Active learning pipeline to automatically identify candidate terms for a CDSS ontology—measures, experiments, and performance

Shailesh Alluri, MS1, Keerthana Komatineni*, MS1, Rohan Goli*, MS1, Nina Hubig, PhD1, Hua Min, PhD2, Yang Gong, PhD3, Dean F. Sittig, PhD3, David Robinson, MD4, Paul Biondich, MD5, Adam Wright, PhD6, Christian Nøhr, PhD7, Timothy Law, DO8, Arild Faxvaag, PhD9, Richard D. Boyce, PhD10, Ronald Gimbel, PhD1, Lior Rennert, PhD1, Xia Jing, PhD1

1Clemson University, Clemson, SC, USA; 2George Mason University, Fairfax, VA, USA; 3University of Texas Health Sciences Center at Houston, Houston, TX, USA; 4Independent Consultant, Cumbria, UK;5 Indiana University School of Medicine, Indianapolis, IN, USA; 6Vanderbilt University Medical Center, Nashville, TN, USA; 7Aalborg University, Aalborg, Denmark; 8Ohio Musculoskeletal and Neurologic Institute, Ohio University, Athens, OH, USA; 9Norwegian University of Science and Technology, Trondheim, Norway; 10University of Pittsburgh, Pittsburgh, PA, USA *Coauthors contributed equally

ABSTRACT:

In this paper, we present an active learning framework designed to enhance the automatic identification of key phrases relevant to a Clinical Decision Support System (CDSS) ontology. CDSS ontology can play a critical role in standardizing medical vocabulary and enabling seamless data integration across healthcare systems. Traditional methods for ontology development are manual, labor-intensive, and require significant domain expertise. Our approach combines a Named Entity Recognition (NER) component based on a BiLSTM-CRF model with an active learning loop to progressively incorporate human experts' feedback to improve its accuracy. We implement uncertainty sampling as our core data selection strategy, prioritizing instances where the model exhibits low confidence for human review. We introduce new uncertainty aggregation methods—KPSum, KPAvg, DOCSum, and DOCAvg—which, in combination with uncertainty measures such as Maximum Token Probability (MTP), Token Entropy (TE), and Margin, are used to calculate document-level confidence scores. These methods improve the selection process in Active Learning, ensuring that the most informative documents are prioritized for annotation. This study underscores the value of active learning in facilitating ontology development.

INTRODUCTION:

Clinical Decision Support Systems (CDSS) have become a cornerstone in modern healthcare, aiding medical practitioners by offering data-driven recommendations to improve patient care[Ref 1]. Ontology is an enabling technology of the Semantic Web technologies and plays a critical role in information interoperability [Ref 2]. A medical ontology, defines the domain concepts and their relationships, can be used in CDSS in multiple ways, such as to develop

reusable CDSS rules. The ontologies are crucial in standardizing clinical vocabulary and ensuring interoperability across different healthcare systems [Ref 3]. However, building and maintaining such an ontology typically is a manual effort and can be a challenging task, as it is labor intensive and it requires domain expertise. In the long term, the ontology requires continuous updates to include newly discovered medical terms, procedures, and treatments. As medical knowledge rapidly evolves, the task of maintaining ontologies manually becomes increasingly difficult, even unrealistic, leading to gaps in outdated information.

This paper addresses these limitations by proposing an active learning-based pipeline that automates the identification of key phrases from unstructured CDSS publications. Our approach leverages deep learning to systematically select and highlight important terms, which are then refined through iterative human feedback. This reduces the reliance on solely manual selection while ensuring the ontology remains accurate and current, which is particularly important in long term CDSS ontology maintenance. By incorporating feedback from human domain experts (HDEs), the model continuously improves its performance, refining its predictions of context-specific phrases and evolving, as the new data becomes available. Our pipeline emphasizes prioritizing instances where the model is uncertain or where consensus is mostly needed, enabling targeted human review to resolve ambiguities effectively.

The key contributions of this work are as follows:

1. **Advanced Aggregation Methods for calculating Confidence Scoring**:
   We propose new document-level confidence aggregation methods—KPSum, KPAvg, DOCSum, and DOCAvg—that prioritize data for expert review more effectively. These methods leverage established uncertainty sampling strategies, such as Maximum Token Probability (MTP), Margin, and Token Entropy[Ref 4], to improve the selection process in Active Learning, ensuring that the most informative documents are chosen for annotation.
2. **Seamless Integration of an Active Learning Pipeline**:
   This work introduces a novel active learning framework for key phrase identification with a continuous expert feedback loop. This approach reduces the effort required for pure manual annotations while iteratively refining the model to adapt to evolving data distributions over time.
3. **Comparison of various Strategies:**
   We conduct a systematic evaluation of multiple Active Learning strategies(e.g. MTP DocAvg Actual, MTP DocAvg Reverse, MTP KPSum Actual, TE KPSum Actual)**.** We also compare these strategies against a Random selection, which was used as the baseline, to assess the advantages of uncertainty-based selection. By analyzing their impact on F1 and recall over successive Active Learning cycles, we provide evidence and insights into the effectiveness of each strategy in improving model performance and annotation efficiency.
4. **Improvement in key phrase Identification**:
   Through iterative Active Learning cycles, our approach achieves faster and more accurate identification of key phrases, reducing the dependence on manual annotation processes.

The remainder of this paper is structured as follows**:** In Section 2, we review related work on key phrase identification and active learning approaches. Section 3(Methodology) describes the integration of human feedback and continuous learning mechanisms. It also presents our proposed methodology, including a brief introduction of the deep learning model architecture with an active learning loop, and our novel confidence score calculation methods using our proposed uncertainty aggregation methods (KPSum, KPAvg, DOCSum, and DOCAvg). Section 4 (Experiments) outlines the experimental setup and presents our results, while Section 5 interprets the primary experiment results with brief discussion and analysis. Section 6 (Conclusion) concludes with the implications of our findings and future work.

Section 2: Literature

Identifying key phrases from textual data is a critical task in natural language processing (NLP), forming the foundation for downstream applications such as ontology development, information retrieval, and text summarization. Traditional methods, including rule-based and statistical approaches like Term Frequency-Inverse Document Frequency and Rapid Automatic Keyword Extraction(Rose et al., 2010) [Ref 5], have been widely adopted for key phrase extraction. However, these methods often fail to generalize across domains, prompting the adoption of machine learning-based models, such as conditional random fields (CRFs) (Lafferty et al., 2001) [Ref 6] and sequence labeling models, for named entity recognition (NER) and key phrase identification.

The advent of deep learning has revolutionized the NER landscape. Models such as

Bidirectional Long Short-Term Memory Network - Conditional Random Fields (Huang et al., 2015)[Ref 7] and transformer-based architectures like Bidirectional Encoder Representations from Transformers(BERT) (Devlin et al., 2019)[Ref 8] have set new benchmarks in NER tasks. Notably, the work by Alzaidy et al. utilized [Ref 9] a BiLSTM-CRF model for key phrase identification, demonstrating its ability to effectively leverage hierarchical context and transfer learning techniques for domain-specific applications. Our work builds on this foundation, extending it with an active learning framework to reduce labeling costs while maintaining high accuracy in key phrase identification.

Active learning (AL) is a subfield of machine learning that aims to achieve high model performance with minimal labeled data by strategically selecting the most informative samples for annotation[Ref 10]. Active learning has proven to be an effective strategy to enhance performance in tasks with limited labeled data, such as NER. Settles (2009) [Ref 11] provides a comprehensive overview of active learning strategies, highlighting uncertainty sampling, query-by-committee, and expected error reduction as key techniques. Shen et al. (2017) [Ref 12] employed uncertainty sampling with deep learning models, effectively reducing the labeling effort required for text classification. Similarly, Siddhant and Lipton (2018) [Ref 13] explored active learning in low-resource NER tasks, demonstrating the benefits of selectively annotating high-value samples.

Integrating active learning into key phrase identification has gained attention in recent years. For instance, Radmard et al. (2021) [Ref 14] utilized active learning to iteratively refine Named Entity Recognition (NER) models, focusing on identifying ambiguous or novel terms in unstructured text. Margatina et al. (2021) [Ref 15] extended active learning techniques to transformer models like BERT, achieving significant performance gains with reduced annotation costs. These studies underscore the potential of active learning in improving the efficiency and accuracy of NER systems. In contrast to prior work, our study begins with a synthetically labeled training set, requiring the model to adapt to real-world patterns over time. Moreover, our use case necessitated document-level uncertainty estimation, for which we introduced novel aggregation strategies tailored to this setup

Our work contributes to this growing body of research by introducing an active learning framework specifically designed to enhance the accuracy of key phrase identification in the clinical decision support domain. By leveraging uncertainty sampling and human-in-the-loop feedback mechanisms, our approach enables the iterative refinement of our BiLSTM-CRF, key phrase identification model, achieving more accurate and scalable key phrase identification pipelines. This framework ultimately supports the broader goal of ontology development by providing high-quality, domain-specific key phrases for HDE to review.

Section 3: METHODOLOGY:

3.1 Active Learning Overview

Active learning (AL) is a subfield of machine learning that aims to achieve high model performance with minimal labeled data by strategically selecting the most informative samples for annotation [Ref 16]. In a typical supervised learning setting, a large amount of labeled data is required to train a model effectively. However, in many real-world applications, such as NER or key phrase identification, obtaining labeled data is expensive and time-consuming due to the need for domain expertise.

Active learning addresses this challenge by iteratively selecting the data points that the model finds most uncertain or informative and presenting them for human annotation. This process reduces the overall annotation effort while maximizing the model's learning efficiency.

3.2 Active Learning Cycle:

An Active Learning Cycle refers to the learning process between two stages, e.g., Mx to Mx+1, where Mx represents the model state after x cycles of Active Learning. Each cycle involves selecting the most informative samples from the data pool, obtaining annotations, and retraining the model with the newly labeled data (Figure 1).

This iterative approach ensures that the model continuously improves by focusing on the most uncertain or informative samples in each cycle. The model will obtain incrementally improving performance through each cycle.

The active learning cycle operates as follows:

At each stage Mx, a BiLSTM-CRF model is used to classify tokens into five possible tags: B-KP (Beginning of a Key phrase), I-KP (Inside a key phrase), O-KP (Outside a key phrase), Start (Start of the sequence), and Stop (End of the sequence). The model assigns probabilities to each token, indicating the likelihood of belonging to one of these categories. These probabilities serve as the foundation for uncertainty measures, which are used to compute confidence scores for each document in the Active Learning data pool.

Once the confidence scores are calculated (sections 3.3.1 and 3.3.2 will introduce the detailed methods about calculations), the documents in the Active Learning data pool are sorted based on these scores. The sorting ensures that the most informative documents—those that the model is most uncertain about—are prioritized for HDE annotation. The order of selection (explained in detail later in section 4.1.1) determines how these documents are chosen. A subset of documents is then selected and sent to HDEs for annotation.

After annotation, the labeled documents are added to the training data and removed from the Active Learning pool to prevent multiple copies of the same document being labeled again. This step is essential to avoid overfitting, as repeatedly training on the same document could reinforce biases rather than improving the model's generalization ability.

Once the newly labeled documents are incorporated, the BiLSTM-CRF model is retrained using the updated dataset. This newly trained model, denoted as Mx+1, then undergoes the same Active Learning selection process in the next cycle. This iterative process continues until a stopping criterion (such as performance saturation) is met.

By systematically incorporating human-labeled data and prioritizing informative samples, this Active Learning cycle enhances the model’s ability to classify key phrases more effectively with each iteration. In this work, we employ active learning to identify key phrases relevant to Clinical Decision Support System (CDSS) ontologies.

