similar documents at the beginning of the relevance feedback process. At the conclusion of the retrieval effort, there will be multiple identical documents retrieved in the separate CAL processes, which is wasted effort (and cost).

2. The extra relevant documents found by "user 1" cannot be used in training the machine-learning algorithm being trained by "user 2", and vice versa.

3. Since we combine the retrieval efforts of both users, the cumulative number of false positives tagged by both users will lower the end-to-end precision.

4. The scope of document review is reduced by half, documents beyond the $B/2^{\text{th}}$ mark will not be reviewed under this strategy as the strategy would have already maxed out its review budget at B.

4.2.2 Lock-Step CAL–Type 1

This strategy attempts to overcome the first two limitations observed in the Separate CAL review strategy discussed above. By enabling two users to review together, on the same system, the relevant documents found by both the users are used for training by the system and are retrieved in the end-to-end retrieved set. The benefit of this combined effort is evident in the resultant user recall and precision rates; for example, 80User(s) will retrieve 82.99% of the system-presented relevant documents correctly and 83.12% of the documents tagged as relevant by the reviewers will be relevant.

Advantages:

1. Both users' relevance feedback help the CAL system to learn, and the number of documents tagged relevant using this strategy should be higher than either user working alone, since we combine all the documents tagged relevant by either user to train the system, and add them to the end-to-end retrieved set.

Limitations:

1. Because we combine the retrieval efforts of both users, the false positives induced by each of them will end up training the system potentially resulting in lower system precision. The cumulative number of false positives tagged by both users will also lower the end-to-end precision.

2. Similar to Separate CAL, the scope of document review is reduced by half, since documents beyond the $B/2^{\text{th}}$ mark will not be reviewed under this strategy, as the strategy would have already maxed out its review budget at B.

4.2.3 Lock-Step CAL–Type 2

This strategy attempts to overcome the first limitation observed in the Lock-Step CAL–Type 1 review strategy. By enabling two users to review together, on the same system, the relevant documents found by both users will contribute to the training of the system, but a document will only be included in the end-to-end retrieval set if the more effective (i.e., higher quality) user tags the document as relevant. This technique seeks to reduce the number of false positives induced by the less effective user. In Graphs 4.9a-4.12c, we do not observe an increase in the resultant user recall and precision rates since we have only engaged users from the same review-quality pool. However, when 90User and 70User provide relevance feedback utilizing the Lock-Step CAL–Type 2 review technique, they obtain 8% higher resultant user recall and user precision rates, and consequently, achieve 6% higher end-to-end recall than for Lock-Step CAL–Type 1. The end-to-end retrieval results, for all the simulated-user combinations, employing Lock-Step CAL–Type 2 can be observed in Tables B.13 - B.16. The benefit of this review strategy is evident only if we are able to identify in advance which of the two users is the superior one.

Advantages:

1. This review strategy prevents the unnecessary inclusion of false positive documents tagged by "user 2" (the user with lower user recall and user precision rates) in the end-to-end retrieved set, which will thereby increase the end-to-end precision value to the user precision value of the more effective user.

Limitations:

1. Any documents missed by "User 1" (i.e., the more effective reviewer) as false negatives will not be captured by this review strategy in the end-to-end retrieved set, even if "User 2" (i.e., the less effective reviewer) manages to correctly tag the document presented by the system as relevant. So, the end-to-end recall cannot be higher than the user recall of the more effective reviewer.

2. Similar to Lock-Step CAL–Type 1, the scope of document review is reduced by half, because documents beyond the $B/2^{th}$ mark will not be reviewed under this strategy, as the strategy would have already maxed out its review budget at B.

4.2.4 Majority-Vote-of-Three

In the Lock-Step CAL–Type 2 review strategy, although we achieve an end-to-end precision equal to the user precision of the most effective user, the end-to-end retrieval effort will not be able to extract all relevant documents presented by the system because of false negative errors. To overcome this drawback, and at the same time maintain Lock-Step CAL–Type 2's high precision, we considered a review strategy that leverages three users and takes the majority vote of their relevance feedback to classify the relevance of the documents presented by the system. The benefit of this review strategy is evident in the large increase in the resultant user recall and precision rates; for example, 80User(s) retrieved 96.25% of the system-presented relevant documents correctly and 96.26% of the documents tagged as relevant by the reviewers were relevant. The number of false negatives and false positives induced in this review strategy is low.

Advantages:

1. The system training accomplished using the majority vote relevance feedback of three imperfect users is almost equal to the system training done by a perfect user.
2. This strategy manages to retrieve almost all of the relevant documents presented by the system for review and thereby has an end-to-end recall almost equal to the system recall.
3. This strategy also has a higher end-to-end precision, since by taking the majority vote of three users, the chances of adding false positive documents into the end-to-end retrieved set is reduced.

Limitations:

1. The system recall for this review strategy is comparatively lower using this strategy as compared to the earlier proposed review strategies, because users review only B/3 documents and relevant documents only within this scope can be retrieved. Thus, the high end-to-end recall and end-to-end precision achieved by this review strategy comes at the cost of review budget; all of the review budget B is spent on reviewing only B/3 documents, and thereby the review scope is reduced to B/3 documents.

4.2.5 CAL with Quality Control–Type 1

The CAL QC–Type 1 review strategy endeavors to increase the user review scope, while at the same time maintaining the Majority-Vote-of-Three review strategy's high user recall and high precision. This strategy achieves resultant user recall and precision rates close to that of the Majority-Vote-of-Three review strategy. When 80User(s) provide relevance feedback using the CAL QC–Type 1 strategy, they retrieved 93.04% of the system-presented relevant documents and 93.08% of the documents tagged as relevant by the reviewer were relevant.

Advantages:

1. Similar to the Majority-Vote-of-Three review strategy, this review strategy results in near-perfect relevance feedback for the first B/3 documents.

2. By establishing a well-trained system using only the first B/3 documents, this strategy can increase the user review scope by saving some of the review budget for the third user, thereby enabling that user to continue to review documents using a well-trained system.

3. Superior recall can be achieved when the reviewers come from a high-expertise reviewer pool.

Limitations:

1. The review scope of this strategy is limited to less than 2B/3 documents. Since the two primary users are imperfect, the third user will have to spend their review budget reviewing some of the documents as to which the two primary reviewers disagreed and will therefore necessarily review fewer than B/3 more documents after the primary assessors have reviewed the initial B/3 documents.

4.2.6 CAL with Quality Control–Type 2

This review strategy seeks to overcome the only major drawback of the CAL QC–Type 1 review strategy: the moderate review scope. By considering the system as one of the users, the system provides quality control to a certain extent. The benefit of this review strategy can be observed in the resultant user recall and precision rates, for example, 80User(s) correctly retrieved 90.05% of the system-presented relevant documents, and 88.53% of the documents tagged as relevant by the reviewer were relevant.

Advantages:

1. This review technique broadens the scope of the review to 2B/3, while maintaining quality control over user feedback.

2. The strategy engages another user ("User 2") to re-review the system-presented documents and capture relevant documents that "User 1" may have missed. This helps

to avoid false negatives, thereby ensuring the system has the maximum opportunity to learn features of all relevant documents presented initially by the system and to identify those documents for the end-to-end retrieved set.

3. By re-reviewing the last B/6 documents tagged relevant by "User 1," this strategy also attempts to avoid false positives being added to the end-to-end retrieved set.

4. High recall can be achieved irrespective of the user(s)'s expertise level(s).

Limitations:

1. Not all documents about which "User 1" and the system disagree are re-reviewed by "User 2" because, only B/3 review budget is allocated to "User 2" to address false negative errors and another B/3 budget to address false positive errors. Therefore, certain documents outside of "User 2"'s scope will have no quality control performed as to them.

The CAL QC–Type 2 review method broadens the review scope with a reasonable level of quality control and retrieves the largest number of relevant documents of all our proposed methods. Nonetheless, this method may still have overlooked some relevant documents during the review process. CAL QC–Type 1, on the other hand, achieves near-perfect relevance feedback, at the expense of reduced review scope. CAL QC–Type 1 appears to be an effective approach for high-expertise users, while it performs closer to the Majority-Vote-of-Three review method when used by low-expertise users, whereas, CAL QC–Type 2 performs well with users of all expertise levels. The CAL QC–Type 2 review method is recommended if we are prepared to forego some relevant documents while still achieving the maximum recall within the allocated review budget. Alternatively, if each document is considered equally significant, the CAL QC–Type 1 review method is suggested. As a result, both quality control strategies appear to perform well and could be applied depending on the requirement for optimal outcomes.

4.2.7 Workload on Reviewers

The greater the number of documents reviewed, the more review time and effort will be required of a user. According to a study published by Wong et. al., long hours of work can affect both the quality of the work and the health of the user [22]. Unlike Single-User CAL, multi-user strategies allow for the distribution of review tasks among multiple users, thereby potentially minimizing stress on a single individual. This distribution of review tasks also helps in accommodating the collaboration of various departments/reviewers who may define relevance differently. In this section, we discuss how each of the multi-user review strategies delegate the review workload to different users, unlike the Single-User review strategy, which allocates the entire workload on a single user.

For review strategies like Separate CAL, Lock-Step CAL–Type 1 and Type 2, the effort spent by each reviewer is halved because the review budget is split into two equal parts. The Majority-Vote-of-Three and CAL QC–Type 1 review strategies further reduce the review burden on each reviewer as they need to review only B/3 documents. The quality of feedback provided by both of the latter are similar, but CAL QC–Type 1 optimally uses the reviewers to achieve a higher recall. In CAL QC–Type 1, it is interesting to note that once B/3 documents have been reviewed, we only require the time and effort of one user for additional review. CAL QC–Type 2 splits the review tasks unequally, thereby placing a small amount of extra strain on one user, but this review load is still lower than that of the Single-User CAL review strategy. Even though, the second user requires only a B/3 review effort, the user is expected to stay until the end of the first user's review budget is exhausted to provide quality control. The structure of engaging reviewers in such a manner appears to help in reducing false negatives and false positives to some extent and thereby achieves higher end-to-end recall and precision results than the Single-User CAL review strategy. The distribution of work guaranteed by multi-user review strategies helps reduce the time and effort taken by each user to perform review, and thereby potentially avoids compromising the quality of work and health of the reviewer.

Chapter 5

Conclusion and Future Work

The widely used single-user review strategy does not itself provide quality control on the relevance feedback used to train the system and is potentially prone to inducing a large number of errors. Human review will invariably contain false positive and false negative errors as the concept of relevance is subjective and the presence of feedback errors limits state-of-the-art TAR systems from achieving maximum performance. This research examines the potential advantage of multi-user-based, hybrid, review strategies to assist reviewers in achieving high recall and high precision technology-assisted review results. Six unique strategies were presented, of which, two review strategies (CAL QC–Type 1 and CAL QC–Type 2) are novel. With a budget constraint to establish a level playing field (B = R, 2R, or 3R), we have compared each of the six multi-user, hybrid, human-computer assessment strategies against the single-user, hybrid, human-computer assessment strategy as a baseline, leveraging 25 unique, simulated users. Across all four datasets used in this study, we observe that our proposed quality control review strategies consistently perform better than the single-user review strategy, with respect to both system recall and precision; as well as end-to-end recall and precision.

