

Article

Learning Selective Deferral Policies for Reliable Medical Text Classification

Tahani Albalawi and Amani Alzahrani * 

Department of Computer Science, College of Computing and Information Technology, Shaqra University, Shaqra 11911, Saudi Arabia; s446500717@std.su.edu.sa

* Correspondence: aalzhrani@su.edu.sa

Abstract

Medical text classification is an important task in biomedical natural language processing, but prediction errors remain problematic in high-stakes settings where reliability matters in addition to accuracy. To address this challenge, this paper proposes a learned selective deferral framework for biomedical sentence classification that allows uncertain predictions to be deferred under constrained review budgets. The framework combines a transformer-based classifier with uncertainty estimation, temperature scaling, and a learned deferral policy that predicts the likelihood of model error from multiple signals, including confidence, entropy, calibration-aware features, and Monte Carlo Dropout descriptors. Deferral decisions are applied under fixed budgets to improve the use of limited review capacity. Experiments on the PubMed 200k RCT dataset show that budget-constrained deferral reduces system-level risk. Using PubMedBERT as the primary backbone, deferring 20% of the highest-risk cases reduces system risk from 0.1108 to 0.0360. Compared with a calibrated confidence-threshold baseline, the learned policy provides modest but generally favorable improvements, with statistical significance observed at the 20% budget. Additional experiments across PubMedBERT, BioBERT, and SciBERT suggest that the framework transfers across biomedical transformer backbones, while calibration improves the reliability of confidence estimates and learned policies outperform random deferral.

Keywords: medical text classification; selective deferral; human-in-the-loop AI; uncertainty estimation; calibration; biomedical NLP

1. Introduction

Machine learning models have become a central component of modern natural language processing (NLP), enabling automated analysis of large volumes of text across many domains. In the medical domain, text classification models support tasks such as clinical decision support, literature screening, and biomedical information extraction [1,2]. By processing large collections of scientific and clinical documents, these systems can reduce manual effort for clinicians and researchers. However, despite major advances in deep learning-based language models, automated systems remain prone to errors. In high-stakes settings such as healthcare, incorrect predictions may have serious consequences, making reliability a critical requirement for deployment [3].

Medical text classification typically involves assigning sentences or documents to predefined semantic categories. In clinical trial literature, for example, sentences may be classified into rhetorical roles such as background, objective, methods, results, or conclusions. The PubMed 200k RCT dataset has become a widely used benchmark for this task,



Academic Editor: Hongying Meng

Received: 16 April 2026

Revised: 6 May 2026

Accepted: 8 May 2026

Published: 13 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

providing structured abstracts from biomedical publications for sequential text classification in medical abstracts [4]. Such datasets enable the development and evaluation of models that can capture the discourse structure of biomedical scientific writing.

The availability of large biomedical corpora has also accelerated the development of domain-specific pretrained language models. Transformer-based architectures trained on biomedical text have achieved strong performance across a range of biomedical NLP tasks [5]. Their contextual representations improve performance in tasks such as named entity recognition, relation extraction, and text classification. Nevertheless, even strong biomedical language models may produce uncertain or incorrect predictions when faced with ambiguous or semantically subtle inputs.

Traditional classification systems assume that every input should receive an automatic prediction. In medical applications, however, this assumption is often inappropriate. Forcing a model to predict under uncertainty can lead to high-risk errors, particularly in safety-critical settings. As a result, there is growing recognition that AI systems in healthcare should not only aim for high predictive accuracy, but also identify when their predictions may be unreliable [6].

One promising direction is to incorporate human expertise into the decision process. Rather than treating all inputs equally, a system can defer uncertain cases to human reviewers. This idea aligns with the broader paradigm of human-in-the-loop AI, in which machine learning systems support rather than replace human decision-makers [7]. Such collaboration is especially valuable in medical applications, where expert interpretation remains important for handling difficult and ambiguous cases.

Existing approaches for handling uncertainty often rely on confidence-based mechanisms derived from model outputs. Under this strategy, low-confidence predictions are treated as less reliable. However, neural network confidence scores are frequently miscalibrated and may not reliably indicate whether a prediction is correct [8]. As a result, simple confidence-threshold strategies may fail to identify the cases for which human review would most effectively reduce overall system risk.

More importantly, confidence-based approaches do not explicitly model deferral decisions under realistic constraints on human review capacity. In practice, expert review is limited, and not all uncertain cases can or should be escalated. The core challenge is therefore not only to detect uncertainty, but to determine which cases should be deferred in order to improve overall system-level reliability.