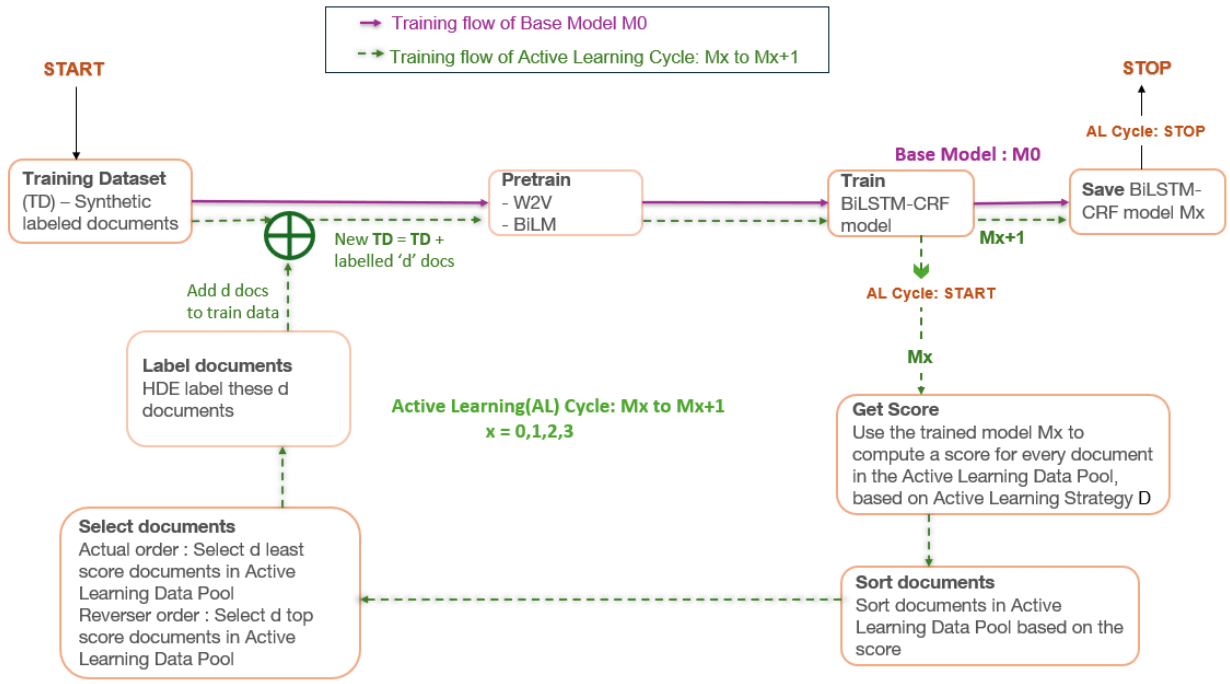

Figure 1: Active Learning Cycle for automatic key phrase identification

3.3 Active Learning Framework

Our framework begins with a BiLSTM-CRF model for Named Entity Recognition (NER), which serves as the foundational model (denoted as M0) for key phrase identification. Initially, M0 is trained on a small set of synthetically labeled medical documents(500 documents)[Ref 9]. Following this, the framework employs a pool-based Active Learning (AL) strategy, as outlined by Settles in the survey [Ref 11], to iteratively enhance the model by selecting and labeling the most informative documents from a larger unlabeled pool, Active Learning data pool(42 Documents). This process allows the model to progressively improve its performance by leveraging uncertainty sampling techniques, focusing on the most uncertain data points to maximize learning efficiency.

Figure 2 illustrates the entire lifecycle of our Active Learning experiment, showing the iterative model learning and improvement process from M0 to M4. Each transition represents an Active Learning cycle, where new human-reviewed documents are incorporated into the training data.

- M0: The initial model (M0) is trained solely on 500 synthetic documents.
- M1: After the first Active Learning cycle, the model is retrained on 500 synthetic documents + 5 selected documents reviewed by HDE.
- M2: The training data expands further, now including 500 synthetic documents + 10 human-reviewed documents (Previous 5 from M1 + 5 new documents).

- M3: The model is retrained on 500 synthetic documents + 15 human-reviewed documents (previous 10 from M1 and M2 + 5 new documents).
- M4: The final cycle includes 500 synthetic documents + 20 human-reviewed documents (previous 15 from M1, M2, and M3 + 5 new documents), continuing the refinement process.

As the model progresses through M0 to M4, it continuously learns from the most informative samples, improving its key phrase identification accuracy.

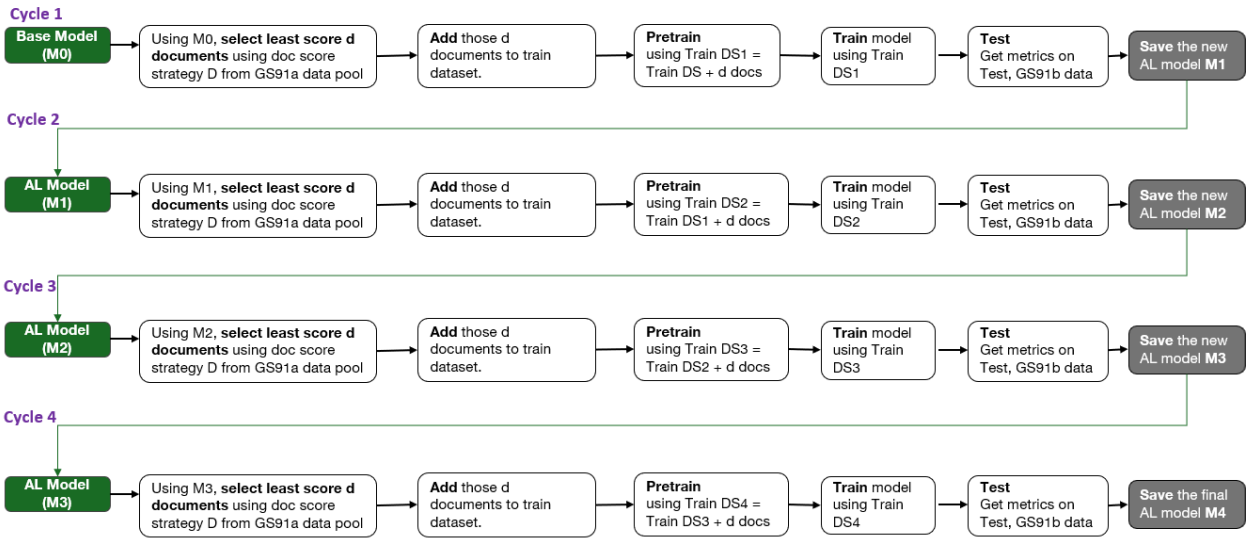

Figure 2: Active Learning life cycle examples from M0 to M3 with elaborated steps

We integrate active learning into this framework by incorporating both established and novel methodologies:

3.3.1 Uncertainty Sampling Strategies: To identify sample documents for HDE annotation, we utilize well-established uncertainty metrics such as MTP(Minimum Token Probability), Margin, and Token Entropy. These metrics help identify tokens or sequences for which the model's predictions exhibit the lowest confidence.

Minimum Token Probability (MTP)

The **Minimum Token Probability (MTP)[Ref 4]** method assesses uncertainty by considering the highest probability assigned to any label for a token. A lower maximum probability indicates greater uncertainty, as the model lacks confidence in any single prediction.

MTP(x)=P(y1|x)∑iP(yi|x)

MTP(x) = \frac{P(y_1 | x)}{\sum_{i} P(y_i | x)}(Latex version)

MTP(x)=∑iP(yi|x)P(y1|x)

$$\mathrm{MTP}(x) = \frac{P(y_1 \mid x)}{\sum_i P(y_i \mid x)}$$

Where:

- MTP(x) represents the model's confidence in the most probable label for token”x”.
- P(y1|x) is the highest predicted probability for token “x”.
- ∑iP(yi|x)\sum_{i} P(y_i | x)∑iP(yi|x) is the sum of probabilities for all possible labels of token “x”.
- Here x is token and yi is the tag for the token x, P(yi|x) is the probability of token x belonging to class yi. The summation goes over each tag/class.

Token Entropy

The **Token Entropy[Ref 17]** method measures uncertainty by calculating the entropy of the probability distribution across all possible labels for each token. Higher entropy indicates greater uncertainty, as the model struggles to assign a confident prediction. By selecting tokens (or documents containing tokens) with the highest entropy, this method effectively identifies the most ambiguous documents for HDE review.
Token Entropy is particularly well-suited for key phrase identification tasks, as it provides a detailed, probability-based view of uncertainty, allowing for precise targeting of areas requiring clarification.

H(x)=−∑iP(yi|x)log□P(yi|x))

$$H(x) = -\sum_i P(y_i \mid x) \log P(y_i \mid x)$$

Where:

- H(x) is the entropy for a token “x”.
- P(yi|x)P(y_i | x)P(yi|x) is the probability of class yi being the correct label for token “x”.
- Here x is token and yi is the tag for the token x, P(yi|x) is the probability of token x belonging to class yi. The summation goes over each tag/class.

Margin Method

The **Margin Method[Ref 18]** calculates uncertainty by examining the difference between the probabilities of the two most likely predicted labels for a token. A smaller margin reflects greater uncertainty, as the model finds it harder to differentiate between the top predictions.
This method is computationally efficient compared to Token Entropy, as it does not require evaluating the full probability distribution. The Margin Method is particularly effective for identifying instances near classification boundaries, where the model’s confidence is weakest.

M(x)=P(y1|x)−P(y2|x)M(x) = P(y_1 | x) - P(y_2 | x)M(x)=P(y1|x)−P(y2|x)

$$M(x) = P(y_1 \mid x) - P(y_2 \mid x)$$

Where:

- M(x) is the margin for token x.
- P(y1|x)P(y_1 | x)P(y1|x) is the highest predicted probability for token “x”.
- P(y2|x)P(y_2 | x)P(y2|x) is the second-highest predicted probability for token “x”.
- Here x is token and yi is the tag for the token x, P(yi|x) is the probability of token x belonging to class yi. The summation goes over each tag/class.

3.3.2 Document-Level Aggregation: We introduce novel methods—KPSum, KPAvg, DOCSum, and DOCAvg [Ref 19]—to aggregate token-level confidence scores into document-level scores. This ensures that entire documents can be prioritized for HDE review, enhancing the efficiency of the annotation process.

DocSum: This method involves summing the model uncertainty (measured by methods such as Token Entropy or Margin, explained above) across all tokens in a document. By considering the cumulative uncertainty of all tokens, **DocSum** identifies documents with the highest overall uncertainty for HDE review. However, **DocSum** tends to favor longer documents, as they naturally contain more tokens, which can lead to a bias toward selecting those documents.

DocAvg: Instead of summing the uncertainty, **DocAvg** calculates the average uncertainty per token in the document. This method ensures that documents with high average uncertainty (rather than just long documents with more tokens) are selected, balancing document length and uncertainty.

KPSum (Key Phrase Sum): Similar to DocSum, this method sums the uncertainty scores of all identified key phrases (instead of all tokens) within a document. It prioritizes documents based on the collective uncertainty of the predicted key phrases, making it suitable when key phrases are the focus rather than all tokens in the document. However, KPSum can be biased toward documents with more key phrases, as it sums the uncertainty associated with each predicted key phrase, so longer lists of key phrases will contribute more to the score.

KPAvg (Key Phrase Avg): This strategy takes the average uncertainty of the predicted key phrases per document, regardless of how many key phrases are in a document. It is particularly effective for documents where only a few key phrases contribute to the overall uncertainty, ensuring that the model focuses on those key phrases that are most uncertain.

Both KPSum and KPAvg focus on the BiLSTM-CRF's the most ambiguous key phrases. These methods prioritize key phrase tokens over all tokens, unlike DOCSum and DOCAvg, which treat all tokens equally. As a result, KPSum and KPAvg ensure that both the model and expert attention are directed toward documents containing the most uncertain key phrases.

Document Sorting and Selection: After computing the document-level confidence scores, all documents in the Active Learning data pool are sorted in ascending order of their uncertainty (i.e., documents with the highest uncertainty are prioritized). This sorting allows for the selection of the most uncertain documents, which are then passed on to HDE for review.

Human Expert Review: The most uncertain documents, as identified through the confidence scores, are reviewed and annotated by the Human Domain Expert (HDE). The review process ensures that the documents are correctly labeled and that any key phrases missed by the model are corrected.

Retraining the Model: Once the documents are labeled by HDE, they are added to the training dataset and removed from the Active Learning data pool. The BiLSTM-CRF model is then retrained using the updated training dataset, which now includes the newly labeled documents. This process allows the model to learn from the continuing updated correct labels and gradually improve its key phrase identification performance.

3.3.3 Comparison of Uncertainty Measures

The main distinction among the three uncertainty measures lies in their focus and computational efficiency. Token Entropy analyzes the entire probability distribution, providing a detailed measure of uncertainty suited for complex, multi-label tasks. In contrast, the Margin Method focuses on the difference between the top two predicted labels, offering a computationally efficient solution for large-scale tasks where quick AL cycles are essential. MTP, on the other hand, simplifies uncertainty assessment by considering only the highest predicted probability, making it useful for flagging low-confidence predictions in simpler tasks while being less computationally intensive than Token Entropy. Table 1 presents side by side comparison of the three measures.

Uncertainty Measure	**Focus**	**Computational Efficiency**	**Best Suited For**
Token Entropy	Analyzes the entire probability distribution	Computationally intensive	Complex, multi-label tasks where full range of potential outcomes is critical
Margin Method	Focuses on the difference between the top two predicted labels	Computationally efficient	Large-scale tasks requiring quick cycles

Minimum Token Probability (MTP)	Considers only the highest predicted probability	Less computationally intensive than Token Entropy, and Margin Method	Identifying low-confidence predictions in simpler tasks or high-throughput scenarios.

Table 1 : Comparison among uncertainty measures used in the active learning pipeline.

4. Experiments

This section outlines the experimental design used to evaluate the effectiveness of various Active Learning strategies for key phrase identification. It begins by detailing the datasets used, the metrics selected for evaluation, and the document selection order strategies applied across Active Learning cycles. We describe the experimental objective, the strategy configurations tested (combinations of uncertainty measures, aggregation methods, and selection order), and explain how each Active Learning cycle was conducted. A comprehensive summary of all experimental settings is provided, along with Table 2, which summarizes the systematic evaluation of different measures and strategies and their combinations used during the experiments. The actual results and performance comparisons are presented in section 4.5.