The multi-user review strategies address several of the major challenges associated with eDiscovery: achieving the maximum possible recall with the limited availability of resources, and accommodating declining budgets. We have also identified the preferred conditions for using a particular multi-user review strategy. When every relevant document is critical and we seek to achieve a reasonably high recall, it is recommended to use the CAL QC–Type 1 review strategy. When we seek to retrieve the maximum number of relevant documents with little to no compromise on missing a few relevant documents, it is advisable to use the CAL QC–Type 2 review strategy to obtain optimal results. From the Paired Two-Sample T-Test results, we conclude that the superior performance in terms of recall of both the CAL QC–Type 1 and Type 2 review strategies to be statistically significant compared to the single-user review strategy. This safely qualifies CAL QC–Type 1 and Type 2 to be potentially better alternatives to the single-user review strategy in achieving high recall results.

Limitations

1. The review budget B used in our study depends on the number of relevant documents (R) present in the corpus associated with a query topic. Allocating the review budget as a factor of R is challenging because we don't know R's value in advance. Hence, the review budget criteria used in this research is not practical to be deployed in real-time. We only utilise this artificial review budget to assist with our simulation.

2. Simulating human relevance feedback error is challenging. Although the number of relevance feedback errors (FNs and FPs) to be induced in each batch are predetermined based upon the reviewer's user recall and user precision, the error-simulation technique used in this study entails inducing a fixed number of errors at random. Because human reviewers do not tend to make a fixed number of random errors, this error-simulation technique does not reflect true human error, which is difficult to model.

3. To simulate the reviewers in our study, we have assumed them to have the same recall and precision rate. For example, when we refer to 80User, we assume the user to have 80% user recall and 80% user precision. Finding and engaging reviewers, in practice, having equal recall and precision rate is unlikely. The impact of the reviewer with non-similar recall and precision rate has not been studied, e.g., a reviewer with 82% user recall and 87% user precision. Hence, our work does not study the impact of different possible combinations in user effectiveness.

4. When leveraging multiple reviewers to review documents, there often exists an overhead cost of management of those reviewers. This study does not address/allocate the budget involved in managing the workflows for each of these multi-user review strategies. A few examples of overhead management costs include engaging a supervisor to monitor/delegate documents to each reviewer and the cost involved in running one or more CAL systems.

Future Work

1. Among all of the multi-user review strategies, the CAL QC - Type 2 review methodology obtains the best end-to-end performance; we chose a depth of 2B/3 for the first user; future experimentation can use other depths to determine the appropriate depth for this strategy to be effectively adopted.

2. From CAL QC–Type 2, we also understand that using the system as one of the users can help achieve high system and end-to-end recall. The BMI system we used employs a logistic regression model to assign a confidence score (ranging from -4 to +6), indicating the document's relevance to the topic and allowing it to rank documents for relevance feedback. Instead of employing the system as a second user, it might be more advantageous to use the system as an adjudicator. For example, if User 1 and User 2 disagree about the relevance of a document, the TAR strategy can automatically tag the document as relevant if the confidence score for the document generated by the system is greater than, say 2.0, and not relevant if it is less than that score.

3. The cost of engaging a human reviewer is expensive but running a system for longer to rank potentially relevant documents is comparatively cost-effective. We can formulate a review strategy that runs the system automatically to tag a set of documents based on the confidence score and attempt to utilize the human review budget efficiently.

References

[1] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. de Vries, and E. Yilmaz. *Relevance assessment: are judges exchangeable and does it matter?* 2008.

[2] William W. Belt, Dennis R. Kiker, and Daryl E. Shetterly. *Technology-assisted document review: Is it defensible.* Berkeley Tech. LJ 35 : 171, 2020.

[3] Carla E. Brodley and Mark A. Friedl. *Identifying mislabeled training data.* 1999.

[4] Shannon Brown. *Peeking inside the black box: A preliminary survey of technology assisted review (tar) and predictive coding algorithms for ediscovery.* 2015.

[5] Stefan Buttcher, Charles LA Clarke, and Gordon V. Cormack. *Information retrieval: Implementing and evaluating search engines.* 2016.

[6] Max W Callaghan and Finn Müller-Hansen. *Statistical stopping criteria for automated screening in systematic reviews.* Systematic Reviews 9, 1, 1–14, 2020.

[7] Ben Carterette and Ian Soboroff. *The effect of assessor errors on IR system evaluation.* 2010.

[8] 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, pages 153–162,, 2014.

[9] Gordon V. Cormack and Maura R. Grossman. *The Grossman-cormack glossary of technology-assisted review.* Federal Courts Law Review, Rev. 7 : 85, 2014.

[10] Gordon V Cormack and Maura R Grossman. *Autonomy and reliability of continuous active learning for technology-assisted review.* arXiv preprint arXiv:1504.06868, 2015.

[11] Gordon V. Cormack and Maura R. Grossman. *Engineering Quality and Reliability in Technology-Assisted Review*. SIGIR '16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pages 75–84, 2016.

[12] Gordon V Cormack and Maura R Grossman. *Scalability of continuous active learning for reliable high-recall text classification.* In Proceedings of the 25th ACM international on conference on information and knowledge management (pp. 1039-1048), 2016.

[13] Gordon V. Cormack and Maura R. Grossman. *Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me.* In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 5–14, 2017.

[14] Gordon V Cormack and Mona Mojdeh. *Machine Learning for Information Retrieval: TREC 2009 web, relevance feedback and legal tracks.* In Proc. TREC-2009, 2009.

[15] David Dowling. *Tarpits: The Sticky Consequences of Poorly Implementing Technology-Assisted Review.* Berkeley Tech. LJ 35 : 171, 2020.

[16] Yang Eugene, Sean MacAvaney, David D. Lewis, and Ophir Frieder. *Goldilocks: Just-right tuning of bert for technology-assisted review.* In European Conference on Information Retrieval, pp. 502-517. Springer, Cham, 2022.

[17] Maura R. Grossman and Gordon V. Cormack. *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review.* Richmond Journal of Law and Technology, Vol. 17, Issue 3, 2011.

[18] Maura R. Grossman and Gordon V. Cormack. *Comments on "The implications of Rule 26 (g) on the use of technology-assisted review.* Federal Courts Law Review 1, 2014.

[19] Maura R. Grossman and Gordon V. Cormack. *Continuous Active Learning for TAR.* The Journal 4.3 : 1-7, 2016.

[20] Bruce Hedin, Stephen Tomlinson, Jason R. Baron, and Douglas W. Oard. *Overview of the TREC 2008 Legal Track.* In TREC, 2008.

[21] Bruce Hedin, Stephen Tomlinson, Jason R. Baron, and Douglas W. Oard. *Overview of the TREC 2009 Legal Track.* In TREC, 2009.

[22] Wong K, Chan AHS, and Ngan SC. *The Effect of Long Working Hours and Overtime on Occupational Health: A Meta-Analysis of Evidence from 1998 to 2018.* Int J Environ Res Public Health, 16(12):2102, 2019.

[23] Emery G. Lee and Thomas E. Willging. *Defining the Problem of Cost in Federal Civil Litigation.* Duke Law Journal, 2010.

[24] Dan Li and Evangelos Kanoulas. *When to Stop Reviewing in Technology Assisted Reviews: Sampling from an Adaptive Distribution to Estimate Residual Relevant Documents.* ACM Transactions on Information Systems (TOIS) 38, 4, 1–36, 2020.

[25] Haotian Zhang Mark D. Smucker Gordon V. Cormack Mustafa Abualsaud, Nimesh Ghelani and Maura R. Grossman. *A system for efficient high-recall retrieval.* The 41st international ACM SIGIR conference on research development in information retrieval, pages 1317–1320, 2018.

[26] Nicholas M. Pace and Laura Zakaras. *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery.* RAND Institute for Civil Justice, 2012.

[27] Taemin Kim Park. *The Nature of Relevance in Information Retrieval: An Empirical Study.* The Library Quarterly: Information, Community, Policy, Vol. 63, No. 3, pages 318-351, The University of Chicago Press, 1993.

[28] G. L. Paul and J. R. Baron. *Information inflation: Can the legal system adapt?* Richmond Journal of Law and Technology, vol. 13, 2007.

[29] Ganesh Ramakrishnan, Krishna Prasad Chitrapura, Raghu Krishnapuram, and Pushpak Bhattarcharyy. *A model for handling approximate, noisy or incomplete labeling in text classification.* 2005.

[30] Adam Roegiest and Gordon V Cormack. *Total Recall Track Tools Architecture Overview.* In Proc. TREC-2015, 2015.

[31] Adam Roegiest, Gordon V. Cormack, Maura R. Grossman, and Charles L.A. Clarke. *TREC 2015 Total Recall Track Overview.* In TREC, 2015.

[32] Adam Roegiest, Gordon V. Cormack, Maura R. Grossman, and Charles L.A. Clarke. *TREC 2016 Total Recall Track Overview.* In TREC, 2016.

[33] Manali Sharma and Mustafa Bilgic. *Evidence-Based Uncertainty Sampling for Active Learning.* Data Mining and Knowledge Discovery 31.1 : 164-202, 2017.

[34] Ellen M. Voorhees. *Variations in relevance judgments and the measurement of retrieval effectiveness.* Information Processing & Management 36(5), 2000.

[35] Ellen M. Voorhees. *The philosophy of information retrieval evaluation.* Workshop of the cross-language evaluation forum for european languages. Springer, Berlin, Heidelberg, 2001.

[36] Ellen M. Voorhees. *The evolution of cranfield.* Information retrieval evaluation in a changing world., Springer, Cham, 45-69, 2019.

[37] Ellen M. Voorhees and Donna K. Harman. *TREC Ad Hoc Experiments.* 2005.

[38] Oard Douglas W. and William Webber. *Information retrieval for e-discovery.* Information Retrieval 7(2-3): 99-237, 2013.

[39] William Webber. *Re-examining the effectiveness of manual review.* 2011.

[40] Hui Yang, John Frank, and Ian Soboroff. *Trec 2015 dynamic domain track overview.* In Proc. TREC-2015, 2015.

APPENDICES

Appendix A

Topic Descriptions

A.1 At-Home4 Dataset Topics

Topic	Title	Description
401	Olympics	Bid to host the Olympic games in Florida.
402	Space	The space industry, space program, space travel, or space science, public and private, in Florida.
403*	Bottled Water	Extraction of water for bottling by commercial enterprises.
404	Eminent domain	Legality or morality of expropriating land for commercial development.
405	Newt Gingrich	Speaker Newt Gingrich or any entities or personnel associated with Newt Gingrich.
406	Felon disenfranchisement	Right of felons to vote, including but not restricted to voter purges and reinstatement of voter rights. Individual clemency cases are not relevant.
407	Faith-based initiatives	Grants or other initiatives to offload social services to so-called faith-based agencies. Services include but are not limited to education, prisons, and emergency relief.
408*	Invasive species	The problem of invasive species – non-native plants or animals that threaten the ecosystem.
409*	Climate change	Climate change, global warming, or carbon emissions.
410	Condos	Rules and organizations governing condominium associations and conflicts between owners and managers. Relevant documents include those concerning the establishment of the office of ombudsman, and issues relating to hiring and firing the ombudsman.
411	Stand your ground	Use of deadly force to protect one's self or one's property.
412	2000 Recount	Contested result of the 2000 presidential election.
413	James V. Crosby	James V. Crosby, including but not limited to his relationship with Gov. Bush before being appointed Secretary of Corrections, his role as Secretary, his firing, and criminal allegations against Mr. Crosby.
414*	Medicaid reform	Efforts to substantially reform Medicaid.
415	George W. Bush	Documents referring to George W. Bush, whether explicitly or by his relationship to Gov. bush.
416*	Marketing	Advertising or marketing efforts undertaken by the Governor's office or institutions of the State of Florida.
417	Movie Gallery	Investments by Florida in Movie Gallery.
418	War preparations	Preparations for the Iraq War undertaken before the March 20, 2003 invasion.
419	Lost foster child	Disappearance of Rilya Wilson and its aftermath.
420	Billboards	Rights and control of billboards. Distinct legislative efforts should be considered to be separate categories.
421	Traffic cameras	Use of unattended cameras to enforce traffic laws.
422*	Non-resident Aliens	Non-resident alien issue. Documents concerning the National Rifle Association are not relevant.
423*	National Rifle Association	The NRA, its members, and its influences.
424	Gulf drilling	Off-shore drilling for oil or gas. Water drilling is not relevant.
425*	Civil Rights Act	Civil Rights Act of 2003.
426	Jeffrey Goldhagen	Jeffrey Goldhagen's role in the administration, his firing, and reinstatement.
427	Slot Machines	Legality/licensing/definition of "slot machines."
428	New Stadiums	Construction of new sports stadiums or arenas.
429*	Cuban Child	Elian Gonzales and his status.
430*	Restraints and Helmets	Seat belt, child seat, and helmet mandates.
431	Agency Ratings	Credit ratings of Florida institutions, particularly those by Standard and Poor's, Fitch, and Moody's.
432	Gay Adoption	Gay adoption issue.
433*	Abstinence	Abstinence and abstinence-only programs to supplant birth control or sex education.
434*	Bacardi Trademark Lobbying	The Jeb Bush administration's involvement in a trademark dispute between Bacardi and the U.S. Patent and Trademark Office.