Learning-based deferral approaches address this challenge by predicting which cases are most likely to benefit from human review. Instead of relying on confidence alone, they can leverage richer uncertainty signals to better approximate prediction risk and allocate limited human attention more effectively [9]. This perspective is particularly relevant for biomedical text classification, where reducing harmful automated errors may be more important than maximizing raw predictive coverage alone.

1.1. Medical Text Classification in High-Stakes Settings

Medical text classification is a core task in biomedical natural language processing, where clinical and scientific text is assigned to predefined semantic categories. Deep learning methods, including convolutional neural networks, have shown strong performance in extracting semantic patterns from clinical text [1]. Transformer-based architectures have further advanced this area, with domain-specific pretrained models such as BioBERT improving performance across biomedical NLP tasks [5].

The availability of annotated biomedical benchmarks has been central to this progress. In particular, the PubMed 200k RCT dataset has become a widely used benchmark for sequential sentence classification in medical abstracts [4]. Later work has focused on

improving classification quality and generalization, including approaches that reduce dependence on large labeled datasets [10,11]. Sentence classification in medical abstracts is inherently sequential, since each sentence depends on surrounding context. Hierarchical neural networks have therefore been proposed to capture both sentence-level and document-level structure, improving performance by modeling inter-sentence dependencies [12]. However, despite these advances, most studies emphasize predictive accuracy rather than system-level reliability or the consequences of incorrect predictions in high-stakes settings.

1.2. Uncertainty, Reliability, and Selective Prediction

A major limitation of deep learning systems is the unreliability of their uncertainty estimates. Neural networks often produce overconfident predictions even when incorrect, which is particularly problematic in healthcare applications [8]. Consequently, a growing body of work focuses on uncertainty estimation. Prior studies evaluate uncertainty methods under challenging conditions such as label noise and domain shift [13], and demonstrate their usefulness in improving reliability in clinical text classification tasks [2]. Ensemble-based approaches and biomedical applications, including hierarchical rejection and clinical prediction systems, have also been explored [14–17].

Despite these advances, most uncertainty estimation methods are designed to quantify confidence rather than guide decision-making under limited human review capacity. Reliability, however, is a critical requirement for deploying AI systems in healthcare, where managing prediction risk is as important as achieving high accuracy [3,6].

Selective prediction addresses this issue by allowing models to abstain from making predictions when uncertainty is high. Early work on reject-option classification established theoretical foundations for balancing prediction and rejection [18,19], while more recent approaches extend these ideas to deep learning. For example, SelectiveNet integrates rejection into end-to-end training [20], and other methods explore distribution-free selective classification and probabilistic reject-option formulations [21,22]. Alternative abstention strategies, such as game-theoretic approaches like Deep Gamblers, have also been proposed [23].

However, many of these methods remain confidence-driven, architecture-specific, or designed for general settings, without explicitly addressing biomedical sentence classification under constrained human review budgets.

1.3. Learning-Based Deferral and Human-in-the-Loop AI

Learning-based deferral is the most closely related paradigm to this work. Rather than relying on fixed confidence thresholds, learning-to-defer approaches train models to decide whether a prediction should be automated or deferred for review [9]. Related research includes constrained optimization for reject-option decisions [24], training-dynamics-based uncertainty identification [25], and deferral strategies in sequential models such as LSTMs [26].

Human-in-the-loop AI extends this paradigm by incorporating expert judgment into the decision process, which is particularly important in healthcare settings. Prior work highlights the role of human involvement in improving system safety and robustness [7], while more recent studies investigate collaborative frameworks that defer uncertain cases to human experts [27].

Nevertheless, many existing approaches rely on rule-based or loosely defined deferral mechanisms, limiting their ability to optimally allocate limited human review resources. This gap motivates the need for learned, uncertainty-aware deferral policies that explicitly prioritize cases based on predicted risk.

1.4. Research Gap

Despite substantial progress in medical text classification, uncertainty estimation, selective prediction, and human-in-the-loop learning, an important gap remains. Existing work often improves predictive accuracy, measures uncertainty, or supports abstention, but does not fully integrate these components into a single framework for decision-making under constrained review capacity. Selective prediction methods frequently rely on simple confidence-based rules, while broader human-in-the-loop frameworks emphasize collaboration without providing a learned mechanism for prioritizing which cases should be reviewed.

Accordingly, the gap addressed in this work is not the absence of uncertainty estimation or selective prediction in general, but the lack of a practically integrated framework for biomedical sentence classification that combines calibrated uncertainty signals, learned error-aware deferral, and explicit budget-constrained evaluation at the system level.