4.1 Experiment preparation

4.1.1 Document Order in Active Learning Cycles

The order in which documents are added during active learning cycles plays a significant role in shaping the model learning trajectory. We evaluate two strategies:

1. **Actual Order**: Documents are prioritized based on their inherent uncertainty (e.g., calculated by Token Entropy, Margin, or Maximum Token Probability). This strategy mimics a real-world scenario where the most uncertain documents are presented first for HDE annotation.
2. **Reverse Order**: Documents are presented in reverse order of their uncertainty. This serves as a control to evaluate whether prioritizing high-uncertainty documents truly enhances model performance compared to random or low-priority sampling.

The "Actual Order" strategy focuses on presenting the most uncertain documents first for HDE annotation, under the assumption that these documents will provide the greatest benefit in improving model performance. During the Actual Order cycles,the documents on the top tend to be outliers, adding them all at once can pose challenges for the model, potentially making it harder to adjust. In contrast, by prioritizing documents in Reverse Order—where the most uncertain documents are presented last—the selection process begins with samples that are more similar to the existing data. This approach allows the model to adjust more smoothly. This consideration is particularly relevant considering our training data is synthetically generated from a small pool of documents. These comparisons allow us to assess the impact of document order on annotation efficiency and overall model performance.

4.1.2 Experiment Strategies: In each experiment, we test a specific strategy, which is a combination of the uncertainty measure(e.g., Margin, MTP, Token Entropy), document aggregator(e.g., DOCSum, DOCAvg, KPSum, KPAvg), and document order(e.g., Actual, Reverse). These combinations collectively determine how documents are prioritized for HDE annotation during Active Learning cycles. Table 2 presents the evaluation settings we implemented in the experiments. We aim to compare the performance of different Active Learning strategies in terms of recall and F1 improvement and to identify the most effective one for improving the model's performance.

Active Learning Strategies Evaluated in the Experiments				
Uncertainty Measure/ Document Aggregator	DOCSum	DOCAvg	KPSum	KPAvg
MTP	A:NA R: NA	A: S6 R: S5	A: S8 R: S7	A: S10 R: S9
MARGIN	A:NA R: NA	A: S2 R: S1	A: S3 R: S4	A:NA R: NA
TOKEN ENTROPY	A:NA R: NA	A: S12 R: S11	A: S14 R: S13	A:NA R: NA

NA: Not Available; A: Actual, R: Reverse, Sx : Strategy x.

Table 2 : All the active learning strategies (i.e., combinations of uncertainty measure and strategy) experimented in this study

.

4.1.3 Dataset Overview

- **Initial Training Dataset**: Consists of 500 documents and annotated with synthetic labels [Ref 9]. This dataset is used to train the base model (M0) before Active Learning begins.
- **Active Learning Pool**: Comprises 42 HDE labeled (Gold Standard data) documents, which are gradually introduced into the training process during Active Learning cycles. These documents are selected based on uncertainty measures and compared with the

HDEs annotation results before being added to the updated training set. The goal of this dataset is to incrementally improve the model's ability to generalize to real-world data by prioritizing the most informative samples.

- **Test Set**: Consists of 49 HDE labeled documents, which are never used in the training process and remain untouched throughout the Active Learning cycles. This dataset serves as the final benchmark for evaluating the model's performance, ensuring that improvements observed during Active Learning are not due to data leakage but reflect the model's genuine generalization capabilities.

  The HDE labeled documents were annotated by three medical domain experts and the details about the annotation have been published [Ref 9, 22, 23].

4.2 Metrics used in the experiments

In this study, we use two evaluation metrics to assess model performance: Recall and F1-score. These metrics are tracked across successive Active Learning cycles to measure how effectively the model improves in identifying key phrases as it incorporates human feedback.

Recall is selected as the primary metric of interest. This decision is motivated by the goal of minimizing the risk of missing any key phrases. A high recall indicates that the model successfully captures as many correct key phrases as possible, even if it occasionally includes false positives. In contrast, F1-score, which balances both precision and recall, is less suitable as a primary metric in this context because it penalizes false positives—an aspect that is not critical to our task.

F1-score provides a more holistic view of model performance. F1-score allows us to distinguish between strategies that achieve similar recall levels but differ in precision, offering a more nuanced comparison of Active Learning strategies.

4.3 Hyperparameter and Experiment Settings

To ensure fair and robust comparisons across all Active Learning strategies, we fine-tuned and standardized the following **hyperparameters** and **experimental configurations**:

- **Active Learning Pool**: 42 gold standard documents
- **Initial Training Data**: 500 synthetically labeled documents [Ref 9]
- **Number of Active Learning Cycles**: 4
- **Documents Added per Cycle**: 5
- **Number of Copies per Document per Cycle**: 1
- **BiLM Retrained**: Yes, the Language Model (BiLM) is retrained after each Active Learning cycle
- **Sequence Length (BiLM)**: 10
- **Learning Rate**: 0.001
- **Dropout Rate**: 0.3

- **Repetitions**: All experiments were run **25 times**, and the **average results** are reported in the results tables

These settings were chosen to ensure consistency, stability, and fairness in evaluating the performance of various uncertainty-based document selection strategies in the Active Learning framework.

4.4 Results

The **average Recall** and **average F1-score** metrics for each Active Learning cycle were reported for each strategy. The results presented here aim to evaluate how different strategies impact the rate and magnitude of these improvements, particularly in the context of recall. The strategy number is corresponding to the combination lists in Table 2.

Strategy 1:

Uncertainty measure : MARGIN

Document Aggregator: DOCAvg

Document Selection Order: REVERSE

Git link (provides the full sets of testing results) :

MARGIN DOC AVG REVERSE ORDER		
	RECALL AVG	**F1 AVG**
M0	0.43	0.5395
M1	0.5226	0.5743
M2	0.5352	0.5735
M3	0.5252	0.5735
M4	0.5226	0.573

Table 3 : Active Learning experiment results for Margin Doc Avg Reverse strategy (n =25)

The results demonstrate a significant improvement in recall, which is our primary metric of interest, as well as a corresponding increase in F1-score, albeit to a lesser extent.

The recall improves from 0.43 at M0 to 0.5226 at M1, highlighting the impact of the active learning cycle in addressing the model’s initial gaps. This substantial early gain is consistent with the general observation that most improvement tends to occur in the first few cycles of

active learning [Ref 20]. After the first cycle, the recall continues to improve gradually, peaking at 0.5352 at M2, before stabilizing with marginal fluctuations in subsequent cycles.

Interestingly, the F1-score also increases, from 0.5396 at M0 to 0.5743 at M1, suggesting that the refinement in recall does not come at the cost of drastically lower precision. MARGIN prioritizes the most uncertain tokens. This ensures that documents selected for annotation contain ambiguous or difficult key phrases, increasing the chances of capturing more key phrases and improving recall. DOC AVG selects documents based on their average uncertainty. Reverse Order selects documents with lower uncertainty first, allowing the model to learn from easier cases before tackling more challenging ones, leading to a steady improvement in recall. Synthetic data limitations – The M0 model, trained only on synthetic data. Even a small number of real documents at M1 helps refine key phrase detection.

However, after M1, the F1-score exhibits a plateau, stabilizing around 0.573, even as recall shows slight improvements. This behavior suggests that while the strategy is effective at capturing more key phrases, the additional data incorporated in later cycles does not significantly improve precision, possibly because the selected documents add noise or repetitive patterns.

Strategy 2:

Uncertainty measure : MARGIN

Document Aggregator: DOCAvg

Document Selection Order: ACTUAL

Git link :

MARGIN DOC AVG ACTUAL ORDER		
	RECALL AVG	**F1 AVG**
M0	0.4456	0.548
M1	0.5208	0.5744
M2	0.5332	0.5784
M3	0.5308	0.5792
M4	0.5252	0.5776

Table 4 : Active Learning experiment results for Margin Doc Avg Actual strategy (n =25)

S2 experiments results indicate a significant improvement in recall, alongside a moderate increase in F1-score.

Recall improves from 0.4456 at M0 to 0.5208 at M1. The steady improvement continues until M2 (0.5332) before stabilizing in later cycles. This suggests that the most informative documents are introduced early, and additional cycles contribute only marginal refinements.

F1-score follows a similar trajectory, increasing from 0.548 at M0 to 0.5744 at M1, demonstrating that the increase in recall does not drastically reduce precision. After M2, the F1-score stabilizes, fluctuating slightly but maintaining overall consistency. This strategy effectively enhances recall while maintaining precision balance.

Strategy 3 (Table 5):

- **Uncertainty measure : MARGIN**
- **Document Aggregator: KPSum**
- **Document Selection Order: ACTUAL**
- **Git link :**

MARGIN KP SUM ACTUAL ORDER		
	RECALL AVG	**F1 AVG**
M0	0.4404	0.546
M1	0.5224	0.5752
M2	0.534	0.5732
M3	0.5252	0.572
M4	0.528	0.5756

Table 5 : Active Learning experiment results for Margin KPSum Actual strategy (n =25)

S3 experiments results show a notable increase in recall, along with a moderate improvement in F1-score.

Recall improves from 0.4404 at M0 to 0.5224 at M1, demonstrating a substantial early gain. The recall continues to increase gradually, reaching 0.534 at M2, after which it stabilizes with minor fluctuations.

F1-score also improves, increasing from 0.546 at M0 to 0.5752 at M1. However, after M1, it fluctuates slightly around the 0.573–0.575 range, indicating that while the strategy enhances recall, additional cycles do not significantly impact precision.

Strategy 4 (Table 6):

- **Uncertainty measure : MARGIN**
- **Document Aggregator: KPSum**
- **Document Selection Order: REVERSE**
- **Git link :**

MARGIN KP SUM REVERSE ORDER		
	RECALL AVG	**F1 AVG**
M0	0.4392	0.542
M1	0.5024	0.5552
M2	0.5184	0.5572
M3	0.5323	0.558
M4	0.5152	0.5576

Table 6 : Active Learning experiment results for Margin KPSum Reverse strategy (n =25)

S4 experiments resulted in noticeable improvements in both recall and F1-score over the active learning cycles, with a more substantial increase in recall compared to F1.

The recall improved from 0.4392 at M0 to 0.5024 at M1, reflecting the positive impact of the active learning cycle in refining the model's understanding of key phrases. The recall continues to increase, reaching a peak of 0.5323 at M3, before stabilizing with slight fluctuations in the later cycles.

The F1-score also increases, from 0.542 at M0 to 0.5576 at M4, though the improvement is more gradual compared to recall. The relatively steady F1-score suggests that while recall-focused strategies are prioritized, they do not significantly harm precision.

The KPSum aggregator typically prioritizes documents with a large number of key phrases when used with actual order, focusing on documents with higher uncertainty in the early cycles. However, with reverse order, the model begins by selecting documents with fewer key phrases. This shift in focus leads to a more modest initial improvement in recall at M1 compared to the strategy using actual order.

Since the reverse order starts with documents that are more certain in terms of key phrase identification, the improvement in recall is slower in the early cycles. These documents contribute less to recall in the initial stages. Over time, as the model progresses, it starts working on documents with higher uncertainty, which drives further improvements in recall.

The modest improvement in F1-score can be explained by this shift in focus. Documents selected early on in the reverse order have fewer key phrases, contributing less to both recall and precision in the beginning, which results in slower F1 improvements.

Strategy 5 (Table 7):

- **Uncertainty measure : MTP**
- **Document Aggregator: DOCAvg**
- **Document Selection Order: REVERSE**
- **Git link :**

MTP DOC AVG REVERSE ORDER		
	RECALL AVG	**F1 AVG**
M0	0.45	0.5468
M1	0.5204	0.5736
M2	0.5432	0.5696
M3	0.5296	0.5796
M4	0.5332	0.5764

Table 7 : Active Learning experiment results for MTP Doc Avg Reverse strategy (n =25)

S5 experiments yielded noticeable improvements in both recall and F1-score, with recall showing a more substantial increase.

The recall improves from 0.45 at M0 to 0.5204 at M1, demonstrating the positive effect of incorporating active learning.The recall continues to increase gradually, reaching a peak of 0.5432 at M2, before stabilizing with minor fluctuations in later cycles.

F1-score also increases, from 0.5468 at M0 to 0.5764 at M4. While the F1-score improvement is more gradual compared to recall, it still reflects the model's ability to balance both precision and recall effectively, with no drastic drop in precision.

This strategy may lead to a steadier and more controlled improvement in recall, as the model is better able to handle the increasing complexity of the data over time.

Overall, the MTP uncertainty measure combined with DOCAvg and REVERSE ORDER demonstrates the ability to improve recall effectively while maintaining a reasonable balance between recall and precision.

Strategy 6 (Table 8):

- **Uncertainty measure : MTP**
- **Document Aggregator: DOCAvg**
- **Document Selection Order: ACTUAL**
- **Git link :**

MTP DOC AVG ACTUAL ORDER		
	RECALL AVG	**F1 AVG**
M0	0.4356	0.542
M1	0.518	0.5728
M2	0.5464	0.5832
M3	0.5192	0.5764
M4	0.5252	0.5696

Table 8 : Active Learning experiment results for MTP Doc Avg Actual strategy (n =25)

S6 experiments resulted in improvements in both recall and F1-score, with recall showing a more substantial increase over the active learning cycles.