Figure A.1: Topics and Topic Descriptions for the At-Home4 Collection

Appendix B

Full Metric Tables

B.1 At-Home1 Dataset Results

Budget B = 3R		Reviewer / System Metrics	IdealUser	60User	70User	80User	90User
Review Models	**Single-User CAL**	Recall	0.9548	0.6790	0.7672	0.8435	0.9150
		Precision	0.3178	0.2260	0.2554	0.2807	0.3046
		F1 Score	0.4769	0.3391	0.3832	0.4212	0.4570
	Separate CAL	Recall		0.6113	0.6969	0.7597	0.8429
		Precision		0.3153	0.3812	0.4341	0.5062
		F1 Score		0.3901	0.4928	0.5525	0.6325
	Lock-Step CAL–Type 1	Recall		0.6262	0.7233	0.7866	0.8576
		Precision		0.3360	0.3936	0.4507	0.5227
		F1 Score		0.4101	0.5098	0.5731	0.6495
	Lock-Step CAL–Type 2	Recall		0.6262	0.7233	0.7866	0.8576
		Precision		0.3360	0.3936	0.4507	0.5227
		F1 Score		0.4101	0.5098	0.5731	0.6495
	Majority-Vote of-Three	Recall		0.5539	0.6324	0.6904	0.7076
		Precision		0.5539	0.6324	0.6904	0.7076
		F1 Score		0.5539	0.6324	0.6904	0.7076
	CAL QC –Type 1	Recall		0.6600	0.7856	0.8532	0.8744
		Precision		0.4784	0.5621	0.6424	0.6783
		F1 Score		0.5547	0.6553	0.7329	0.7640
	CAL QC –Type 2	Recall		0.7282	0.8094	0.8787	0.8961
		Precision		0.3365	0.3878	0.4136	0.4403
		F1 Score		0.4603	0.5244	0.5625	0.5905

Table B.1: **At-Home1 Dataset: System–Retrieval Metrics at Budget B=3R**

Budget B = 3R		Reviewer / User Metrics	IdealUser	60User	70User	80User	90User
Review Models	Single-User CAL	Recall	1.0000	0.6082	0.6999	0.7993	0.9067
		Precision	1.0000	0.6038	0.7096	0.8063	0.9002
		F1 Score	1.0000	0.6060	0.7047	0.8028	0.9034
	Separate CAL	Recall		0.5894	0.7034	0.8044	0.9018
		Precision		0.6024	0.6947	0.8050	0.9047
		F1 Score		0.5958	0.6990	0.8047	0.9032
	Lock-Step CAL–Type 1	Recall		0.6113	0.7289	0.8299	0.9271
		Precision		0.6286	0.7209	0.8312	0.9320
		F1 Score		0.6198	0.7249	0.8305	0.9296
	Lock-Step CAL–Type 2	Recall		0.5850	0.6933	0.8037	0.9037
		Precision		0.6031	0.6954	0.8072	0.9057
		F1 Score		0.5939	0.6944	0.8055	0.9047
	Majority-Vote of-Three	Recall		0.7702	0.8833	0.9625	0.9975
		Precision		0.7708	0.8826	0.9626	0.9977
		F1 Score		0.7705	0.8830	0.9625	0.9976
	CAL QC –Type 1	Recall		0.6968	0.8245	0.9304	0.9796
		Precision		0.6965	0.8251	0.9306	0.9795
		F1 Score		0.6967	0.8248	0.9305	0.9796
	CAL QC –Type 2	Recall		0.7254	0.8327	0.9005	0.9550
		Precision		0.7261	0.8336	0.8853	0.9560
		F1 Score		0.7257	0.8332	0.8929	0.9555

Table B.2: **At-Home1 Dataset: User–Retrieval Metrics at Budget B=3R**

Budget B = 3R		Reviewer / End-To-End Metrics	IdealUser	60User	70User	80User	90User
Review Models	Single-User CAL	Recall	0.9548	0.4130	0.5370	0.6742	0.8296
		Precision	1.0000	0.6038	0.7096	0.8063	0.9002
		F1 Score	0.9769	0.4905	0.6114	0.7344	0.8635
	Separate CAL	Recall		0.3603	0.4902	0.6111	0.7601
		Precision		0.6024	0.6947	0.8050	0.9047
		F1 Score		0.4096	0.5748	0.6948	0.8261
	Lock-Step CAL–Type 1	Recall		0.3828	0.5272	0.6528	0.7951
		Precision		0.6286	0.7209	0.8312	0.9320
		F1 Score		0.4352	0.6090	0.7313	0.8581
	Lock-Step CAL–Type 2	Recall		0.3663	0.5015	0.6322	0.7750
		Precision		0.6031	0.6954	0.8072	0.9057
		F1 Score		0.4150	0.5827	0.7091	0.8353
	Majority-Vote of-Three	Recall		0.4266	0.5586	0.6645	0.7058
		Precision		0.7708	0.8826	0.9626	0.9977
		F1 Score		0.5492	0.6842	0.7862	0.8267
	CAL QC –Type 1	Recall		0.4599	0.6477	0.7938	0.8566
		Precision		0.6965	0.8251	0.9306	0.9795
		F1 Score		0.5540	0.7257	0.8568	0.9139
	CAL QC –Type 2	Recall		0.5282	0.6740	0.7913	0.8558
		Precision		0.7261	0.8336	0.8853	0.9560
		F1 Score		0.6115	0.7454	0.8357	0.9031

Table B.3: **At-Home1 Dataset: End-To-End–Retrieval Metrics at Budget B=3R**

B.2 At-Home2 Dataset Results

Budget B = 3R		Reviewer / System Metrics	IdealUser	60User	70User	80User	90User
Review Models	Single-User CAL	Recall	0.9183	0.6530	0.7379	0.8113	0.8800
		Precision	0.2955	0.2074	0.2356	0.2599	0.2828
		F1 Score	00.4471	0.3148	0.3572	0.3937	0.4280
	Separate CAL	Recall		0.5932	0.6721	0.7327	0.8130
		Precision		0.3054	0.3693	0.4205	0.4904
		F1 Score		0.3772	0.4767	0.5343	0.6118
	Lock-Step CAL–Type 1	Recall		0.6075	0.6976	0.7587	0.8271
		Precision		0.3255	0.3813	0.4366	0.5064
		F1 Score		0.3966	0.4931	0.5543	0.6282
	Lock-Step CAL–Type 2	Recall		0.6075	0.6976	0.7587	0.8271
		Precision		0.3255	0.3813	0.4366	0.5064
		F1 Score		0.3966	0.4931	0.5543	0.6282
	Majority-Vote of-Three	Recall		0.5252	0.5997	0.6547	0.6710
		Precision		0.5252	0.5997	0.6547	0.6710
		F1 Score		0.5252	0.5997	0.6547	0.6710
	CAL QC –Type 1	Recall		0.6247	0.7555	0.8205	0.8409
		Precision		0.4718	0.5543	0.6335	0.6689
		F1 Score		0.5412	0.6394	0.7150	0.7451
	CAL QC –Type 2	Recall		0.6913	0.7684	0.8342	0.8507
		Precision		0.3213	0.3703	0.3949	0.4204
		F1 Score		0.4387	0.4998	0.5360	0.5627

Table B.4: **At-Home2 Dataset: System–Retrieval Metrics at Budget B=3R**

Budget B = 3R		Reviewer / User Metrics	IdealUser	60User	70User	80User	90User
Review Models	**Single-User CAL**	Recall	1.0000	0.6084	0.7000	0.7991	0.9067
		Precision	1.0000	0.6017	0.7075	0.8042	0.8981
		F1 Score	1.0000	0.6050	0.7037	0.8016	0.9024
	Separate CAL	Recall		0.5883	0.7027	0.8037	0.9010
		Precision		0.6031	0.6954	0.8057	0.9054
		F1 Score		0.5956	0.6990	0.8047	0.9032
	Lock-Step CAL–Type 1	Recall		0.6105	0.7289	0.8298	0.9272
		Precision		0.6291	0.7214	0.8317	0.9325
		F1 Score		0.6197	0.7251	0.8308	0.9299
	Lock-Step CAL–Type 2	Recall		0.5845	0.6934	0.8036	0.9038
		Precision		0.6021	0.6944	0.8062	0.9047
		F1 Score		0.5932	0.6939	0.8049	0.9042
	Majority-Vote of-Three	Recall		0.7696	0.8826	0.9618	0.9969
		Precision		0.7703	0.8821	0.9621	0.9972
		F1 Score		0.7700	0.8824	0.9620	0.9970
	CAL QC –Type 1	Recall		0.7085	0.8250	0.9309	0.9803
		Precision		0.6968	0.8254	0.9309	0.9798
		F1 Score		0.7026	0.8252	0.9309	0.9800
	CAL QC –Type 2	Recall		0.7269	0.8345	0.9024	0.9570
		Precision		0.7242	0.8317	0.8834	0.9541
		F1 Score		0.7255	0.8331	0.8928	0.9555

Table B.5: **At-Home2 Dataset: User–Retrieval Metrics at Budget B=3R**

Budget B = 3R		Reviewer End-To-End Metrics	IdealUser	60User	70User	80User	90User
Review Models	Single-User CAL	Recall	0.9183	0.3973	0.5165	0.6483	0.7979
		Precision	1.0000	0.6017	0.7075	0.8042	0.8981
		F1 Score	0.9574	0.4786	0.5971	0.7179	0.8450
	Separate CAL	Recall		0.3490	0.4723	0.5889	0.7325
		Precision		0.6031	0.6954	0.8057	0.9054
		F1 Score		0.3998	0.5625	0.6804	0.8098
	Lock-Step CAL–Type 1	Recall		0.3709	0.5085	0.6296	0.7669
		Precision		0.6291	0.7214	0.8317	0.9325
		F1 Score		0.4250	0.5965	0.7167	0.8416
	Lock-Step CAL–Type 2	Recall		0.3551	0.4837	0.6097	0.7475
		Precision		0.6021	0.6944	0.8062	0.9047
		F1 Score		0.4049	0.5702	0.6943	0.8186
	Majority-Vote of-Three	Recall		0.4042	0.5293	0.6297	0.6689
		Precision		0.7703	0.8821	0.9621	0.9972
		F1 Score		0.5302	0.6616	0.7612	0.8007
	CAL QC –Type 1	Recall		0.4426	0.6233	0.7638	0.8243
		Precision		0.6968	0.8254	0.9309	0.9798
		F1 Score		0.5413	0.7103	0.8391	0.8953
	CAL QC –Type 2	Recall		0.5025	0.6412	0.7528	0.8141
		Precision		0.7242	0.8317	0.8834	0.9541
		F1 Score		0.5933	0.7241	0.8129	0.8786