To better position the contribution of this work relative to prior research, Table 1 compares the proposed framework with representative learning-to-defer and selective prediction approaches. Existing methods such as those of Hemmer et al. focus on learning-based deferral but are primarily evaluated in general machine learning settings and do not explicitly integrate calibration or rich uncertainty representations. Similarly, approaches such as SelectiveNet and Deep Gamblers incorporate rejection mechanisms within deep models but do not explicitly consider constrained human review budgets or calibration-aware decision-making.

Table 1. Comparison with Prior Learning-to-Defer and Selective Prediction Approaches.

Method	Domain	Budget Constraints	Calibration Integration	Uncertainty Features	Learning-Based Deferral	Biomedical Focus
Hemmer et al. [9]	General ML	Limited	Not explicit	Limited (confidence-based)	Yes	No
SelectiveNet [20]	General DL	No	No	Implicit	Yes (end-to-end)	No
Deep Gamblers [23]	General DL	No	No	Implicit	Yes	No
Confidence Thresholding [18,22]	General	Yes	optional	Confidence only	No (rule-based)	No
Proposed Framework	Biomedical NLP	Explicit fixed budgets (5–20%)	Temperature scaling	Rich (confidence, entropy, MC Dropout, calibration-aware, engineered features)	Yes (error-aware learned policy)	Yes

In contrast, the proposed framework introduces a practically integrated pipeline tailored to biomedical text classification, combining (i) explicit budget-constrained deferral, (ii) post hoc calibration via temperature scaling, and (iii) a rich set of uncertainty features, including deterministic, calibration-aware, and Monte Carlo Dropout-based descriptors. Furthermore, unlike confidence-threshold baselines, the deferral decision is formulated as a learned error prediction problem, enabling more effective prioritization of high-risk cases under limited review capacity. This combination distinguishes the present work from prior approaches and addresses a gap in reliability-oriented biomedical NLP systems.

1.5. Aim of the Study and Main Contributions

Motivated by this problem, the present work proposes a learned selective deferral framework for medical text classification using the PubMed 200k RCT dataset. The framework combines transformer-based prediction, uncertainty estimation, calibration, and a

learned error-aware deferral policy under fixed review budgets. The goal is to improve system-level reliability by identifying which cases should remain automated and which should be deferred for review.

The main contributions of this study are as follows:

1. A learned selective deferral framework for reliable biomedical sentence classification is proposed.
2. The framework integrates transformer-based prediction, calibration, deterministic uncertainty features, and Monte Carlo Dropout-based uncertainty descriptors within a unified pipeline.
3. A learned error-aware deferral policy is developed under fixed review budgets in order to prioritize cases for review more effectively than simple confidence-threshold ranking.
4. The framework is evaluated on the PubMed 200k RCT benchmark using PubMedBERT as the primary backbone.
5. Additional analyses are provided through comparison with confidence-threshold and random baselines, bootstrap significance testing, cross-backbone evaluation, cost-aware analysis, imperfect-review sensitivity analysis, ablation study, and feature-group permutation importance analysis.

2. Materials and Methods

2.1. Dataset

This study uses the PubMed 200k RCT dataset [4] for all experiments in the proposed framework. In the implementation, the dataset is loaded from the Hugging Face repository under the identifier `pietrolesci/pubmed-200k-RCT`, and the official train, validation, and test splits are used directly. Each example consists of a sentence extracted from a randomized controlled trial abstract together with its sentence-level label.

Label metadata are read directly from the dataset object. The implementation constructs `id2label` and `label2id` mappings from the embedded label names, and the total number of labels is derived from these mappings. The dataset is then enriched with additional fields including cleaned sentence text, numeric label identifier, and label name.

2.2. Data Preprocessing and Tokenization

Preprocessing is intentionally kept light in order to preserve the original biomedical text for transformer-based modeling. Each sentence is converted to string format, line breaks and tab characters are replaced with spaces, repeated spaces are normalized, and leading and trailing whitespace are removed. No stemming, aggressive normalization, or domain-specific rewriting is applied.

Sentences are tokenized using the tokenizer associated with each backbone, with a maximum sequence length of 128 tokens. Truncation is enabled, while fixed-length padding is disabled so that padding can be handled dynamically during batching. After tokenization, each example is assigned a target field equal to its `label_id`. The tokenized dataset retains both metadata and model inputs, including `uid`, `text`, `clean_text`, `label_id`, `label_name`, `target`, `input_ids`, and `attention_mask`. If present, `token_type_ids` are also retained.