The recall improved from 0.4356 at M0 to 0.518 at M1, indicating a positive effect of the active learning cycle. Recall continues to improve, reaching a peak of 0.5464 at M2, before stabilizing at 0.5252 at M4, with some fluctuations in the later cycles.

The F1-score also increased, from 0.542 at M0 to 0.5696 at M4, though the improvement was more gradual compared to recall. This gradual increase in F1-score reflects the model's ability to maintain a reasonable balance between precision and recall.

The actual order for document selection may contribute to faster improvements in recall as the model focuses on more challenging key phrases in the early cycles.

Overall, the combination of MTP uncertainty measure, DOCAvg document aggregator, and ACTUAL ORDER demonstrates effective improvements in recall and a modest increase in F1-score, suggesting the strategy's capability to refine the model's performance while maintaining a balance between recall and precision.

Strategy 7 (Table 9):

- **Uncertainty measure : MTP**
- **Document Aggregator: KPSum**
- **Document Selection Order: REVERSE**
- **Git link :**

MTP KP SUM REVERSE ORDER		
	RECALL AVG	**F1 AVG**
M0	0.475	0.5608

M1	0.5254	0.5708
M2	0.5404	0.5754
M3	0.5367	0.5758
M4	0.5342	0.5746

Table 9 : Active Learning experiment results for MTP KPSum Reverse strategy (n =25)

S7 experiments resulted in a steady improvement in both recall and F1-score, with recall exhibiting a notable increase early on before stabilizing in later cycles.

The recall improved from 0.475 at M0 to 0.5254 at M1, reflecting the early benefits of active learning. The recall continued to increase, reaching a peak of 0.5404 at M2, before showing slight fluctuations in later cycles, stabilizing around 0.5342 at M4. The F1-score followed a similar trend, increasing from 0.5608 at M0 to 0.5746 at M4, indicating that the strategy maintained a balance between recall and precision.

KPSum and the reverse order likely contribute to the more gradual improvements in recall compared to other strategies where highly uncertain documents are prioritized early. The model first learns from documents with lower key phrase density before transitioning to more informative ones, which could explain the steady rather than sharp increases in recall.

Overall, this strategy demonstrates the effectiveness of MTP with KPSum in improving recall while keeping F1-score stable. The use of reverse order alters the document selection dynamics, potentially leading to a more incremental learning process.

Strategy 8 (Table 10):

- **Uncertainty measure : MTP**
- **Document Aggregator: KPSum**
- **Document Selection Order: ACTUAL**
- **Git link :**

MTP KP SUM ACTUAL ORDER		
	RECALL AVG	**F1 AVG**
M0	0.4388	0.5432
M1	0.5324	0.5768
M2	0.5264	0.5772
M3	0.5436	0.586
M4	0.5344	0.5856

Table 10 : Active Learning experiment results for MTP KPSum Actual strategy (n =25)

S8 experiments led to a notable increase in recall, particularly in the early active learning cycles, while maintaining a stable F1-score.

The recall improved significantly from 0.4388 at M0 to 0.5324 at M1, highlighting the impact of selecting highly uncertain documents with a large number of key phrases early in the process. The recall peaked at 0.5436 at M3 before slightly declining to 0.5344 at M4. The F1-score followed a consistent upward trend, increasing from 0.5432 at M0 to 0.5856 at M4, suggesting that the model maintained a good balance between recall and precision.

Using KPSum as the document aggregator in actual order prioritizes documents with a higher number of key phrases, ensuring that the model is exposed to more informative instances in the early stages. This likely contributes to the sharp initial improvement in recall. The fluctuations in later cycles may be due to the limited diversity in the remaining document pool, reducing the model's ability to gain new insights.

Overall, this strategy effectively leverages MTP and KPSum to enhance recall while keeping F1-score stable. The actual order selection ensures that the most informative documents are chosen early, accelerating the learning process.

Strategy 9 (Table 11):

- **Uncertainty measure : MTP**
- **Document Aggregator: KPAvg**
- **Document Selection Order: REVERSE**
- **Git link :**

MTP KP AVG REVERSE ORDER		
	RECALL AVG	**F1 AVG**
M0	0.4232	0.5328
M1	0.524	0.5748
M2	0.5388	0.578
M3	0.5288	0.5748
M4	0.5292	0.5732

Table 11 : Active Learning experiment results for MTP KPAvg Reverse strategy (n =25)

S9 experimentsresulted in a notable increase in recall in the early active learning cycles, with a relatively stable F1-score throughout.

The recall improved significantly from 0.4232 at M0 to 0.524 at M1, showing that the model benefited from the initial learning phase. The recall peaked at 0.5388 at M2 before slightly

declining in later cycles, which may indicate that the model quickly learned from the most informative documents available early on. The F1-score followed a consistent trend, increasing from 0.5328 at M0 to 0.578 at M2, then stabilizing around 0.5732 at M4.

Since KPAvg selects documents based on the average key phrase uncertainty rather than the total number of key phrases, this strategy may have provided a more balanced selection of documents. However, the use of reverse order means that the model started with documents that were initially more certain, the significant improvement for both recall and F1-score from M0 to M1 indicate mixed signals about the order of documents.

Overall, this strategy effectively improves recall early while keeping the F1-score stable. The use of KPAvg in reverse order provides a more uniform document selection process, though the later-stage fluctuations suggest that the diversity of available documents might become a limiting factor over time.

Strategy 10 (Table 12):

- **Uncertainty measure : MTP**
- **Document Aggregator: KPAvg**
- **Document Selection Order: ACTUAL**
- **Git link :**

MTP KP AVG ACTUAL ORDER		
	RECALL AVG	**F1 AVG**
M0	0.4233	0.5296
M1	0.5213	0.5688
M2	0.5767	0.5333
M3	0.5292	0.5779
M4	0.5017	0.5654

Table 12 : Active Learning experiment results for MTP KPAvg Actual strategy (n =25)

S10 experiments led to an initial rise in recall and F1-score, followed by fluctuations in later cycles.

The recall improved from 0.4233 at M0 to 0.5213 at M1, showing the model’s early learning from highly uncertain documents. The recall peaked at 0.5767 at M2 but then declined in the

following cycles, suggesting that the newly selected documents might not have contributed as effectively to further improving recall. The F1-score followed a different trajectory, rising from 0.5296 at M0 to 0.5688 at M1, stabilizing near 0.5779 at M3, and then slightly dropped at M4.

Since KPAvg selects documents based on the average uncertainty of key phrases, it balances the focus across different types of documents. However, with the actual order, the model was consistently exposed to the most uncertain documents first, which may have led to initial gains but also caused later fluctuations when the remaining pool became less diverse.

Overall, this strategy initially boosted recall and F1, but the later cycles showed instability, possibly due to limitations in the available document pool. Further investigation into the impact of uncertainty-based selection on later-stage performance could provide insights into stabilizing the learning process.

Strategy 11 (Table 13):

- **Uncertainty measure : Token Entropy**
- **Document Aggregator: DOCAvg**
- **Document Selection Order: REVERSE**
- **Git link :**

Token Entropy DOC AVG REVERSE ORDER		
	RECALL AVG	**F1 AVG**
M0	0.4313	0.5342
M1	0.5325	0.5746
M2	0.5171	0.5688
M3	0.5367	0.5779
M4	0.5188	0.5725

Table 13 : Active Learning experiment results for Token Entropy Doc Avg Reverse strategy (n =25)

S11 experiments indicate an initial boost in recall and F1-score, followed by fluctuations in later cycles.

The recall improved from 0.43125 at M0 to 0.5325 at M1, showing a strong early learning phase. However, it dropped to 0.5171 at M2 before recovering slightly at M3 (0.5367) and then declining again at M4 (0.5188). This pattern suggests that the model benefits initially from the reverse order selection of documents but struggles with stability in later cycles.

Similarly, the F1-score increased from 0.5342 at M0 to 0.5746 at M1, showing that the early selected documents contributed well to improving both recall and precision. However, as seen

with recall, F1 fluctuated in later cycles, peaking at 0.5779 at M3 before slightly declining at M4 (0.5725), as shown in Table 13.

Since DOCAvg aggregates uncertainty across the document, it selects documents that exhibit overall high entropy. In combination with reverse order, this means that documents with relatively lower uncertainty were selected earlier, while more uncertain documents were encountered later. This might explain the recall fluctuations, as the initially selected documents may not have provided a strong enough foundation for sustained improvement. Overall, this strategy led to early gains but lacked stability in later cycles.

Strategy 12 (Table 14):

- **Uncertainty measure : Token Entropy**
- **Document Aggregator: DOCAvg**
- **Document Selection Order: ACTUAL**
- **Git link :**

Token Entropy DOC AVG ACTUAL ORDER		
	RECALL AVG	**F1 AVG**
M0	0.4333	0.5379
M1	0.5279	0.5763
M2	0.5383	0.5763
M3	0.5292	0.5708
M4	0.5188	0.5713

Table 14 : Active Learning experiment results for Token Entropy Doc Avg Actual order strategy (n =25)

S12 experiments show an early improvement in both recall and F1-score, followed by slight fluctuations in later cycles.

The recall improved from 0.4333 at M0 to 0.5279 at M1, indicating a strong early learning phase as the most uncertain documents were prioritized. The recall peaked at 0.5383 at M2, after which it declined slightly in M3 (0.5292) and M4 (0.5188). This suggests that while the initial selection of uncertain documents contributed effectively to recall, later selections of documents did not provide a sustained boost.

The F1-score followed a similar pattern, rising from 0.5379 at M0 to 0.5763 at M1,maintaining this value at M2 (0.5763), and then showing minor fluctuations in M3 (0.5708) and M4 (0.5713). The stable F1-score suggests that precision remained relatively steady while recall varied.

Since DOCAvg aggregates uncertainty across an entire document, it selects documents that exhibit high overall entropy. With actual order, the most uncertain documents were selected first, leading to stronger improvements early on. However, as the most uncertain documents were exhausted, the remaining selections of documents did not contribute as effectively to continued improvement, explaining the stabilization and slight decline in later cycles.

Overall, this strategy demonstrated strong early gains, but the diminishing improvements in later cycles highlight potential limitations in selecting only the most uncertain documents first. Future experiments could explore balancing uncertainty with additional selection criteria to sustain performance across cycles.

Strategy 13 (Table 15):

- **Uncertainty measure : Token Entropy**
- **Document Aggregator: KPSum**
- **Document Selection Order: REVERSE**
- **Git link :**

Token Entropy KP SUM REVERSE ORDER		
	RECALL AVG	**F1 AVG**
M0	0.428	0.5368
M1	0.5184	0.572
M2	0.5396	0.5772
M3	0.5356	0.5812
M4	0.5216	0.5704

Table 15 : Active Learning experiment results for Token Entropy KPSum Reverse order strategy (n =25)

S13 experiments indicate steady improvements in recall and F1-score in the early cycles, followed by slight fluctuations in the later stages.

The recall increased from 0.428 at M0 to 0.5184 at M1, showing a strong early gain as the model learned from documents with lower uncertainty due to the reverse selection order. The recall continued to improve, peaking at 0.5396 at M2, before slightly declining in M3 (0.5356) and M4 (0.5216).

The F1-score followed a similar trend, starting at 0.5368 at M0, rising steadily to 0.5812 at M3, and then slightly decreasing to 0.5704 at M4. The F1-score's relatively stable trajectory suggests that while the recall fluctuated, precision was not significantly affected.

Since KPSum prioritizes documents with a high number of key phrases, using reverse order means that the model initially selects documents with fewer key phrases. This likely contributes to the moderate initial recall improvement at M1, as the selected documents provide a more controlled learning process. However, by later cycles, as the pool shrinks, the model has fewer high-uncertainty, high-key phrase documents to learn from, leading to stabilization and minor fluctuations.

Overall, this strategy demonstrates that Token Entropy with KPSum in reverse order provides consistent improvements in recall and F1-score in the early cycles but exhibits diminishing returns as the document pool depletes. Future explorations could examine hybrid selection approaches to maintain performance gains across later cycles.

Strategy 14 (Table 16):

- **Uncertainty measure : Token Entropy**
- **Document Aggregator: KPSum**
- **Document Selection Order: ACTUAL**
- **Git link :**

Token Entropy KP SUM ACTUAL ORDER		
	RECALL AVG	**F1 AVG**
M0	0.4392	0.5433
M1	0.5308	0.5767
M2	0.5254	0.5771
M3	0.5442	0.5863
M4	0.5358	0.5867

Table 16 : Active Learning experiment results for Token Entropy KPSum Actual order strategy (n =25)

S14 experiments indicate strong early improvements in recall and F1-score, with performance stabilizing in later cycles.

The recall improved from 0.4392 at M0 to 0.5308 at M1, demonstrating a significant initial gain as the model prioritized high-uncertainty documents with a large number of key phrases. The recall continued to fluctuate, dropping slightly at M2 (0.5254) before peaking at 0.5442 at M3, and then stabilizing at 0.5358 at M4.

Similarly, the F1-score showed steady improvements, increasing from 0.5433 at M0 to 0.5767 at M1. The F1-score remained stable in the following cycles, peaking at 0.5867 at M4, suggesting that the model maintained a balance between recall and precision throughout the process.