Table B.6: **At-Home2 Dataset: End-To-End–Retrieval Metrics at Budget B=3R**

B.3 At-Home3 Dataset Results

Budget B = 3R		Reviewer / System Metrics	IdealUser	60User	70User	80User	90User
Review Models	Single-User CAL	Recall	0.9016	0.6411	0.7245	0.7965	0.8640
		Precision	0.3918	0.2786	0.3149	0.3461	0.3755
		F1 Score	0.5462	0.3884	0.4390	0.4825	0.5235
	Separate CAL	Recall		0.6035	0.6863	0.7481	0.8301
		Precision		0.3654	0.4417	0.5031	0.5866
		F1 Score		0.4235	0.5375	0.6016	0.6874
	Lock-Step CAL–Type 1	Recall		0.6182	0.7123	0.7746	0.8445
		Precision		0.3894	0.4562	0.5223	0.6058
		F1 Score		0.4447	0.5562	0.6239	0.7055
	Lock-Step CAL–Type 2	Recall		0.6182	0.7123	0.7746	0.8445
		Precision		0.3894	0.4562	0.5223	0.6058
		F1 Score		0.4447	0.5562	0.6239	0.7055
	Majority-Vote of-Three	Recall		0.6110	0.6977	0.7616	0.7806
		Precision		0.6110	0.6977	0.7616	0.7806
		F1 Score		0.6110	0.6977	0.7616	0.7806
	CAL QC –Type 1	Recall		0.6233	0.7419	0.8057	0.8257
		Precision		0.5349	0.6286	0.7183	0.7585
		F1 Score		0.5757	0.6806	0.7595	0.7907
	CAL QC –Type 2	Recall		0.6962	0.7849	0.8522	0.8690
		Precision		0.4204	0.4846	0.5167	0.5501
		F1 Score		0.5271	0.5992	0.6433	0.6737

Table B.7: **At-Home3 Dataset: System–Retrieval Metrics at Budget B=3R**

Budget B = 3R		Reviewer / User Metrics	IdealUser	60User	70User	80User	90User
Review Models	Single-User CAL	Recall	1.0000	0.6083	0.6999	0.7992	0.9067
		Precision	1.0000	0.6011	0.7069	0.8036	0.8975
		F1 Score	1.0000	0.6047	0.7034	0.8014	0.9021
	Separate CAL	Recall		0.5884	0.7023	0.8034	0.9007
		Precision		0.6022	0.6945	0.8048	0.9045
		F1 Score		0.5952	0.6984	0.8041	0.9026
	Lock-Step CAL–Type 1	Recall		0.6110	0.7289	0.8300	0.9272
		Precision		0.6304	0.7227	0.8330	0.9338
		F1 Score		0.6205	0.7258	0.8315	0.9305
	Lock-Step CAL–Type 2	Recall		0.5848	0.6934	0.8038	0.9037
		Precision		0.6036	0.6959	0.8077	0.9062
		F1 Score		0.5940	0.6946	0.8057	0.9050
	Majority-Vote of-Three	Recall		0.7709	0.8839	0.9632	0.9982
		Precision		0.7713	0.8831	0.9631	0.9982
		F1 Score		0.7711	0.8835	0.9632	0.9982
	CAL QC –Type 1	Recall		0.7613	0.8383	0.9307	0.9800
		Precision		0.6967	0.8253	0.9308	0.9797
		F1 Score		0.7276	0.8317	0.9308	0.9799
	CAL QC –Type 2	Recall		0.7822	0.8365	0.9046	0.9594
		Precision		0.7254	0.8329	0.8846	0.9553
		F1 Score		0.7528	0.8347	0.8945	0.9573

Table B.8: **At-Home3 Dataset: User–Retrieval Metrics at Budget B=3R**

Budget B = 3R		Reviewer End-To-End Metrics	IdealUser	60User	70User	80User	90User
Review Models	Single-User CAL	Recall	0.9016	0.3900	0.5071	0.6366	0.7834
		Precision	1.0000	0.6011	0.7069	0.8036	0.8975
		F1 Score	0.9483	0.4731	0.5906	0.7104	0.8366
	Separate CAL	Recall		0.3551	0.4820	0.6010	0.7477
		Precision		0.6022	0.6945	0.8048	0.9045
		F1 Score		0.4050	0.5691	0.6881	0.8187
	Lock-Step CAL–Type 1	Recall		0.3777	0.5192	0.6429	0.7830
		Precision		0.6304	0.7227	0.8330	0.9338
		F1 Score		0.4312	0.6043	0.7257	0.8518
	Lock-Step CAL–Type 2	Recall		0.3615	0.4939	0.6226	0.7632
		Precision		0.6036	0.6959	0.8077	0.9062
		F1 Score		0.4109	0.5778	0.7032	0.8286
	Majority-Vote of-Three	Recall		0.4710	0.6167	0.7336	0.7792
		Precision		0.7713	0.8831	0.9631	0.9982
		F1 Score		0.5849	0.7262	0.8328	0.8752
	CAL QC –Type 1	Recall		0.4745	0.6219	0.7499	0.8092
		Precision		0.6967	0.8253	0.9308	0.9797
		F1 Score		0.5352	0.7093	0.8306	0.8863
	CAL QC –Type 2	Recall		0.5446	0.6566	0.7709	0.8337
		Precision		0.7254	0.8329	0.8846	0.9553
		F1 Score		0.6021	0.7343	0.8238	0.8904

Table B.9: **At-Home3 Dataset: End-To-End–Retrieval Metrics at Budget B=3R**

B.4 At-Home4 Dataset Results

Budget B = 3R		Reviewer / System Metrics	IdealUser	60User	70User	80User	90User
Review Models	Single-User CAL	Recall	0.9153	0.6509	0.7355	0.8086	0.8771
		Precision	0.3051	0.2170	0.2452	0.2695	0.2924
		F1 Score	0.4577	0.3255	0.3678	0.4043	0.4386
	Separate CAL	Recall		0.5993	0.6805	0.7418	0.8231
		Precision		0.3038	0.3673	0.4183	0.4878
		F1 Score		0.3778	0.4771	0.5349	0.6126
	Lock-Step CAL–Type 1	Recall		0.6138	0.7063	0.7681	0.8374
		Precision		0.3238	0.3793	0.4343	0.5037
		F1 Score		0.3973	0.4936	0.5549	0.6290
	Lock-Step CAL–Type 2	Recall		0.6138	0.7063	0.7681	0.8374
		Precision		0.3238	0.3793	0.4343	0.5037
		F1 Score		0.3973	0.4936	0.5549	0.6290
	Majority-Vote of-Three	Recall		0.5273	0.6135	0.6697	0.6864
		Precision		0.5273	0.6135	0.6697	0.6864
		F1 Score		0.5273	0.6135	0.6697	0.6864
	CAL QC –Type 1	Recall		0.6327	0.7531	0.8179	0.8382
		Precision		0.4726	0.5553	0.6346	0.6701
		F1 Score		0.5411	0.6392	0.7147	0.7448
	CAL QC –Type 2	Recall		0.7048	0.7834	0.8505	0.8673
		Precision		0.3284	0.3785	0.4036	0.4297
		F1 Score		0.4480	0.5104	0.5474	0.5747

Table B.10: **At-Home4 Dataset: System–Retrieval Metrics at Budget B=3R**

Budget B = 3R		Reviewer / User Metrics	IdealUser	60User	70User	80User	90User
Review Models	Single-User CAL	Recall	1.0000	0.5986	0.6937	0.7891	0.9051
		Precision	1.0000	0.5950	0.7008	0.7975	0.8914
		F1 Score	1.0000	0.5968	0.6972	0.7933	0.8982
	Separate CAL	Recall		0.5880	0.7022	0.8032	0.9005
		Precision		0.6028	0.6951	0.8054	0.9051
		F1 Score		0.5953	0.6986	0.8043	0.9028
	Lock-Step CAL–Type 1	Recall		0.6108	0.7289	0.8299	0.9272
		Precision		0.6292	0.7215	0.8318	0.9326
		F1 Score		0.6199	0.7252	0.8309	0.9299
	Lock-Step CAL–Type 2	Recall		0.5846	0.6934	0.8037	0.9037
		Precision		0.6028	0.6951	0.8069	0.9054
		F1 Score		0.5935	0.6942	0.8053	0.9045
	Majority-Vote of-Three	Recall		0.7844	0.8828	0.9621	0.9971
		Precision		0.7705	0.8823	0.9623	0.9974
		F1 Score		0.7774	0.8826	0.9622	0.9972
	CAL QC –Type 1	Recall		0.6969	0.8246	0.9306	0.9798
		Precision		0.6966	0.8252	0.9307	0.9796
		F1 Score		0.6967	0.8249	0.9306	0.9797
	CAL QC –Type 2	Recall		0.7253	0.8327	0.9005	0.9550
		Precision		0.7249	0.8324	0.8841	0.9548
		F1 Score		0.7251	0.8325	0.8922	0.9549

Table B.11: **At-Home4 Dataset: User–Retrieval Metrics at Budget B=3R**

Budget B = 3R		Reviewer End-To-End Metrics	IdealUser	60User	70User	80User	90User
Review Models	Single-User CAL	Recall	0.9153	0.3896	0.5102	0.6381	0.7939
		Precision	1.0000	0.5950	0.7008	0.7975	0.8914
		F1 Score	0.9558	0.4709	0.5905	0.7090	0.8398
	Separate CAL	Recall		0.3524	0.4778	0.5958	0.7412
		Precision		0.6028	0.6951	0.8054	0.9051
		F1 Score		0.4028	0.5663	0.6849	0.8150
	Lock-Step CAL–Type 1	Recall		0.3749	0.5148	0.6375	0.7764
		Precision		0.6292	0.7215	0.8318	0.9326
		F1 Score		0.4286	0.6009	0.7218	0.8474
	Lock-Step CAL–Type 2	Recall		0.3588	0.4897	0.6173	0.7568
		Precision		0.6028	0.6951	0.8069	0.9054
		F1 Score		0.4084	0.5746	0.6995	0.8244
	Majority-Vote of-Three	Recall		0.4136	0.5416	0.6443	0.6844
		Precision		0.7705	0.8823	0.9623	0.9974
		F1 Score		0.5383	0.6712	0.7718	0.8118
	CAL QC –Type 1	Recall		0.4409	0.6210	0.7611	0.8213
		Precision		0.6966	0.8252	0.9307	0.9796
		F1 Score		0.5400	0.7087	0.8374	0.8935
	CAL QC –Type 2	Recall		0.5112	0.6523	0.7659	0.8283
		Precision		0.7249	0.8324	0.8841	0.9548
		F1 Score		0.5996	0.7314	0.8208	0.8871