2.3. Dataloaders and Base Classifiers

The tokenized datasets are converted to PyTorch format using the tensor fields required for model training. Dynamic padding is implemented with `DataCollatorWithPadding`, using the longest sequence in each batch. A custom collate function renames the target field to `labels`, matching the interface expected by Hugging Face sequence classification models. The implementation was conducted using Python 3.12.13, PyTorch 2.10.0+cu128, Hugging Face Transformers 5.0.0, Hugging Face Datasets 4.0.0, and scikit-learn 1.6.1.

The training dataloader uses a batch size of 16 with shuffling, whereas the validation and test dataloaders use a batch size of 32 without shuffling. All dataloaders use `num_workers = 2`, and `pin_memory` is enabled when CUDA is available.

The primary base classifier is implemented using PubMedBERT through `AutoModelForSequenceClassification`. For comparative backbone experiments, the same sequence classification pipeline is also instantiated with BioBERT [5] and SciBERT [28].

2.4. Training Configuration and Procedure

The shared training configuration uses a maximum sequence length of 128, training batch size of 16, evaluation batch size of 32, learning rate of 2×10^{-5} , weight decay of 0.01, and training for 5 epochs. The dropout probability is set to 0.1, the warmup ratio to 0.10, the maximum gradient norm to 1.0, and gradient accumulation steps to 1. The same configuration also defines Monte Carlo Dropout passes, Expected Calibration Error (ECE) binning, bootstrap iterations, and deferral budgets of 5%, 10%, and 20%.

PubMedBERT is trained as the primary backbone using the AdamW optimizer with learning rate 2×10^{-5} and weight decay 0.01. A linear warmup scheduler is implemented through `get_linear_schedule_with_warmup`, with the number of warmup steps set to 10% of the total training steps. During training, gradients are clipped to a maximum norm of 1.0. Validation is performed at the end of each epoch, and the best model is selected according to validation macro F1-score. The same overall training procedure is then applied to BioBERT and SciBERT for comparative backbone evaluation.

2.5. Prediction Output Extraction, Uncertainty Estimation, and Calibration

After training, the best checkpoint is loaded and used to extract per-example prediction outputs on the validation and test splits. For each example, the implementation stores sentence metadata, true label ID and name, predicted label ID and name, correctness indicator, logits, and class probabilities. These outputs form the basis for uncertainty estimation, calibration, and later deferral modeling.

Three basic uncertainty features are computed directly from the predicted probability distribution of the base classifier [2,13]: confidence, defined as the maximum class probability; entropy, computed as the negative sum of the probability distribution multiplied by its logarithm, with numerical clipping for stability; and margin, defined as the difference between the highest and second-highest class probabilities.

To capture epistemic uncertainty, Monte Carlo Dropout is applied during inference [29]. Repeated stochastic forward passes are performed, and multiple uncertainty descriptors are derived from the resulting predictive distributions. These include MC confidence, predictive entropy, expected entropy, mutual information, and variance-based measures such as mean variance and maximum variance.

Post hoc calibration is performed using temperature scaling [30]. A learnable temperature parameter is fitted on the validation logits and validation labels using a `TemperatureScaler` module. Optimization is carried out with LBFGS using a learning rate of 0.01, a maximum of 200 iterations, and cross-entropy loss as the objective. Once the temperature is learned on the validation set, it is applied to both validation and test logits. Calibrated logits, calibrated probabilities, calibrated predicted labels, calibrated correctness, and calibrated confidence are then stored.

2.6. Deferral Feature Construction and Learned Deferral Policy

A dedicated deferral training dataset is constructed from validation-set outputs, while the corresponding deferral evaluation dataset is constructed from test-set outputs. A binary error label is defined as $1 - \text{correct}$, so that the deferral policy explicitly learns to identify likely prediction failures.

To support consistent comparison across backbones, deferral feature construction follows a unified uncertainty feature family for PubMedBERT, BioBERT, and SciBERT. The final feature set combines deterministic uncertainty measures, calibration-aware signals, Monte Carlo Dropout-based descriptors, engineered confidence-derived features, and class-aware indicators. Specifically, the feature representation includes confidence, entropy, margin, calibrated confidence, MC confidence, predictive entropy, expected entropy, mutual information, mean variance, and maximum variance, together with engineered features such as top-two confidence, relative margin, confidence gap, a mid-confidence region flag, and one-hot encoded predicted-class indicators.