Since KPSum selects documents with a high number of key phrases, and actual order prioritizes high-uncertainty documents, this strategy ensures that the model learns from challenging, high-information instances early in the active learning process. The slight dip in recall at M2 suggests that some selected documents may have introduced noise or ambiguous key phrase annotations, but the overall trend shows effective learning.

Overall, this strategy demonstrates that Token Entropy with KPSum in actual order effectively boosts recall and F1-score, particularly in early cycles. The stabilization in later cycles indicates that the model benefits from high-information documents early on, leading to consistent performance improvements across the active learning process.

Strategy 15 (Table 17):

In the S15 experiments, documents were added randomly during each Active Learning cycle, as opposed to the previous experiments where documents were selected based on their confidence scores. This approach serves as a baseline to assess the effectiveness of uncertainty-based document selection strategies.

To ensure consistency and comparability with the other experiments, this experiment is also conducted over 4 Active Learning cycles, with the language model retrained at the end of each cycle. All other hyperparameters, including the initial training dataset size, document pool size, and learning rate, are kept identical to those in the previous experiments.

Git link :

RANDOM		
	RECALL AVG	**F1 AVG**
M0	0.44	0.5432
M1	0.5236	0.5796
M2	0.5876	0.5776
M3	0.5196	0.5712
M4	0.5096	0.574

Table 17 : Active Learning experiment results for Random strategy (n =25)

In the random experiments, documents were selected randomly rather than based on uncertainty measures or document aggregation heuristics. The results show a strong initial improvement in recall and F1-score, followed by fluctuations in later cycles.

The recall increased significantly from 0.44 at M0 to 0.5236 at M1, indicating that the model benefited from diverse training instances in the early stages. Interestingly, recall peaked at 0.5876 at M2 but then dropped to 0.5196 at M3 and 0.5096 at M4, as shown in Table 17,

suggesting that the random selection introduced significant variability in learning effectiveness across cycles.

The F1-score followed a similar trend, starting at 0.5432 at M0 and reaching 0.5796 at M1, before reaching at 0.5776 at M2. However, it declined slightly at M3 (0.5712) and M4 (0.574). This suggests that while recall fluctuated more, precision remained relatively stable across cycles.

Unlike uncertainty-based selection methods, random sampling does not guarantee that the most informative documents are chosen early, leading to inconsistencies in model improvement across cycles. The spike in recall at M2 suggests that the model happened to receive more informative samples, but the decline afterward indicates that subsequent cycles may have included less informative documents.

Overall, random sampling provides a reasonable baseline, demonstrating that performance can improve without targeted document selection. However, the fluctuations in recall and F1-score highlight the potential inefficiency of this approach compared to uncertainty-driven methods, which prioritize more informative samples systematically.

5 Discussion:

This section presents the results across different strategies and observations from our Active Learning experiments, highlighting how different uncertainty measures, aggregation methods, and document selection orders affect model performance. Across all strategies, we observe that the majority of performance improvement occurs between M0 and M1, particularly in recall, which is the primary metric of this study. We also note that Token Entropy and MTP produce similar document rankings, leading to comparable outcomes across cycles.

Among all evaluated strategies, KP SUM Actual consistently leads to higher recall improvements in the early cycles, attributed to its bias toward selecting documents rich in key phrases. These documents not only help the model capture more key phrase patterns but also improve its general learning, reflected in gains in both recall and F1 scores. In contrast, KP SUM Reverse, while more stable, shows relatively smaller recall gains due to its tendency to select documents with fewer key phrases.

We also analyze F1 and recall dynamics in KP AVG strategies, noting that KP AVG Actual tends to prioritize uncertain key phrases and boosts recall early on, while KP AVG Reverse focuses on relatively certain key phrases, helping reduce false positives and slightly improving F1. Finally, we identify saturation points in later cycles where gains taper off, and propose stopping criteria to ensure efficient use of labeling efforts. Based on these findings, we select KP SUM Actual as the most effective strategy for our use case, balancing both performance improvement and annotation efficiency.

5.1 Similarity Between MTP and Token Entropy Results

We observed that the results for MTP and Token Entropy are strikingly similar while comparing results across strategies. This is because both methods calculate uncertainty in a conceptually similar way, even though their formulas differ. MTP considers the maximum probability, while Token Entropy evaluates the entire probability distribution. However, these measures are directly proportional, meaning the relative ranking of documents based on their uncertainty remains almost the same. As a result, the system's performance using these methods is nearly identical.

5.2 Majority of Performance Gains Occur from M0 to M1:.

A notable observation in our experiments is that the majority of performance improvement occurs during the transition from M0 (the initial model without active learning) to M1 (the first cycle of active learning) regardless of the strategies. This pattern aligns with findings from previous studies [Ref 20], which show that significant gains are often achieved in the early stages of active learning.

This behavior highlights the importance of the initial rounds of feedback, as they address the most critical gaps in the model’s performance, leading to rapid improvements. It emphasizes the need to prioritize high-quality and diverse data during the initial cycles to maximize the impact of the active learning process.

Our use case differs from standard Active Learning scenarios in a crucial way: the training data for M0 is purely synthetic. As a result, when real data labeled by HDE are introduced in M1, the model gains exposure to real-world data patterns. This suggests that the performance improvement from M0 to M1 is not solely due to Active Learning’s uncertainty-based selection but also because the model is learning more representative key phrase structures from the real-world data.

As supporting evidence, in figure 4, we observe that even when documents are added randomly, there is still a significant improvement from M0 to M1. This indicates that while Active Learning strategies optimize sample selection, the introduction of real data itself is a major factor in the observed performance boost. Therefore, a substantial portion of the improvement can be attributed to the transition from synthetic labels to human-labeled data, reinforcing the importance of integrating real annotations early in the learning process to improve generalization.

5.3 Identifying Saturation in Active Learning Cycles:

Determining an effective stopping criterion is crucial for ensuring that Active Learning remains efficient while avoiding unnecessary annotations. A well-defined stopping condition prevents excessive labeling costs and mitigates the risk of overfitting the active learning model.

One of the most straightforward approaches is to monitor the convergence of learning by evaluating the generalization error on a test dataset. This approach, as discussed in H

Ishibashi, et al *[Ref 21]*, provides a way to assess when learning has stabilized. However, obtaining a sufficiently large and representative test set may not always be feasible, making alternative stopping strategies necessary.

A practical alternative is to define saturation-based stopping criteria, where the Active Learning process halts once performance improvements fall below a predefined threshold (e.g., less than 1% improvement in recall or F1-score) or when key evaluation metrics begin to fluctuate around a stable value. Additionally, tracking model uncertainty over multiple cycles can help identify when the most informative data points have already been labeled, signaling a natural stopping point. By establishing clear stopping criteria, Active Learning can be conducted efficiently, ensuring that the process runs until meaningful gains are no longer observed.

As shown in Figure 4, recall continues to improve until M2, but after M2, the recall fluctuates around a stable value, indicating that the model is no longer making significant progress. This suggests that the most informative samples have already been incorporated, and adding more labeled data may not provide substantial gains.

However, it is important to note that we may be observing saturation earlier in our experiments due to the limited size of our Active Learning data pool. With only 42 documents available for Active Learning, the selection choices become increasingly restricted in later cycles, which may contribute to the observed plateau in recall and other performance metrics. In a larger Active Learning dataset, the expected saturation point may occur at a later stage, allowing for more extended model improvement.

By establishing clear stopping criteria, we ensure that Active Learning resources are utilized efficiently, stopping the experiment once meaningful gains are no longer observed while recognizing the limitations imposed by our restricted Active Learning data pool.

5.4 Impact of KP SUM Strategies on Recall Improvement in Active Learning

The KP SUM Actual strategy demonstrates the maximum improvement in recall during early Active Learning cycles due to its inherent bias toward documents containing more key phrases. KP SUM is calculated solely based on key phrases, assigning higher uncertainty scores to documents with a higher density of key phrases. By prioritizing such documents for learning, the model is systematically exposed to key phrase-rich documents, enhancing its ability to learn the specific patterns associated with key phrases. This targeted exposure ensures that the model gains a deeper learning of the structural and contextual cues that define key phrases. Consequently, as the learning progresses, the model becomes increasingly adept at identifying key phrases, leading to consistent recall improvements.

While the recall improvement per Active Learning cycle is slightly greater for KP SUM Actual, it does not show a clear-cut advantage over other strategies. We think that with a larger Active Learning data pool, KP SUM Actual may demonstrate a more significant and consistent improvement.

In Figures 3 and 4, a comparison between MTP KP Sum Actual, MTP DOC AVG Actual, MTP KP Avg Actual, and Random selection is presented. As seen in Figure 3, F1 for MTP KP Sum Actual has shown a gradual and more consistent improvement when compared to all other strategies, highlighting the effectiveness of MTP KP SUM Actual in maintaining steady performance gains over Active Learning cycles.

Additionally, in Figure 4, we observe that recall improvement for KP SUM Actual is not only high but also relatively stable, exhibiting fewer fluctuations compared to other strategies. This stability indicates that KP SUM Actual helps the model learn key phrase patterns more reliably, reinforcing its advantage in producing steady improvements in recall over multiple cycles.

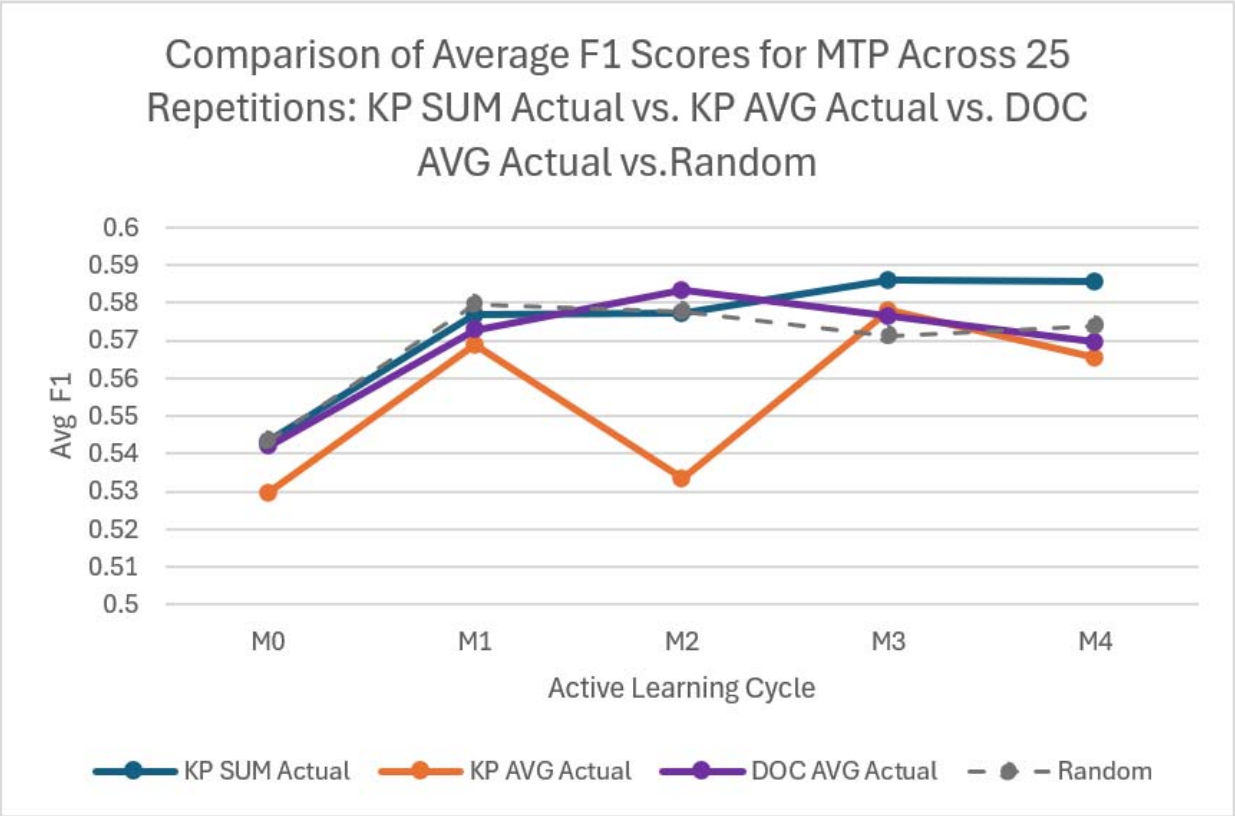

Fig 3: F1 comparison of MTP KP SUM Actual, MTP DOC Avg Actual, MTP KP Avg Actual, and Random Strategies