Table B.12: **At-Home4 Dataset: End-To-End–Retrieval Metrics at Budget B=3R**

B.5 Lock-Step CAL–Type 2 Results

User 2 / User 1	**End-to-End Recall@B=R**					**End-to-End Recall@B=2R**					**End-to-End Recall@B=3R**				
	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	0.4632	0.4632	0.4632	0.4632	0.4632	0.7114	0.7114	0.7114	0.7114	0.7114	0.8673	0.8673	0.8673	0.8673	0.8673
90User	0.4179	0.4115	0.3951	0.3883	0.3848	0.6410	0.5878	0.5678	0.5599	0.5543	0.7827	0.7750	0.7541	0.7454	0.7347
80User	0.3708	0.3537	0.3443	0.3182	0.3075	0.5673	0.5066	0.4616	0.4345	0.4252	0.6929	0.6674	0.6322	0.6013	0.5956
70User	0.3260	0.3033	0.2751	0.2503	0.2240	0.5013	0.4383	0.3793	0.3312	0.3089	0.6082	0.5771	0.5251	0.5015	0.4834
60User	0.2793	0.2565	0.2319	0.1922	0.1629	0.4254	0.3710	0.3198	0.2633	0.2448	0.5186	0.4921	0.4443	0.4145	0.3163

User 2 / User 1	**End-to-End Precision@B=R**					**End-to-End Precision@B=2R**					**End-to-End Precision@B=3R**				
	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
90User	0.9065	0.8985	0.9078	0.8999	0.9010	0.9026	0.9025	0.8998	0.8999	0.9017	0.9044	0.9057	0.9059	0.9018	0.8992
80User	0.8037	0.8054	0.8025	0.8015	0.8030	0.7990	0.8029	0.8099	0.8022	0.8041	0.8009	0.8020	0.8072	0.8015	0.8025
70User	0.7048	0.7055	0.7007	0.6958	0.6996	0.7062	0.7048	0.7006	0.6955	0.7052	0.7032	0.6987	0.7002	0.6954	0.7014
60User	0.5990	0.6019	0.6054	0.5996	0.5986	0.5995	0.6040	0.6051	0.6013	0.6048	0.6000	0.6029	0.5992	0.6017	0.6031

User 2 / User 1	**End-to-End F1-Score@B=R**					**End-to-End F1-Score@B=2R**					**End-to-End F1-Score@B=3R**				
	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	0.6331	0.6331	0.6331	0.6331	0.6331	0.8314	0.8314	0.8314	0.8314	0.8314	0.9289	0.9289	0.9289	0.9289	0.9289
90User	0.5720	0.5645	0.5506	0.5425	0.5393	0.7497	0.7119	0.6963	0.6903	0.6865	0.8391	0.8353	0.8230	0.8162	0.8087
80User	0.5075	0.4915	0.4818	0.4555	0.4447	0.6635	0.6212	0.5880	0.5637	0.5563	0.7430	0.7285	0.7091	0.6871	0.6837
70User	0.4458	0.4242	0.3951	0.3681	0.3393	0.5864	0.5405	0.4922	0.4487	0.4296	0.6522	0.6321	0.6001	0.5828	0.5723
60User	0.3810	0.3597	0.3353	0.2911	0.2561	0.4977	0.4596	0.4185	0.3662	0.3485	0.5564	0.5419	0.5103	0.4908	0.4150

Table B.13: **At-Home1 Dataset: End-To-End–Retrieval Metrics obtained using Lock-Step CAL–Type 2 review strategy for budget B={R, 2R, 3R}**

	End-to-End Recall@B=R					**End-to-End Recall@B=2R**					**End-to-End Recall@B=3R**				
User 2 / User 1	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	0.4470	0.4470	0.4470	0.4470	0.4470	0.6753	0.6753	0.6753	0.6753	0.6753	0.8365	0.8365	0.8365	0.8365	0.8365
90User	0.4032	0.3971	0.3813	0.3747	0.3714	0.6085	0.5580	0.5390	0.5315	0.5261	0.7549	0.7475	0.7273	0.7189	0.7086
80User	0.3578	0.3413	0.3322	0.3071	0.2967	0.5386	0.4809	0.4382	0.4124	0.4037	0.6683	0.6437	0.6097	0.5799	0.5744
70User	0.3146	0.2927	0.2655	0.2415	0.2161	0.4759	0.4161	0.3601	0.3144	0.2932	0.5866	0.5566	0.5065	0.4837	0.4662
60User	0.2695	0.2476	0.2238	0.1855	0.1572	0.4038	0.3521	0.3036	0.2499	0.2323	0.5002	0.4746	0.4285	0.3998	0.3051

	End-to-End Precision@B=R					**End-to-End Precision@B=2R**					**End-to-End Precision@B=3R**				
User 2 / User 1	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
90User	0.9055	0.8975	0.9068	0.8989	0.9000	0.9016	0.9015	0.8988	0.8989	0.9007	0.9034	0.9047	0.9049	0.9008	0.8982
80User	0.8027	0.8044	0.8015	0.8005	0.8020	0.7980	0.8019	0.8089	0.8012	0.8031	0.7999	0.8010	0.8062	0.8005	0.8015
70User	0.7038	0.7045	0.6997	0.6948	0.6986	0.7052	0.7038	0.6996	0.6945	0.7042	0.7022	0.6977	0.6992	0.6944	0.7004
60User	0.5980	0.6009	0.6044	0.5986	0.5976	0.5985	0.6030	0.6041	0.6003	0.6038	0.5990	0.6019	0.5982	0.6007	0.6021

	End-to-End F1-Score@B=R					**End-to-End F1-Score@B=2R**					**End-to-End F1-Score@B=3R**				
User 2 / User 1	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	0.6178	0.6178	0.6178	0.6178	0.6178	0.8062	0.8062	0.8062	0.8062	0.8062	0.9110	0.9110	0.9110	0.9110	0.9110
90User	0.5580	0.5506	0.5369	0.5289	0.5258	0.7266	0.6893	0.6739	0.6680	0.6643	0.8225	0.8186	0.8064	0.7996	0.7922
80User	0.4950	0.4793	0.4697	0.4439	0.4332	0.6431	0.6012	0.5684	0.5445	0.5373	0.7282	0.7138	0.6943	0.6726	0.6692
70User	0.4348	0.4136	0.3849	0.3585	0.3301	0.5683	0.5230	0.4755	0.4329	0.4140	0.6392	0.6192	0.5874	0.5702	0.5598
60User	0.3716	0.3507	0.3266	0.2832	0.2489	0.4823	0.4446	0.4041	0.3529	0.3356	0.5452	0.5307	0.4994	0.4801	0.4049

Table B.14: **At-Home2 Dataset: End-To-End–Retrieval Metrics obtained using Lock-Step CAL–Type 2 review strategy for budget B={R, 2R, 3R}**

User 2 / User 1	**End-to-End Recall@B=R**					**End-to-End Recall@B=2R**					**End-to-End Recall@B=3R**				
	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	0.4753	0.4753	0.4753	0.4753	0.4753	0.7840	0.7840	0.7840	0.7840	0.7840	0.8541	0.8541	0.8541	0.8541	0.8541
90User	0.4288	0.4223	0.4055	0.3984	0.3949	0.7065	0.6478	0.6258	0.6171	0.6108	0.7707	0.7632	0.7426	0.7340	0.7235
80User	0.3805	0.3629	0.3533	0.3265	0.3155	0.6252	0.5583	0.5087	0.4788	0.4686	0.6823	0.6572	0.6226	0.5921	0.5865
70User	0.3345	0.3112	0.2823	0.2568	0.2298	0.5525	0.4831	0.4181	0.3650	0.3404	0.5989	0.5684	0.5171	0.4939	0.4760
60User	0.2866	0.2632	0.2379	0.1972	0.1672	0.4688	0.4088	0.3524	0.2901	0.2697	0.5108	0.4846	0.4376	0.4082	0.3115

User 2 / User 1	**End-to-End Precision@B=R**					**End-to-End Precision@B=2R**					**End-to-End Precision@B=3R**				
	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
90User	0.9070	0.8990	0.9083	0.9004	0.9015	0.9031	0.9030	0.9003	0.9004	0.9022	0.9049	0.9062	0.9064	0.9023	0.8997
80User	0.8042	0.8059	0.8030	0.8020	0.8035	0.7995	0.8034	0.8104	0.8027	0.8046	0.8014	0.8025	0.8077	0.8020	0.8030
70User	0.7053	0.7060	0.7012	0.6963	0.7001	0.7067	0.7053	0.7011	0.6960	0.7057	0.7037	0.6992	0.7007	0.6959	0.7019
60User	0.5995	0.6024	0.6059	0.6001	0.5991	0.6000	0.6045	0.6056	0.6018	0.6053	0.6005	0.6034	0.5997	0.6022	0.6036

User 2 / User 1	**End-to-End F1-Score@B=R**					**End-to-End F1-Score@B=2R**					**End-to-End F1-Score@B=3R**				
	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	0.6443	0.6443	0.6443	0.6443	0.6443	0.8789	0.8789	0.8789	0.8789	0.8789	0.9213	0.9213	0.9213	0.9213	0.9213
90User	0.5823	0.5746	0.5607	0.5524	0.5492	0.7928	0.7544	0.7383	0.7323	0.7285	0.8324	0.8286	0.8164	0.8095	0.8020
80User	0.5166	0.5005	0.4907	0.4641	0.4531	0.7017	0.6588	0.6250	0.5998	0.5923	0.7371	0.7226	0.7032	0.6813	0.6779
70User	0.4538	0.4320	0.4025	0.3752	0.3460	0.6201	0.5734	0.5238	0.4789	0.4592	0.6471	0.6270	0.5951	0.5778	0.5673
60User	0.3878	0.3664	0.3417	0.2968	0.2614	0.5264	0.4878	0.4456	0.3915	0.3732	0.5520	0.5375	0.5060	0.4866	0.4109

Table B.15: **At-Home3 Dataset: End-To-End–Retrieval Metrics obtained using Lock-Step CAL–Type 2 review strategy for budget B={R, 2R, 3R}**

	End-to-End Recall@B=R					**End-to-End Recall@B=2R**					**End-to-End Recall@B=3R**				
User 1 \ User 2	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	0.4538	0.4538	0.4538	0.4538	0.4538	0.6900	0.6900	0.6900	0.6900	0.6900	0.8469	0.8469	0.8469	0.8469	0.8469
90User	0.4094	0.4032	0.3871	0.3804	0.3770	0.6218	0.5702	0.5507	0.5431	0.5376	0.7642	0.7568	0.7363	0.7278	0.7174
80User	0.3633	0.3465	0.3373	0.3117	0.3013	0.5503	0.4913	0.4477	0.4214	0.4125	0.6766	0.6517	0.6173	0.5872	0.5816
70User	0.3193	0.2971	0.2695	0.2452	0.2194	0.4862	0.4251	0.3679	0.3213	0.2996	0.5938	0.5636	0.5128	0.4897	0.4720
60User	0.2736	0.2513	0.2272	0.1883	0.1596	0.4126	0.3598	0.3102	0.2553	0.2374	0.5064	0.4805	0.4339	0.4047	0.3088