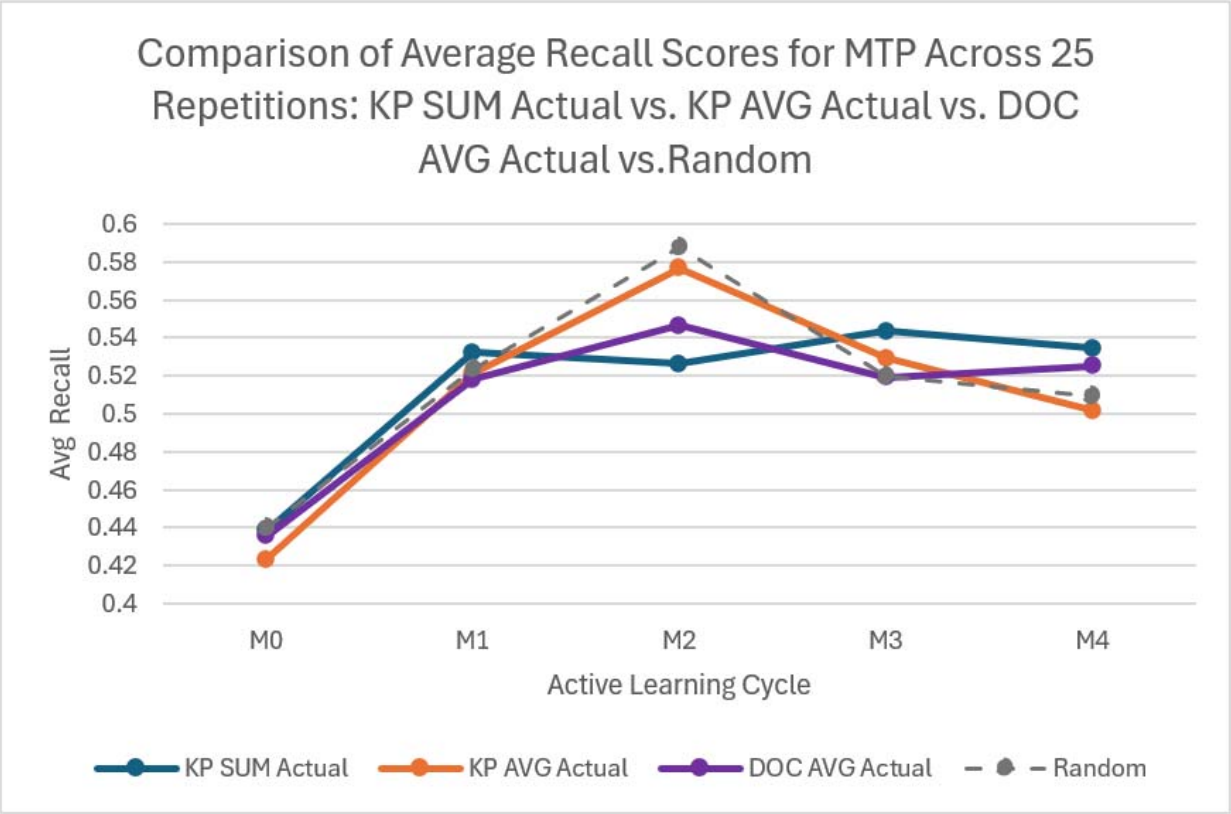

Fig 4: Recall comparison of MTP KP SUM Actual, MTP DOC Avg Actual, MTP KP Avg Actual, and Random Strategies

5.5 Recall Gains for KP SUM strategies:

Another key observation in our experiments is that the recall improvement from M0 to M1 in the KP SUM Actual strategy is consistently greater than in KP SUM Reverse, regardless of the uncertainty measure used. In the figure 6 MTP uncertainty measure is used. This suggests that selecting documents in their actual order leads to more informative updates in the early stages of learning compared to selecting them in reverse order.

One possible explanation for this trend is that KP SUM Actual prioritizes key phrase-rich documents in a way that aligns more naturally with the distribution of the dataset, allowing the model to learn key phrase patterns more effectively in the first cycle. In contrast, the reverse selection order might introduce documents in a less optimal sequence, leading to slightly slower gains in recall.

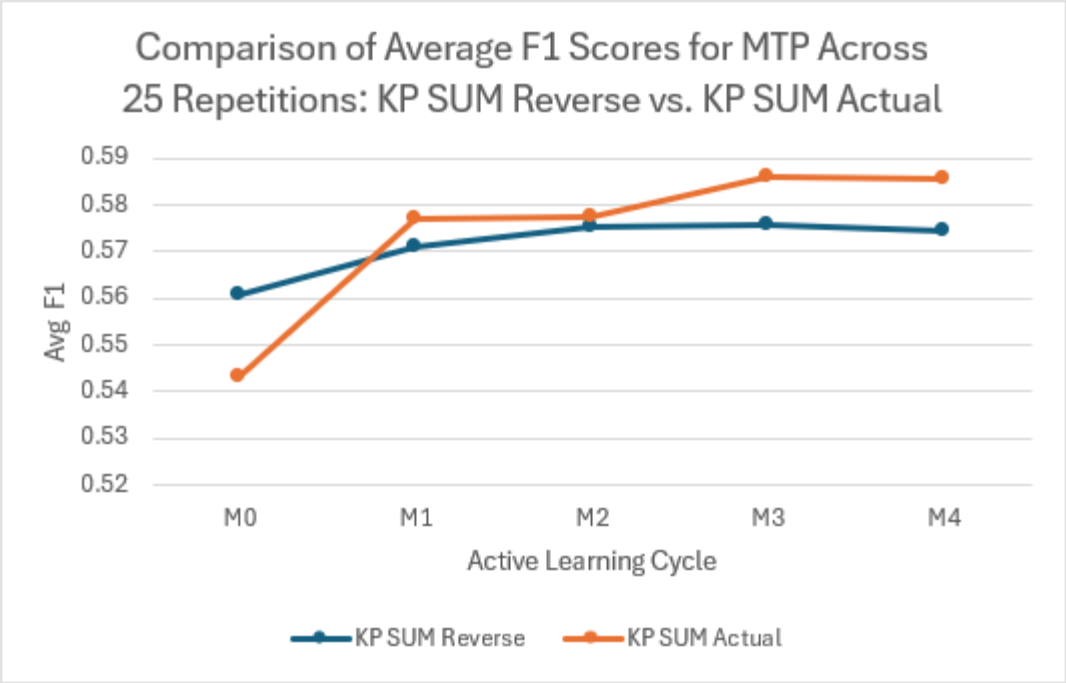

Fig 5: F1 score comparison between MTP KPSum reverse and MTP KPSum Actual

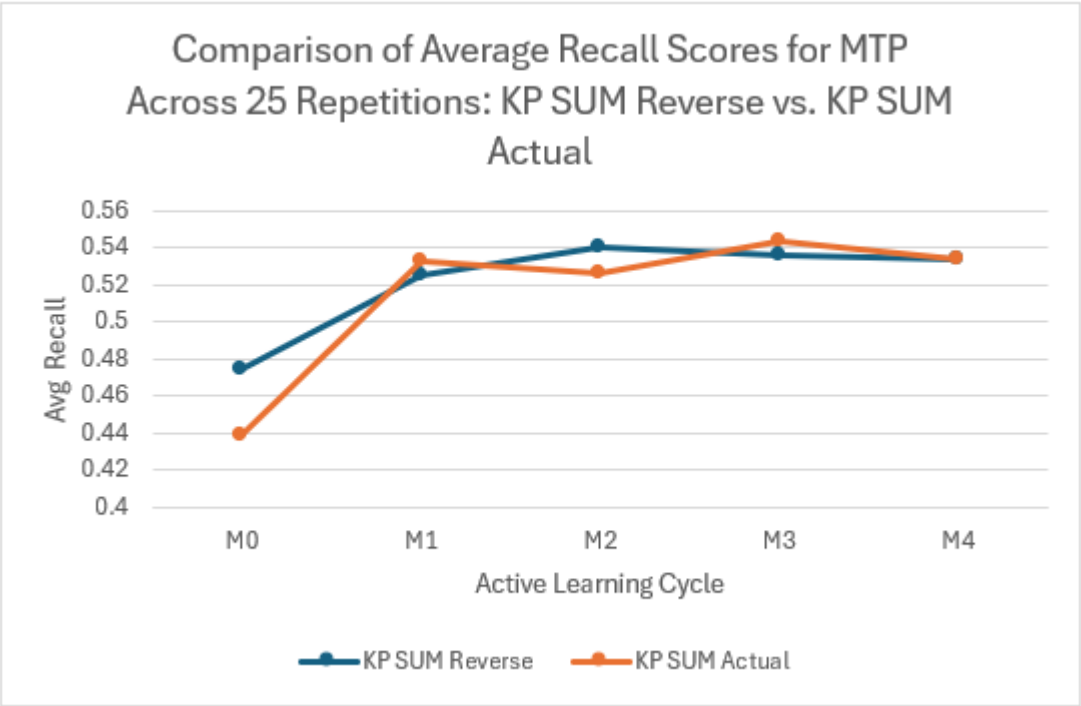

Fig 6: Recall score comparison between MTP KPSum reverse and MTP KPSum Actual

5.6 F1 Score Improvement in KP SUM Strategies:

The increase in F1 score, as shown in figure 5, can be attributed to the following factor: KP SUM Actual naturally selects longer documents. This inadvertently provides the model with more training data per sample, improving its ability to generalize beyond individual key phrases and better differentiate between key phrases and non-key phrases.

In KP SUM Reverse, the model tends to select shorter documents, which in turn has a notable impact on its performance. The key reason for this selection of shorter documents lies in the way key phrases are distributed across documents: shorter documents generally contain fewer key phrases, and as a result, they tend to have fewer tokens overall.

This selection bias towards shorter documents means that the model receives less comprehensive training data from each document for each cycle. With fewer tokens to analyze, the model has less contextual information to learn, reducing its ability to generalize from the data. Since longer documents tend to provide richer contexts and more diverse key phrase patterns, training on shorter documents does not allow the model to gain a robust learning.

As a result, this limitation in context and diversity leads to minimal improvements in the F1 scores under KP SUM Reverse. The model struggles to differentiate key phrases from non-key phrases effectively when provided with such limited information.

5.7 F1 and Recall Dynamics in KP AVG Strategies: Actual vs. Reverse

KP AVG Actual – This strategy tends to select documents with higher key phrase uncertainty, meaning the model is exposed to cases where it is uncertain about key phrase boundaries. As a result, recall improves in early cycles since the model learns to capture more key phrases. Additionally, as more documents are added, the model's general understanding of the task improves, leading to an increase in F1 score.

KP AVG Reverse – This strategy prioritizes documents with lower key phrase uncertainty, meaning the model is exposed to cases where it learns more about non-key phrases and general text patterns. As a result, F1 improves more than in KP AVG Actual during early cycles because the model better differentiates between key phrases and non-key phrases, reducing false positives. Due to the nature of the training data, recall also improves in early cycles, though the primary driver of improvement is better precision.

The figure 7 illustrates the patterns described above. As shown, the F1 score improvement for MTP KP AVG Reverse was higher than that of MTP KP AVG Actual in the earlier cycles, particularly from M0 to M1. The dip in F1 score for MTP KP AVG Actual from M1 to M2 can be attributed to the model needing time to adjust to the more extreme documents added in Actual order, as explained in the Reverse vs. Actual Section. Since Actual order introduces more diverse and uncertain samples, the model takes longer to adapt, leading to fluctuations in F1 during the adjustment phase, as observed in the figure 7.

Similarly, figure 8 also highlights that the recall improvement for MTP KP AVG Actual was higher than that of MTP KP AVG Reverse in the earlier cycles, particularly from M0 to M2. This suggests that Actual order, despite initial fluctuations in F1, helps the model better capture key phrases over multiple cycles, reinforcing the observed trend.

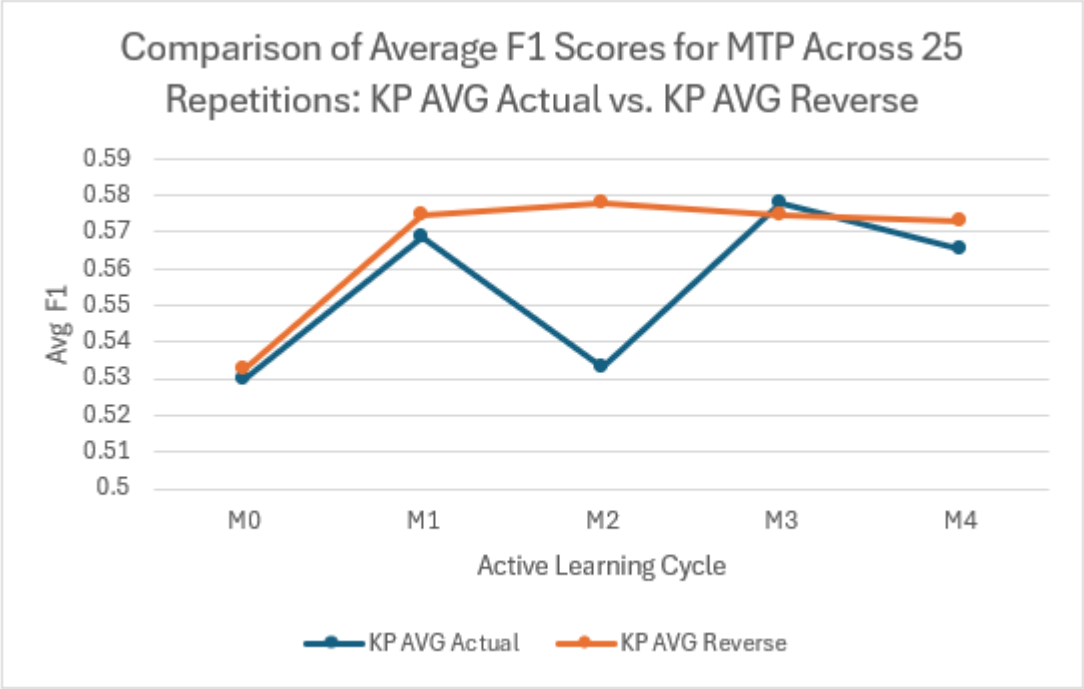

Fig 7: Avg F1 comparison of MTP KpAvg Actual and MTP KpAvg Reverse Strategy.