	End-to-End Precision@B=R					**End-to-End Precision@B=2R**					**End-to-End Precision@B=3R**				
User 1 \ User 2	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
90User	0.9062	0.8982	0.9075	0.8996	0.9007	0.9023	0.9022	0.8995	0.8996	0.9014	0.9041	0.9054	0.9056	0.9015	0.8989
80User	0.8034	0.8051	0.8022	0.8012	0.8027	0.7987	0.8026	0.8096	0.8019	0.8038	0.8006	0.8017	0.8069	0.8012	0.8022
70User	0.7045	0.7052	0.7004	0.6955	0.6993	0.7059	0.7045	0.7003	0.6952	0.7049	0.7029	0.6984	0.6999	0.6951	0.7011
60User	0.5987	0.6016	0.6051	0.5993	0.5983	0.5992	0.6037	0.6048	0.6010	0.6045	0.5997	0.6026	0.5989	0.6014	0.6028

	End-to-End F1-Score@B=R					**End-to-End F1-Score@B=2R**					**End-to-End F1-Score@B=3R**				
User 1 \ User 2	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User	IdealUser	90User	80User	70User	60User
IdealUser	0.6243	0.6243	0.6243	0.6243	0.6243	0.8166	0.8166	0.8166	0.8166	0.8166	0.9171	0.9171	0.9171	0.9171	0.9171
90User	0.5640	0.5565	0.5427	0.5347	0.5315	0.7362	0.6987	0.6832	0.6773	0.6735	0.8283	0.8244	0.8122	0.8054	0.7980
80User	0.5003	0.4845	0.4749	0.4488	0.4381	0.6516	0.6095	0.5766	0.5525	0.5452	0.7334	0.7189	0.6995	0.6777	0.6743
70User	0.4395	0.4181	0.3892	0.3626	0.3340	0.5758	0.5303	0.4824	0.4394	0.4204	0.6438	0.6238	0.5919	0.5746	0.5642
60User	0.3756	0.3545	0.3303	0.2865	0.2520	0.4887	0.4509	0.4101	0.3584	0.3409	0.5491	0.5347	0.5032	0.4838	0.4084

Table B.16: **At-Home4 Dataset: End-To-End–Retrieval Metrics obtained using Lock-Step CAL–Type 2 review strategy for budget B={R, 2R, 3R}**

B.6 T-Test Results

Budget B = 3R	**Review Strategy**						
Topics	**Single User CAL**	**Separate CAL**	**Lock-Step CAL–Type 1**	**Lock-Step CAL–Type 2**	**Majority Vote of Three**	**CAL QC–Type 1**	**CAL QC–Type 2**
athome100	0.4560	0.3556	0.3781	0.3616	0.4719	0.5052	0.5735
athome101	0.4382	0.3332	0.3557	0.3392	0.4495	0.4828	0.5511
athome102	0.3776	0.2759	0.2983	0.2819	0.3922	0.4255	0.4938
athome103	0.4885	0.3836	0.4061	0.3896	0.4999	0.5332	0.6015
athome104	0.3770	0.2541	0.2765	0.2601	0.3704	0.4037	0.4720
athome105	0.4137	0.3106	0.3331	0.3166	0.4269	0.4602	0.5285
athome106	0.4321	0.3304	0.3529	0.3364	0.4467	0.4800	0.5483
athome107	0.3501	0.2474	0.2699	0.2534	0.3637	0.3970	0.4653
athome108	0.4397	0.3455	0.3679	0.3514	0.4617	0.4951	0.5634
athome109	0.3632	0.2673	0.2898	0.2733	0.3836	0.4169	0.4852
athome2052	0.4250	0.3393	0.3608	0.3900	0.4441	0.4826	0.5424
athome2108	0.3605	0.2538	0.2752	0.3044	0.3585	0.3970	0.4569
athome2129	0.3980	0.2998	0.3212	0.3504	0.4045	0.4430	0.5029
athome2130	0.4395	0.3457	0.3671	0.3963	0.4504	0.4889	0.5488
athome2134	0.4688	0.3725	0.3940	0.4232	0.4773	0.5158	0.5756
athome2158	0.2281	0.1149	0.1363	0.1655	0.2196	0.2581	0.3180
athome2225	0.3625	0.2515	0.2729	0.3021	0.3562	0.3947	0.4546
athome2322	0.4757	0.3814	0.4029	0.4321	0.4862	0.5247	0.5845
athome2333	0.3971	0.2991	0.3206	0.3498	0.4039	0.4424	0.5023
athome2461	0.4186	0.3366	0.3580	0.3872	0.4413	0.4798	0.5397
athome3089	0.3734	0.2809	0.3036	0.2874	0.4478	0.4104	0.4904
athome3133	0.4525	0.3691	0.3917	0.3756	0.5359	0.4986	0.5786
athome3226	0.5077	0.4221	0.4447	0.4285	0.5889	0.5515	0.6316
athome3290	0.2451	0.1545	0.1772	0.1610	0.3214	0.2840	0.3640
athome3357	0.3329	0.2367	0.2594	0.2432	0.4036	0.3662	0.4463
athome3378	0.4304	0.3426	0.3653	0.3491	0.5095	0.4721	0.5521
athome3423	0.4376	0.3507	0.3734	0.3572	0.5176	0.4802	0.5602
athome3431	0.4082	0.3216	0.3442	0.3280	0.4884	0.4511	0.5311
athome3481	0.3178	0.2255	0.2482	0.2320	0.3924	0.3550	0.4350
athome3484	0.4505	0.3474	0.3700	0.3539	0.5142	0.4769	0.5569

Table B.17: **T-Test Table for 60User End-to-End–Retrieval: Topic-wise recall percentage for At-Home1, At-Home2, and At-Home3 dataset**

Budget B = 3R	Review Strategy						
Topics	Single User CAL	Separate CAL	Lock-Step CAL–Type 1	Lock-Step CAL–Type 2	Majority Vote of Three	CAL QC–Type 1	CAL QC–Type 2
athome401	0.3903	0.3028	0.3253	0.3091	0.4140	0.4413	0.5116
athome402	0.4318	0.3467	0.3692	0.3530	0.4579	0.4852	0.5555
athome403	0.4611	0.3756	0.3980	0.3819	0.4867	0.5140	0.5844
athome404	0.2204	0.1279	0.1504	0.1342	0.2391	0.2664	0.3367
athome405	0.4645	0.3778	0.4002	0.3841	0.4889	0.5162	0.5866
athome406	0.3531	0.2622	0.2847	0.2685	0.3734	0.4007	0.4710
athome407	0.4466	0.3625	0.3849	0.3688	0.4736	0.5009	0.5713
athome408	0.5017	0.4194	0.4419	0.4258	0.5306	0.5579	0.6282
athome409	0.4971	0.4121	0.4345	0.4184	0.5233	0.5506	0.6209
athome410	0.3269	0.2341	0.2566	0.2404	0.3453	0.3726	0.4429
athome411	0.0562	0.0204	0.0225	0.0225	0.0449	0.0787	0.1236
athome412	0.4203	0.3334	0.3558	0.3397	0.4445	0.4718	0.5422
athome413	0.4266	0.3452	0.3676	0.3515	0.4563	0.4836	0.5540
athome414	0.3903	0.3033	0.3258	0.3096	0.4145	0.4418	0.5121
athome415	0.4318	0.3487	0.3712	0.3550	0.4599	0.4872	0.5575
athome416	0.4022	0.3190	0.3414	0.3253	0.4301	0.4574	0.5278
athome417	0.3118	0.2229	0.2453	0.2292	0.3340	0.3613	0.4317
athome418	0.4445	0.3448	0.3672	0.3511	0.4559	0.4832	0.5536
athome419	0.4244	0.3386	0.3610	0.3449	0.4497	0.4770	0.5474
athome420	0.4387	0.3527	0.3751	0.3590	0.4638	0.4911	0.5615
athome421	0.0562	0.0000	0.0000	0.0000	0.0449	0.0952	0.1236
athome422	0.4655	0.3802	0.4026	0.3865	0.4913	0.5186	0.5890
athome423	0.3891	0.3019	0.3244	0.3082	0.4131	0.4404	0.5107
athome424	0.4669	0.3830	0.4054	0.3893	0.4941	0.5214	0.5918
athome425	0.3943	0.3073	0.3298	0.3136	0.4185	0.4458	0.5161
athome426	0.4146	0.3285	0.3510	0.3348	0.4397	0.4670	0.5373
athome427	0.3996	0.3118	0.3342	0.3181	0.4229	0.4502	0.5206
athome428	0.3983	0.3112	0.3337	0.3175	0.4224	0.4497	0.5200
athome429	0.3719	0.2844	0.3069	0.2907	0.3956	0.4229	0.4932
athome430	0.4855	0.3999	0.4224	0.4062	0.5111	0.5384	0.6087
athome431	0.4619	0.3758	0.3982	0.3821	0.4869	0.5142	0.5846
athome432	0.4844	0.4018	0.4242	0.4081	0.5129	0.5402	0.6106
athome433	0.4798	0.3953	0.4177	0.4016	0.5064	0.5337	0.6041
athome434	0.3110	0.2197	0.2422	0.2261	0.3309	0.3582	0.4285

Table B.18: **Continued, T-Test Table for 60User End-to-End–Retrieval: Topic-wise recall percentage for At-Home4 dataset**

Topics	Single User CAL	Separate CAL	Lock-Step CAL–Type 1	Lock-Step CAL–Type 2	Majority Vote of Three	CAL QC–Type 1	CAL QC–Type 2
Mean (m)	0.402195313	0.304978219	0.327172653	0.326015479	0.434275439	0.445742243	0.515158119
Sample size (n)	30	30	30	30	30	30	30
Variance (σ)	0.00416412	0.004605142	0.004609333	0.004889011	0.005406026	0.004700338	0.004698688
Pooled Standard Deviation (S_p)		0.066216547	0.066232369	0.067279754	0.069174221	0.066574988	0.06656879
Test Static (t)		5.686194798	4.387001731	4.385319696	-1.796128566	-2.533331784	-6.572195
Degree of Freedom (ν)		58	58	58	58	58	58
P-value		0.00000	0.00005	0.00005	0.07768	0.01402	0.00000

Table B.19: **Continued, T-Test Table for 60User End-to-End–Retrieval: Statistical Significance Calculation at 95% confidence interval for At-Home1, At-Home2, and At-Home3 dataset**

Budget B = 3R	Review Strategy						
Topics	**Single User CAL**	**Separate CAL**	**Lock-Step CAL–Type 1**	**Lock-Step CAL–Type 2**	**Majority Vote of Three**	**CAL QC–Type 1**	**CAL QC–Type 2**
athome100	0.5799	0.5354	0.5724	0.5468	0.6039	0.6930	0.7197
athome101	0.5621	0.5131	0.5500	0.5244	0.5815	0.6706	0.6973
athome102	0.5015	0.4557	0.4927	0.4671	0.5242	0.6133	0.6400
athome103	0.6123	0.5635	0.6005	0.5748	0.6319	0.7211	0.7478
athome104	0.5009	0.4339	0.4709	0.4453	0.5024	0.5915	0.6182
athome105	0.5376	0.4905	0.5275	0.5018	0.5589	0.6481	0.6748
athome106	0.5559	0.5103	0.5473	0.5216	0.5787	0.6678	0.6945
athome107	0.4740	0.4273	0.4643	0.4386	0.4957	0.5849	0.6115
athome108	0.5635	0.5253	0.5623	0.5366	0.5937	0.6829	0.7096
athome109	0.4870	0.4472	0.4841	0.4585	0.5156	0.6047	0.6314
athome2052	0.5442	0.5122	0.5484	0.5237	0.5692	0.6633	0.6811
athome2108	0.4797	0.4266	0.4628	0.4381	0.4837	0.5777	0.5955
athome2129	0.5172	0.4726	0.5089	0.4841	0.5297	0.6237	0.6415
athome2130	0.5587	0.5185	0.5547	0.5300	0.5756	0.6696	0.6874
athome2134	0.5880	0.5454	0.5816	0.5569	0.6024	0.6965	0.7143
athome2158	0.3473	0.2877	0.3239	0.2992	0.3448	0.4388	0.4566
athome2225	0.4818	0.4243	0.4605	0.4358	0.4814	0.5754	0.5932
athome2322	0.5950	0.5543	0.5905	0.5658	0.6113	0.7054	0.7232
athome2333	0.5163	0.4720	0.5082	0.4835	0.5290	0.6231	0.6409
athome2461	0.5378	0.5094	0.5456	0.5209	0.5665	0.6605	0.6783
athome3089	0.4850	0.4587	0.4950	0.4698	0.5926	0.5978	0.6325
athome3133	0.5641	0.5469	0.5832	0.5579	0.6807	0.6860	0.7207
athome3226	0.6192	0.5999	0.6362	0.6109	0.7337	0.7389	0.7736
athome3290	0.3567	0.3323	0.3686	0.3433	0.4661	0.4714	0.5061
athome3357	0.4444	0.4146	0.4509	0.4256	0.5484	0.5536	0.5883
athome3378	0.5419	0.5204	0.5567	0.5314	0.6542	0.6595	0.6942
athome3423	0.5492	0.5285	0.5648	0.5395	0.6623	0.6676	0.7023
athome3431	0.5197	0.4994	0.5357	0.5104	0.6332	0.6385	0.6732
athome3481	0.4293	0.4033	0.4396	0.4143	0.5371	0.5424	0.5771
athome3484	0.5620	0.5252	0.5615	0.5362	0.6590	0.6643	0.6990

Table B.20: **T-Test Table for 70User End-to-End–Retrieval: Topic-wise recall percentage for At-Home1, At-Home2, and At-Home3 dataset**

Budget B = 3R	Review Strategy						
Topics	Single User CAL	Separate CAL	Lock-Step CAL–Type 1	Lock-Step CAL–Type 2	Majority Vote of Three	CAL QC–Type 1	CAL QC–Type 2
athome401	0.5108	0.4782	0.5151	0.4901	0.5420	0.6221	0.6527
athome402	0.5524	0.5220	0.5590	0.5340	0.5859	0.6660	0.6966
athome403	0.5816	0.5509	0.5879	0.5628	0.6148	0.6949	0.7255
athome404	0.3409	0.3033	0.3402	0.3152	0.3671	0.4472	0.4778
athome405	0.5851	0.5531	0.5901	0.5650	0.6170	0.6971	0.7277
athome406	0.4736	0.4376	0.4745	0.4495	0.5014	0.5815	0.6121
athome407	0.5672	0.5378	0.5748	0.5497	0.6017	0.6818	0.7124
athome408	0.6223	0.5948	0.6318	0.6067	0.6586	0.7388	0.7693
athome409	0.6177	0.5874	0.6244	0.5994	0.6513	0.7314	0.7620
athome410	0.4475	0.4095	0.4464	0.4214	0.4733	0.5534	0.5840
athome411	0.1080	0.0680	0.0899	0.0798	0.1046	0.1798	0.2047
athome412	0.5408	0.5087	0.5457	0.5206	0.5726	0.6527	0.6833
athome413	0.5471	0.5205	0.5575	0.5324	0.5844	0.6645	0.6951
athome414	0.5108	0.4787	0.5156	0.4906	0.5425	0.6226	0.6532
athome415	0.5524	0.5240	0.5610	0.5360	0.5879	0.6680	0.6986
athome416	0.5228	0.4943	0.5313	0.5062	0.5582	0.6383	0.6689
athome417	0.4324	0.3982	0.4352	0.4101	0.4621	0.5422	0.5728
athome418	0.5651	0.5201	0.5571	0.5320	0.5840	0.6641	0.6947
athome419	0.5450	0.5139	0.5509	0.5258	0.5777	0.6579	0.6885
athome420	0.5592	0.5280	0.5650	0.5399	0.5919	0.6720	0.7026
athome421	0.1508	0.1076	0.1445	0.1195	0.1714	0.2515	0.2821
athome422	0.5860	0.5555	0.5925	0.5674	0.6194	0.6995	0.7301
athome423	0.5096	0.4773	0.5142	0.4892	0.5411	0.6212	0.6518
athome424	0.5874	0.5583	0.5953	0.5702	0.6222	0.7023	0.7329
athome425	0.5149	0.4827	0.5196	0.4946	0.5465	0.6266	0.6572
athome426	0.5352	0.5039	0.5408	0.5158	0.5677	0.6478	0.6784
athome427	0.5201	0.4871	0.5241	0.4990	0.5510	0.6311	0.6617
athome428	0.5188	0.4866	0.5235	0.4985	0.5504	0.6305	0.6611
athome429	0.4925	0.4598	0.4967	0.4717	0.5236	0.6037	0.6343
athome430	0.6060	0.5752	0.6122	0.5872	0.6391	0.7192	0.7498
athome431	0.5824	0.5511	0.5881	0.5630	0.6150	0.6951	0.7257
athome432	0.6050	0.5771	0.6141	0.5890	0.6410	0.7211	0.7517
athome433	0.6004	0.5706	0.6076	0.5825	0.6345	0.7146	0.7452
athome434	0.4316	0.3951	0.4321	0.4070	0.4589	0.5391	0.5696

Table B.21: **Continued, T-Test Table for 70User End-to-End–Retrieval: Topic-wise recall percentage for At-Home4 dataset**

Topics	**Single User CAL**	**Separate CAL**	**Lock-Step CAL–Type 1**	**Lock-Step CAL–Type 2**	**Majority Vote of Three**	**CAL QC–Type 1**	**CAL QC–Type 2**
Mean (m)	0.520393443	0.481813953	0.518320033	0.493092156	0.568247796	0.631059823	0.657463016
Sample size (n)	30	30	30	30	30	30	30
Variance (σ)	0.004262151	0.004640577	0.004645199	0.004639381	0.005949284	0.00472985	0.004775729
Pooled Standard Deviation (S_p)		0.004451364	0.004453675	0.004450766	0.005105718	0.004496001	0.00451894
Test Static (t)		2.239523156	0.120329389	1.584934693	-2.59381306	-6.392167564	-7.897113886
Degree of Freedom (ν)		58	58	58	58	58	58
P-value		0.02897	0.90464	0.11842	0.01200	0.00000	0.00000

Table B.22: **Continued, T-Test Table for 70User End-to-End–Retrieval: Statistical Significance Calculation at 95% confidence interval for At-Home1, At-Home2, and At-Home3 dataset**

Budget B = 3R	**Review Strategy**						
Topics	**Single User CAL**	**Separate CAL**	**Lock-Step CAL–Type 1**	**Lock-Step CAL–Type 2**	**Majority Vote of Three**	**CAL QC–Type 1**	**CAL QC–Type 2**
athome100	0.7166	0.6563	0.6980	0.6774	0.7097	0.8391	0.8366
athome101	0.6988	0.6340	0.6757	0.6551	0.6874	0.8167	0.8142
athome102	0.6383	0.5766	0.6183	0.5977	0.6300	0.7594	0.7569
athome103	0.7491	0.6844	0.7261	0.7055	0.7378	0.8671	0.8646
athome104	0.6376	0.5548	0.5965	0.5759	0.6082	0.7376	0.7351
athome105	0.6743	0.6114	0.6531	0.6325	0.6648	0.7941	0.7916
athome106	0.6927	0.6312	0.6729	0.6523	0.6846	0.8139	0.8114
athome107	0.6108	0.5482	0.5899	0.5693	0.6016	0.7309	0.7284
athome108	0.7003	0.6462	0.6879	0.6673	0.6996	0.8289	0.8264
athome109	0.6238	0.5681	0.6098	0.5892	0.6215	0.7508	0.7483
athome2052	0.6760	0.6288	0.6695	0.6496	0.6696	0.8038	0.7928
athome2108	0.6115	0.5432	0.5840	0.5640	0.5841	0.7182	0.7072
athome2129	0.6490	0.5892	0.6300	0.6100	0.6301	0.7642	0.7532
athome2130	0.6905	0.6351	0.6758	0.6559	0.6760	0.8101	0.7991
athome2134	0.7198	0.6620	0.7027	0.6828	0.7028	0.8370	0.8260
athome2158	0.4791	0.4043	0.4451	0.4251	0.4452	0.5793	0.5683
athome2225	0.6136	0.5409	0.5817	0.5617	0.5818	0.7159	0.7049
athome2322	0.7268	0.6709	0.7116	0.6917	0.7117	0.8459	0.8349
athome2333	0.6481	0.5886	0.6293	0.6094	0.6294	0.7636	0.7526
athome2461	0.6696	0.6260	0.6667	0.6468	0.6669	0.8010	0.7900
athome3089	0.6145	0.5769	0.6188	0.5984	0.7095	0.7258	0.7468
athome3133	0.6936	0.6650	0.7069	0.6866	0.7976	0.8140	0.8350
athome3226	0.7487	0.7180	0.7599	0.7396	0.8506	0.8669	0.8879
athome3290	0.4862	0.4504	0.4923	0.4720	0.5830	0.5994	0.6204
athome3357	0.5739	0.5327	0.5746	0.5542	0.6653	0.6816	0.7026
athome3378	0.6714	0.6385	0.6805	0.6601	0.7711	0.7875	0.8085
athome3423	0.6787	0.6466	0.6885	0.6682	0.7792	0.7956	0.8166
athome3431	0.6492	0.6175	0.6594	0.6391	0.7501	0.7664	0.7874
athome3481	0.5588	0.5214	0.5633	0.5430	0.6540	0.6704	0.6914
athome3484	0.6915	0.6433	0.6852	0.6649	0.7759	0.7923	0.8133

Table B.23: **T-Test Table for 80User End-to-End–Retrieval: Topic-wise recall percentage for At-Home1, At-Home2, and At-Home3 dataset**