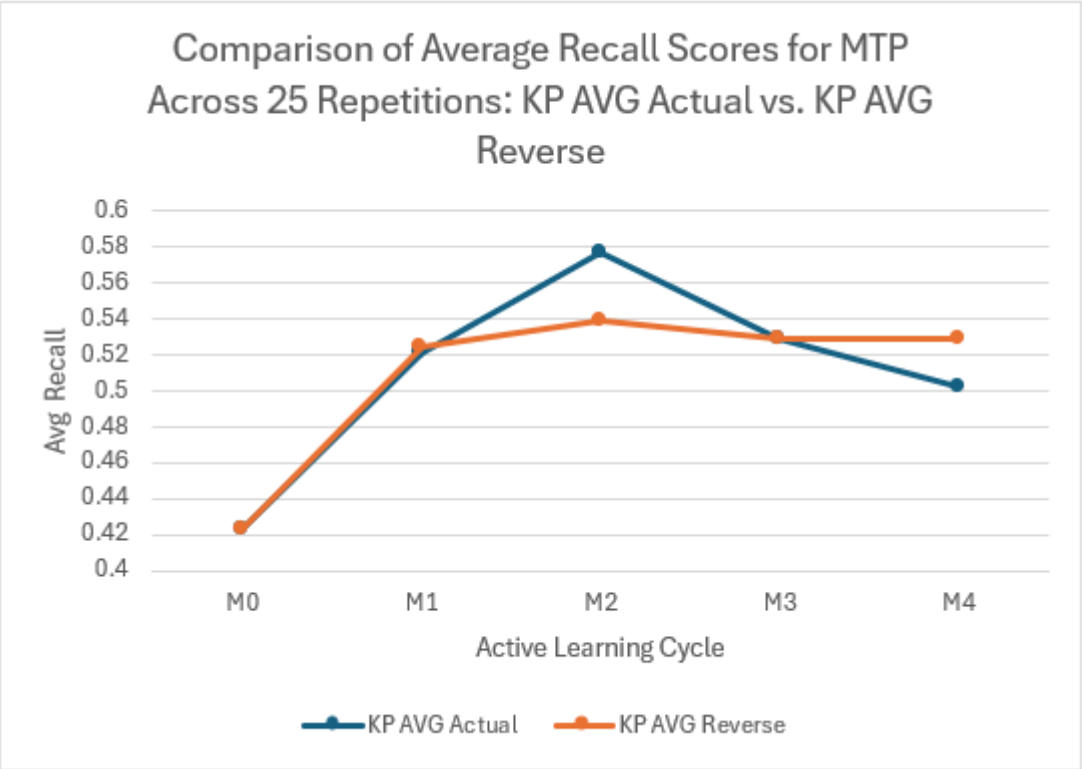

Fig 8: Avg Recall comparison of MTP KpAvg Actual and MTP KpAvg Reverse Strategy.

5.8 Comparing Early-Stage Improvements: Reverse vs. Actual Order

We observe that M0 to M1 improvements in both F1 and Recall are greater for Reverse order, whereas M0 to M2 (or sometimes M3) improvement is greater for Actual order. This pattern is reflected in Figures 9 and 10. Reverse Order prioritizes documents that are most similar to the existing training data (M0). As a result, the model can quickly integrate these familiar patterns, leading to greater initial improvements in F1 and Recall from M0 to M1. Since these documents align more closely with what the model has already seen, both recall and precision improve early on, causing a rapid increase in F1-score in the first cycle.

Actual Order, on the other hand, introduces more extreme or diverse data in the early cycles. The model requires more time to adjust to these differences, resulting in a slower initial improvement from M0 to M1. However, once the model has adapted, it begins to generalize better, leading to a greater improvement from M0 to M2. In some cases, the adjustment may take even longer, and the most significant improvement is observed only by M3, when the model has had more exposure to diverse examples.

This trend suggests that Reverse Order provides a quick boost to F1 by reinforcing familiar patterns early on, while Actual Order leads to more sustained improvements, with delayed but stronger gains in later cycles. This emphasizes the trade-off between fast adaptation and long-term generalization, with Reverse Order excelling in short-term improvements and Actual Order yielding better performance over extended cycles.

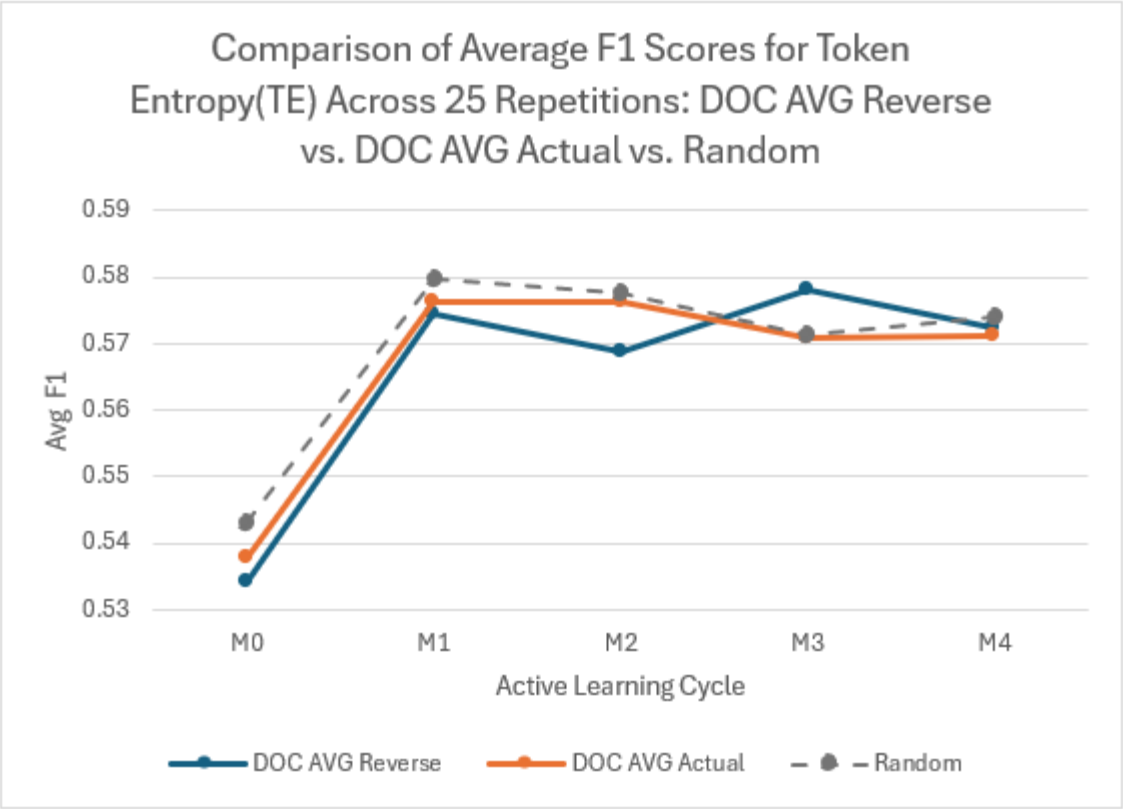

Fig 9: Avg F1 comparison of TE DocAvg Reverse strategy and TE DocAvg Actual strategy

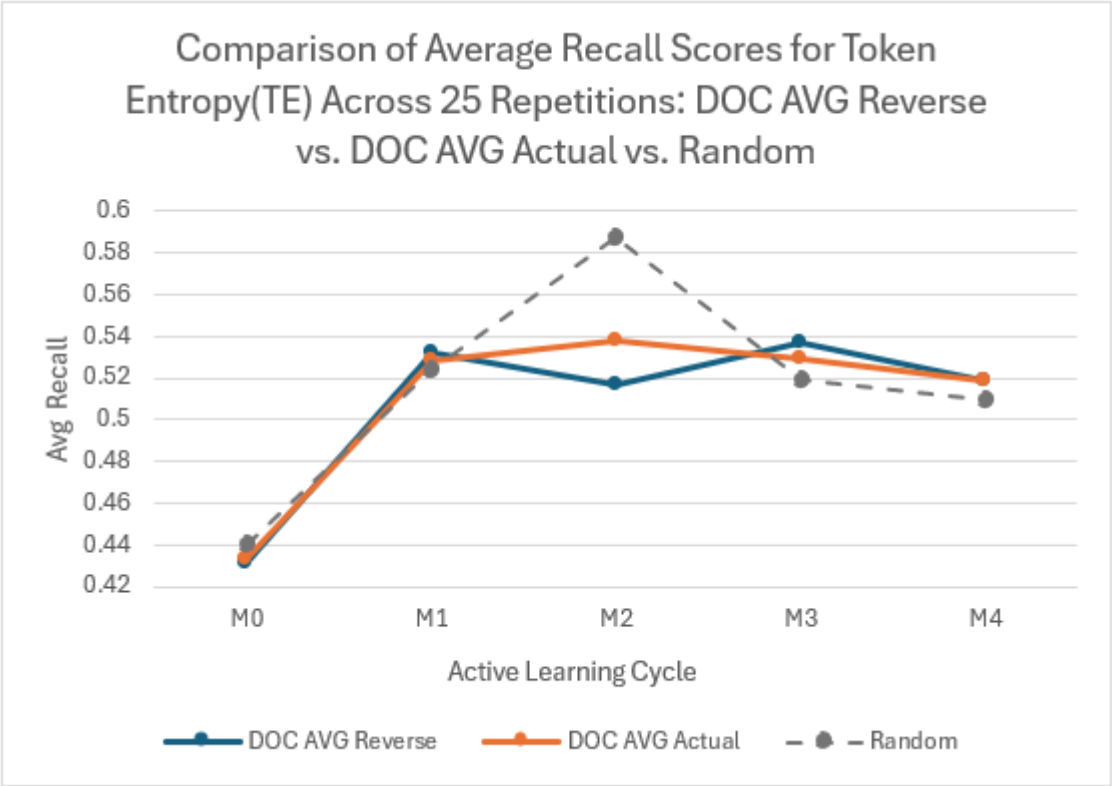

Fig 10: Avg Recall comparison of TE DocAvg Reverse strategy and TE DocAvg Actual strategy

5.9 Active Learning vs. Random Selection: Impact on Recall and Model Stability

In our experiments, we observed a significant difference in the behavior of recall and other performance metrics between Active Learning (AL) and randomly adding documents. Active Learning exhibited rapid improvements in the first two cycles, followed by a plateau in recall as the learning process saturated. This plateau suggests that AL effectively prioritizes the most informative documents early on, reducing uncertainty in the model's predictions and leading to more stable performance.

In contrast, when documents were added randomly, the system initially showed some improvement but soon experienced a steep decline in recall. This decline, accompanied by fluctuations in performance metrics, reflects the inefficiency of random sampling in addressing the model's knowledge gaps. Unlike AL, which systematically selects high-value documents, random selection often introduces less informative data, leading to unstable learning. While the model may eventually stabilize given a sufficiently large dataset. However it seems evident that the lack of a structured selection process causes inconsistencies and delays in performance gains.

These findings highlight the efficiency of Active Learning in leveraging a limited number of human-labeled documents to maximize performance. While random selection may eventually

yield improvements, its inconsistent learning trajectory and performance fluctuations make it less favorable compared to our documentation selection strategies implemented in the Active Learning in practical applications.

In Figures 11 and 12, we compare F1 and Recall scores over Active Learning cycles between an Active Learning strategy (MTP) and Random selection. It is evident that the Active Learning strategy demonstrates consistent improvement across cycles, whereas the Random selection exhibits fluctuating patterns with no clear trend. This highlights the effectiveness of uncertainty-based selection in systematically improving model performance over time, as opposed to the unpredictable behavior seen with randomly added documents.