Budget B = 3R	Review Strategy						
Topics	**Single User CAL**	**Separate CAL**	**Lock-Step CAL–Type 1**	**Lock-Step CAL–Type 2**	**Majority Vote of Three**	**CAL QC–Type 1**	**CAL QC–Type 2**
athome401	0.6388	0.5962	0.6378	0.6176	0.6447	0.7615	0.7663
athome402	0.6803	0.6400	0.6817	0.6615	0.6886	0.8054	0.8102
athome403	0.7096	0.6689	0.7105	0.6904	0.7175	0.8342	0.8391
athome404	0.4689	0.4212	0.4629	0.4427	0.4698	0.5866	0.5914
athome405	0.7130	0.6711	0.7127	0.6926	0.7197	0.8364	0.8413
athome406	0.6016	0.5555	0.5972	0.5770	0.6041	0.7209	0.7257
athome407	0.6951	0.6558	0.6974	0.6773	0.7044	0.8211	0.8260
athome408	0.7503	0.7128	0.7544	0.7343	0.7613	0.8781	0.8829
athome409	0.7456	0.7054	0.7471	0.7269	0.7540	0.8708	0.8756
athome410	0.5754	0.5275	0.5691	0.5490	0.5760	0.6928	0.6976
athome411	0.1599	0.1157	0.1573	0.1371	0.1642	0.2810	0.2858
athome412	0.6688	0.6267	0.6683	0.6482	0.6753	0.7920	0.7969
athome413	0.6751	0.6385	0.6801	0.6600	0.6871	0.8038	0.8087
athome414	0.6388	0.5967	0.6383	0.6181	0.6452	0.7620	0.7668
athome415	0.6803	0.6420	0.6837	0.6635	0.6906	0.8074	0.8122
athome416	0.6507	0.6123	0.6539	0.6338	0.6609	0.7776	0.7824
athome417	0.5603	0.5162	0.5579	0.5377	0.5648	0.6815	0.6864
athome418	0.6930	0.6381	0.6797	0.6596	0.6867	0.8034	0.8083
athome419	0.6729	0.6319	0.6735	0.6534	0.6804	0.7972	0.8020
athome420	0.6872	0.6460	0.6876	0.6675	0.6946	0.8113	0.8161
athome421	0.2788	0.2255	0.2672	0.2470	0.2741	0.3909	0.3957
athome422	0.7140	0.6735	0.7152	0.6950	0.7221	0.8388	0.8437
athome423	0.6376	0.5952	0.6369	0.6167	0.6438	0.7606	0.7654
athome424	0.7154	0.6763	0.7179	0.6978	0.7249	0.8416	0.8465
athome425	0.6429	0.6006	0.6423	0.6221	0.6492	0.7660	0.7708
athome426	0.6632	0.6219	0.6635	0.6433	0.6704	0.7872	0.7920
athome427	0.6481	0.6051	0.6467	0.6266	0.6537	0.7704	0.7752
athome428	0.6468	0.6045	0.6462	0.6260	0.6531	0.7699	0.7747
athome429	0.6204	0.5778	0.6194	0.5993	0.6263	0.7431	0.7479
athome430	0.7340	0.6932	0.7349	0.7147	0.7418	0.8586	0.8634
athome431	0.7104	0.6691	0.7107	0.6906	0.7177	0.8344	0.8393
athome432	0.7329	0.6951	0.7367	0.7166	0.7437	0.8604	0.8653
athome433	0.7284	0.6886	0.7302	0.7101	0.7372	0.8539	0.8587
athome434	0.5595	0.5131	0.5547	0.5346	0.5616	0.6784	0.6832

Table B.24: **Continued, T-Test Table for 80User End-to-End–Retrieval: Topic-wise recall percentage for At-Home4 dataset**

Topics	Single User CAL	Separate CAL	Lock-Step CAL–Type 1	Lock-Step CAL–Type 2	Majority Vote of Three	CAL QC–Type 1	CAL QC–Type 2
Mean (m)	0.653089798	0.600362606	0.641802308	0.62151312	0.675967522	0.769239822	0.771740765
Sample size (n)	30	30	30	30	30	30	30
Variance (σ)	0.004351145	0.00466987	0.004677998	0.004672701	0.006513835	0.004931164	0.004839942
Pooled Standard Deviation (S_p)		0.004510508	0.004514572	0.004511923	0.00543249	0.004641155	0.004595544
Test Static (t)		3.040657809	0.650630993	1.820669958	-1.202150873	-6.603161288	-6.778731931
Degree of Freedom (ν)		58	58	58	58	58	58
P-value		0.00354	0.51785	0.07382	0.23419	0.00000	0.00000

Table B.25: **Continued, T-Test Table for 80User End-to-End–Retrieval: Statistical Significance Calculation at 95% confidence interval for At-Home1, At-Home2, and At-Home3 dataset**

Budget B = 3R	Review Strategy						
Topics	**Single User CAL**	**Separate CAL**	**Lock-Step CAL–Type 1**	**Lock-Step CAL–Type 2**	**Majority Vote of Three**	**CAL QC–Type 1**	**CAL QC–Type 2**
athome100	0.8720	0.8053	0.8403	0.8202	0.7510	0.9018	0.9010
athome101	0.8542	0.7829	0.8179	0.7978	0.7286	0.8794	0.8786
athome102	0.7936	0.7256	0.7606	0.7405	0.6713	0.8221	0.8213
athome103	0.9045	0.8334	0.8684	0.8483	0.7791	0.9299	0.9291
athome104	0.7930	0.7038	0.7388	0.7187	0.6495	0.8003	0.7995
athome105	0.8297	0.7604	0.7954	0.7753	0.7061	0.8569	0.8561
athome106	0.8481	0.7802	0.8152	0.7951	0.7259	0.8767	0.8759
athome107	0.7661	0.6972	0.7322	0.7121	0.6429	0.7937	0.7929
athome108	0.8556	0.7952	0.8302	0.8101	0.7409	0.8917	0.8909
athome109	0.7791	0.7170	0.7520	0.7319	0.6627	0.8135	0.8127
athome2052	0.8255	0.7724	0.8068	0.7875	0.7088	0.8642	0.8540
athome2108	0.7610	0.6869	0.7212	0.7019	0.6233	0.7787	0.7684
athome2129	0.7985	0.7329	0.7672	0.7479	0.6693	0.8247	0.8144
athome2130	0.8401	0.7788	0.8131	0.7938	0.7152	0.8706	0.8603
athome2134	0.8693	0.8056	0.8400	0.8207	0.7420	0.8974	0.8872
athome2158	0.6286	0.5480	0.5823	0.5630	0.4844	0.6398	0.6295
athome2225	0.7631	0.6846	0.7189	0.6996	0.6210	0.7764	0.7661
athome2322	0.8763	0.8145	0.8489	0.8296	0.7509	0.9063	0.8961
athome2333	0.7976	0.7322	0.7666	0.7473	0.6686	0.8240	0.8138
athome2461	0.8191	0.7697	0.8040	0.7847	0.7061	0.8615	0.8512
athome3089	0.7612	0.7236	0.7596	0.7391	0.7550	0.7851	0.8095
athome3133	0.8403	0.8117	0.8478	0.8273	0.8432	0.8732	0.8977
athome3226	0.8955	0.8647	0.9008	0.8802	0.8962	0.9262	0.9507
athome3290	0.6329	0.5971	0.6332	0.6127	0.6286	0.6587	0.6831
athome3357	0.7207	0.6794	0.7154	0.6949	0.7108	0.7409	0.7654
athome3378	0.8182	0.7852	0.8213	0.8008	0.8167	0.8468	0.8712
athome3423	0.8254	0.7933	0.8294	0.8089	0.8248	0.8549	0.8793
athome3431	0.7960	0.7642	0.8003	0.7797	0.7957	0.8257	0.8502
athome3481	0.7056	0.6681	0.7042	0.6837	0.6996	0.7297	0.7541
athome3484	0.8383	0.7900	0.8261	0.8056	0.8215	0.8515	0.8760

Table B.26: **T-Test Table for 90User End-to-End–Retrieval: Topic-wise recall percentage for At-Home1, At-Home2, and At-Home3 dataset**

Budget B = 3R	**Review Strategy**						
Topics	**Single User CAL**	**Separate CAL**	**Lock-Step CAL–Type 1**	**Lock-Step CAL–Type 2**	**Majority Vote of Three**	**CAL QC–Type 1**	**CAL QC–Type 2**
athome401	0.7945	0.7416	0.7767	0.7572	0.6848	0.8217	0.8287
athome402	0.8361	0.7854	0.8206	0.8011	0.7287	0.8656	0.8726
athome403	0.8653	0.8143	0.8495	0.8299	0.7575	0.8944	0.9014
athome404	0.6246	0.5667	0.6018	0.5823	0.5099	0.6468	0.6538
athome405	0.8688	0.8165	0.8517	0.8322	0.7598	0.8966	0.9037
athome406	0.7573	0.7009	0.7361	0.7166	0.6442	0.7811	0.7881
athome407	0.8509	0.8012	0.8364	0.8169	0.7444	0.8813	0.8884
athome408	0.9060	0.8582	0.8934	0.8738	0.8014	0.9383	0.9453
athome409	0.9014	0.8508	0.8860	0.8665	0.7941	0.9310	0.9380
athome410	0.7312	0.6729	0.7080	0.6885	0.6161	0.7530	0.7600
athome411	0.3156	0.2611	0.2962	0.2767	0.2043	0.3412	0.3482
athome412	0.8245	0.7721	0.8073	0.7877	0.7153	0.8522	0.8592
athome413	0.8308	0.7839	0.8191	0.7996	0.7272	0.8640	0.8711
athome414	0.7945	0.7421	0.7772	0.7577	0.6853	0.8222	0.8292
athome415	0.8361	0.7874	0.8226	0.8031	0.7307	0.8676	0.8746
athome416	0.8065	0.7577	0.7929	0.7733	0.7009	0.8378	0.8448
athome417	0.7161	0.6616	0.6968	0.6773	0.6049	0.7418	0.7488
athome418	0.8488	0.7835	0.8187	0.7992	0.7268	0.8636	0.8707
athome419	0.8287	0.7773	0.8125	0.7929	0.7205	0.8574	0.8644
athome420	0.8429	0.7914	0.8266	0.8070	0.7346	0.8715	0.8785
athome421	0.4346	0.3710	0.4061	0.3866	0.3142	0.4511	0.4581
athome422	0.8698	0.8189	0.8541	0.8346	0.7622	0.8991	0.9061
athome423	0.7934	0.7407	0.7758	0.7563	0.6839	0.8208	0.8278
athome424	0.8711	0.8217	0.8569	0.8374	0.7650	0.9018	0.9089
athome425	0.7986	0.7461	0.7812	0.7617	0.6893	0.8262	0.8332
athome426	0.8189	0.7673	0.8024	0.7829	0.7105	0.8474	0.8544
athome427	0.8039	0.7505	0.7857	0.7661	0.6937	0.8306	0.8376
athome428	0.8025	0.7500	0.7851	0.7656	0.6932	0.8301	0.8371
athome429	0.7762	0.7232	0.7583	0.7388	0.6664	0.8033	0.8103
athome430	0.8898	0.8386	0.8738	0.8543	0.7819	0.9188	0.9258
athome431	0.8661	0.8145	0.8497	0.8301	0.7577	0.8946	0.9016
athome432	0.8887	0.8405	0.8757	0.8562	0.7837	0.9206	0.9277
athome433	0.8841	0.8340	0.8692	0.8496	0.7772	0.9141	0.9211
athome434	0.7153	0.6585	0.6937	0.6741	0.6017	0.7386	0.7456

Table B.27: **Continued, T-Test Table for 90User End-to-End–Retrieval: Topic-wise recall percentage for At-Home4 dataset**

Topics	Single User CAL	Separate CAL	Lock-Step CAL–Type 1	Lock-Step CAL–Type 2	Majority Vote of Three	CAL QC–Type 1	CAL QC–Type 2
Mean (m)	0.803636209	0.746801506	0.781945289	0.761959739	0.717995063	0.830075679	0.834547652
Sample size (n)	30	30	30	30	30	30	30
Variance (σ)	0.004480841	0.004715875	0.004723322	0.004715059	0.006757493	0.004987876	0.004884579
Pooled Standard Deviation (S_p)		0.067811195	0.067838644	0.067808185	0.074961104	0.068806674	0.06843033
Test Static (t)		3.246069631	1.238358666	2.380424646	4.424784543	-1.488222301	-1.749509332
Degree of Freedom (ν)		58	58	58	58	58	58
P-value		0.00195	0.22057	0.02060	0.00004	0.14211	0.08549

Table B.28: **Continued, T-Test Table for 90User End-to-End–Retrieval: Statistical Significance Calculation at 95% confidence interval for At-Home1, At-Home2, and At-Home3 dataset**