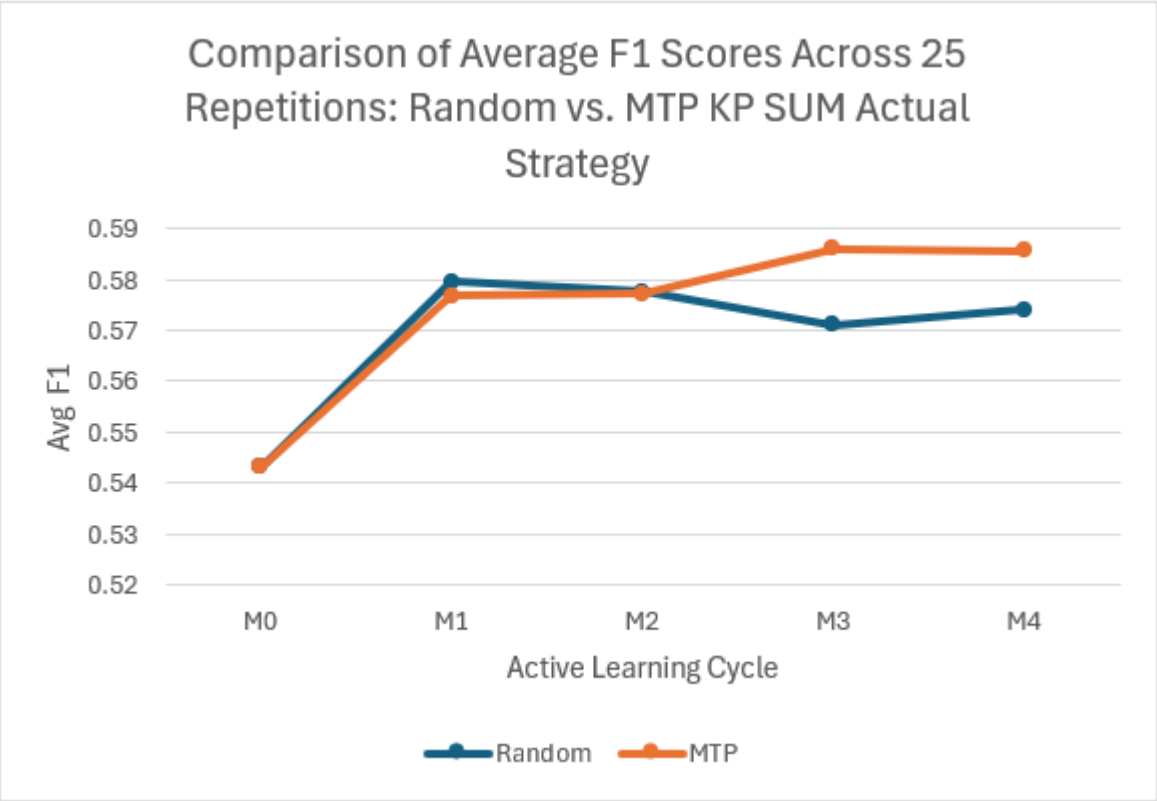

Fig 11: Avg F1 comparison between Random and —---- Strategy

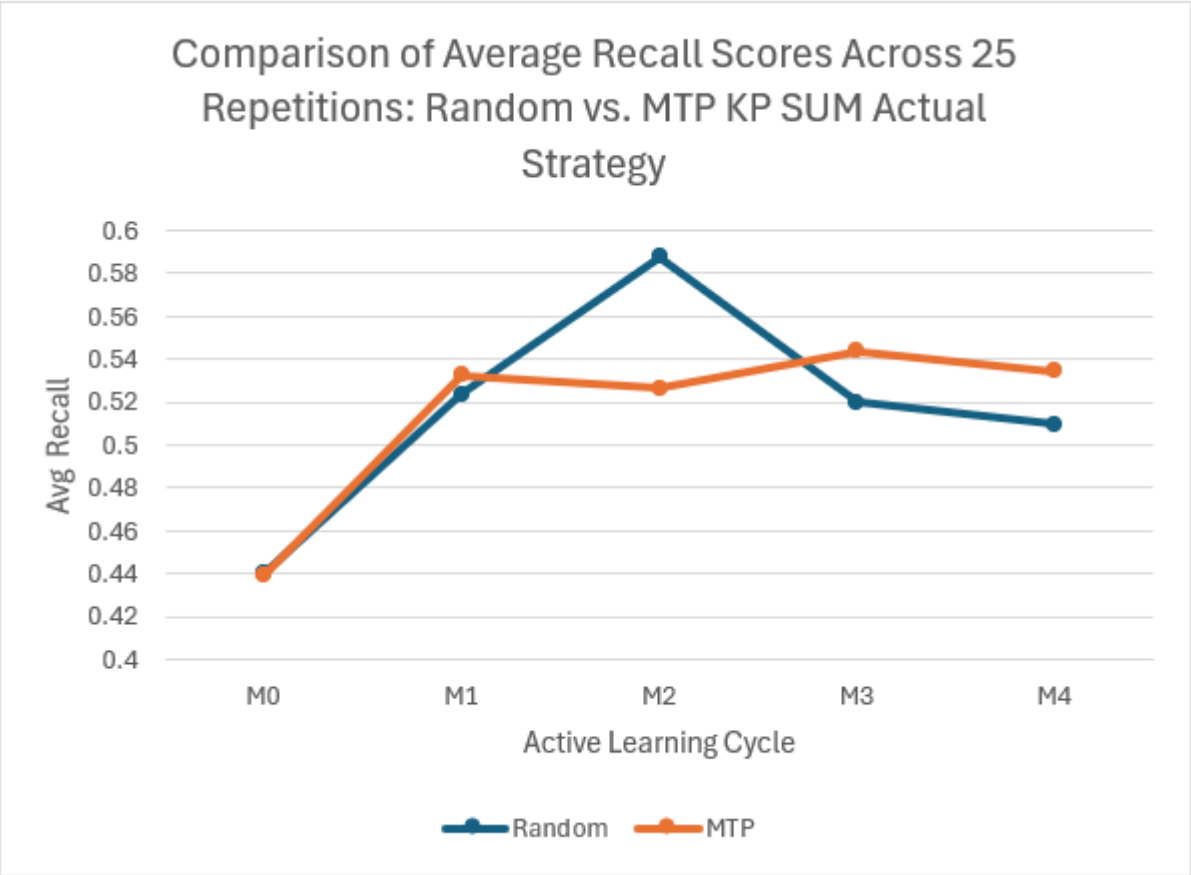

Fig 12 : Avg Recall comparison between Random and MTP KPSum Actual Strategy

5.10 Exploratory Experiments and Alternate Strategies

We attempted topic clustering of the documents to explore whether adding a document from a particular cluster would improve the model's performance on other documents within the same topic. However, this approach did not yield significant results. We think the limited number of documents might make it difficult to form reliable and meaningful clusters. With a larger dataset, topic clustering could be more feasible to test whether adding one document could impact performance across different clusters.

Additionally, we considered the use of a weighted mean as an aggregation strategy. Conceptually, this would serve as a compromise between DocAvg (which averages uncertainty across all tokens) and KPAvg (which focuses only on keyphrases). By adjusting the weight ratio between keyphrase and non-keyphrase tokens, one can control the emphasis placed on keyphrase uncertainty. Though promising, we were unable to experiment with this due to time constraints.

It is important to note that the observed differences across methods were relatively small, likely due to our limited Active Learning data pool. With a larger pool, more pronounced and reliable trends might emerge.

5.11 Future Work

In this study, we explored basic aggregation strategies for estimating document-level uncertainty, primarily focusing on sum-based (e.g., KPSum, DocSum) and average-based (e.g., KPAvg, DocAvg) methods. Future work could explore alternative aggregation techniques such as trimmed mean, which reduces the influence of outlier uncertainty scores, or weighted mean, where keyphrase scores are weighted based on contextual importance. Techniques like median, mode, or box-plot-based selection (e.g., interquartile range filtering) may also provide more robust document representations, especially in skewed distributions.

Section 6: CONCLUSION

In this study, we proposed a generalizable Active Learning framework for identifying candidate terms for a CDSS ontology using a BiLSTM-CRF model, by incorporating HDE feedback. A key contribution of our work is the design and evaluation of novel document-level uncertainty aggregation strategies (KPSum, KPAvg, DocSum, and DocAvg), which enable more effective document selection for annotation. Based on our experimental results, KPSum (Actual) emerged as the most effective strategy for improving keyphrase identification performance in our setting. Meanwhile, the other aggregation methods presented in this manuscript provide a foundation for researchers to compare and select the optimal approach for their own tasks and datasets. Through systematic comparison of various Active Learning strategies and uncertainty measures, this study not only improves keyphrase identification performance but also offers practical benchmarks for future research in similar domains. While our experiments centered on keyphrase extraction, the proposed framework is domain-agnostic and can be adapted to other sequence labeling tasks. Overall, this work contributes to the underexplored application of Active Learning in deep learning-based keyphrase identification, providing a scalable and adaptive foundation for efficient annotation in NLP.

Acknowledgement

This study is made possible by funding from the National Institute of General Medical Sciences (R01GM138589) at the National Institutes of Health and partially through P20GM121342.

REFERENCES:

1. Greenes, R. and G.D. Fiol, Clinical Decision Support and Beyond: Progress and Opportunities in Knowledge-Enhanced Health and Healthcare. 3rd Edition ed. 2023, Cambridge, MA: Elsevier: Academic Press.[q]

2. Davies, J., D. Fensel, and F.V. Harmelen, Towards the semantic web: Ontology-driven knowledge management. 2003, Chichester: John Wiley & Sons.[r]

3. Supekar, K., et al., Knowledge zone: A public respiratory of peer-reviewed biomedical ontologies in MEDINFO 2007, K.e.a. Kuhn, Editor. 2007, IOS Press: Brisbane, Australia. p. 812-816. Rector, A.L. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. in K-CAP'03. 2003. Sanibel Island, Florida, USA.[s]

4. Liu, M., Tu, Z., Zhang, T. *et al.* LTP: A New Active Learning Strategy for CRF-Based Named Entity Recognition. *Neural Process Lett* 54, 2433–2454 (2022). https://doi.org/10.1007/s11063-021-10737-x [MTP, TE, MARGIN][a]

5. Rose, S.J., Engel, D.W., Cramer, N., & Cowley, W. (2010). 1 Automatic keyword extraction from individual documents. (TF IDF for Keyword retrieval) [b]

6. Lafferty, J.D., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *International Conference on Machine Learning*. (CRF for sequence tagging) [c]

7. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *ArXiv, abs/1508.01991*.(BiLSTM CRF Model) [d]

8. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*.(BERT) [e]

9. Goli R, Hubig N, Min H, Gong Y, Sittig D, Robinson D, et al. Keyphrase Identification Using Minimal Labeled Data with Hierarchical Context and Transfer Learning. MedRxiv. 2023.[f]

10. Monarch, R.M., Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI. 2021, Shelter Island, NY: Manning Publications.[t]

11. Settles, B. (2009). Active Learning Literature Survey. [g]

12. Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. In Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 252–256, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2630. URL https://www.aclweb.org/anthology/ W17-2630. [h]

13. Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2904–2909, Brussels, Belgium. Association for Computational Linguistics. [i]

14. Puria Radmard, Yassir Fathullah, and Aldo Lipani. 2021. Subsequence Based Deep Active Learning for Named Entity Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4310–4321, Online. Association for Computational Linguistics.[j].

15. Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active Learning by Acquiring Contrastive Examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. [k]

16. S. -J. Huang, R. Jin and Z. -H. Zhou, "Active Learning by Querying Informative and Representative Examples," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 10, pp. 1936-1949, Oct. 2014, doi: 10.1109/TPAMI.2014.2307881.keywords: {Measurement uncertainty;Uncertainty;Algorithm design and analysis;Correlation;Labeling;Clustering algorithms;Kernel;Active learning;learning with unlabeled data;multi-label learning;informativeness;representativeness}.[u]

17. Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active Learning with Sampling by Uncertainty and Density for Word Sense Disambiguation and Text Classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, Manchester, UK. Coling 2008 Organizing Committee. [l]

18. Dara Bahri, Heinrich Jiang, Tal Schuster, and Afshin Rostamizadeh. 2022. Is margin all you need? An extensive empirical study of active learning on tabular data, arxiv. DOI: https://arxiv.org/abs/2210.03822[(Margin uncertainty boundary)][m]

19. Neha Gupta, Hari Narasimhan, Ankit Singh Rawat, Wittawat Jitkrittum, Aditya Menon, and Sanjiv Kumar. 2024. Language Model Cascades: Token-Level Uncertainty And Beyond.

@inproceedings{1019493,title = {Language Model Cascades: Token-Level Uncertainty And Beyond},author = {Neha Gupta and Hari Narasimhan and Ankit Singh Rawat and Wittawat Jitkrittum and Aditya Menon and Sanjiv Kumar},year = {2024},URL = {https://openreview.net/forum?id=KgaBScZ4VI},booktitle = {International Conference on Learning Representations}}(Uncertainty summing Doc Sum and Doc Avg, KP sum and KP avg) [n]

20. Mosqueira-Rey, E., Alonso-Ríos, D. & Baamonde-Lozano, A. Integrating Iterative Machine Teaching and Active Learning into the Machine Learning Loop. *Procedia Comput. Sci.* 192, 553–562 (2021) (Random)[o].

21. Ishibashi, H., & Hino, H. (2020). Stopping criterion for active learning based on deterministic generalization bounds. *ArXiv, abs/2005.07402*.[p]

22. Jing X, Goli R, Komatineni K, Alluri S, Hubig N, Min H, et al. Active Learning Pipeline to Identify Candidate Terms for a CDSS Ontology. Stud Health Technol Inform. 2024;316:1338-42.[v]

23. Jing X, Indani A, Hubig NC, Min H, Gong Y, Cimino JJ, et al. A systematic approach to configuring MetaMap for optimal performance. Methods Inf Med. 2022.[x]