The proposed deferral policy is formulated as a binary error-prediction problem. It is trained on validation-split outputs using a gradient-boosted decision model, specifically HistGradientBoostingClassifier, to estimate the probability that the base classifier's prediction is incorrect. Because prediction errors are less frequent than correct predictions, class-balanced sample weights are applied during training. The raw output of the deferral model is the predicted probability of error, which serves as the main learned risk signal. Engineered features such as mid-confidence indicators and confidence gap are included to explicitly capture regions where predictions are ambiguous and confidence alone may be insufficient. These features are designed to complement standard uncertainty measures and improve the identification of high-risk cases.

2.7. Deferral Scoring, Budgets, and Baselines

The learned error score is blended with a calibration-aware uncertainty signal defined as $1 - \text{calibrated confidence}$ in order to improve ranking stability under fixed budgets. For a candidate blending coefficient α , the blended score is defined as:

$$\text{Blended score} = \alpha \cdot \text{Model score} + (1 - \alpha) \cdot (1 - \text{Calibrated confidence})$$

A small set of candidate α values is evaluated on the validation split, and the final coefficient is selected using a budget-weighted validation objective over the target budgets of 0.05, 0.10, and 0.20.

Once the final risk score is computed, deferral is applied under fixed review budgets of 5%, 10%, and 20% [9,24]. For a given budget, the test set is sorted in descending order of `deferral_risk_score`. The highest-risk examples are marked as deferred until the budget is exhausted, while the remaining examples are treated as automated predictions. The blending coefficient α controls the trade-off between the learned error probability and the calibration-based uncertainty signal. Its value is selected using a budget-weighted validation objective, in which candidate α values are evaluated across the target deferral budgets (0.05, 0.10, and 0.20). The selected value maximizes overall selective risk reduction across these budgets. In practice, the framework is relatively stable for moderate α values, while extreme settings (e.g., α approaching 0 or 1) effectively reduce the method to confidence-based ranking or purely learned ranking, respectively.

Two baseline methods are used for comparison. The confidence-threshold baseline defers the lowest-confidence examples first, using calibrated confidence as the ranking signal [18,22]. The random baseline defers examples uniformly at random under the same fixed budgets. This random process is repeated 100 times for each budget, and the mean, standard deviation, minimum, and maximum system risk are computed across random runs.

2.8. Evaluation Metrics and Statistical Analysis

The framework is evaluated at both the classification and selective-prediction levels. For the base classifier, training and validation monitoring includes accuracy, macro preci-

sion, macro recall, macro F1, weighted precision, weighted recall, and weighted F1. For the learned deferral classifier, binary evaluation metrics include accuracy, precision, recall, F1-score, ROC-AUC, and average precision.

At the selective-prediction level, the code computes coverage, deferral rate, selective risk, and system risk [18,22]. Coverage is the fraction of inputs that remain automated. Deferral rate is the fraction of inputs sent for review. Selective risk is the error rate among automated predictions only. System risk is defined as the total number of automated errors divided by the full test set size, under an idealized assumption that deferred cases are correctly resolved by human review. In the present study, human review is simulated using ground-truth labels rather than real expert annotations. It is important to note that this definition assumes that all deferred cases are correctly resolved, representing an idealized best-case scenario. Therefore, system risk should be interpreted as a lower-bound estimate, and the observed improvements correspond to an upper bound on achievable performance under perfect review conditions.

To assess the sensitivity of this assumption, we additionally simulate imperfect human review by assuming reviewer accuracy levels of 95% and 90%. Under this setting, adjusted system risk includes both the remaining automated errors and the expected residual errors among deferred cases. This analysis provides an estimate of how the system may behave when human review is beneficial but not perfectly accurate.

To quantify uncertainty in selective performance estimates, the implementation computes stratified bootstrap confidence intervals for the learned policy and the confidence-threshold baseline. Bootstrap analysis is performed on the test set using 10,000 runs for budgets 0.05, 0.10, and 0.20 [31]. Sampling is stratified by the binary error label. For each bootstrap sample, selective risk is recomputed for both the learned policy and the threshold baseline. The procedure reports mean selective risk together with the 2.5th and 97.5th percentiles as empirical 95% confidence intervals.

Figure 1 illustrates the overall architecture of the proposed selective deferral framework. The pipeline begins with biomedical text input from the PubMed 200k RCT dataset, which is processed by a transformer-based classifier (PubMedBERT) to generate class probabilities. These predictions are then calibrated using temperature scaling to improve the reliability of confidence estimates.

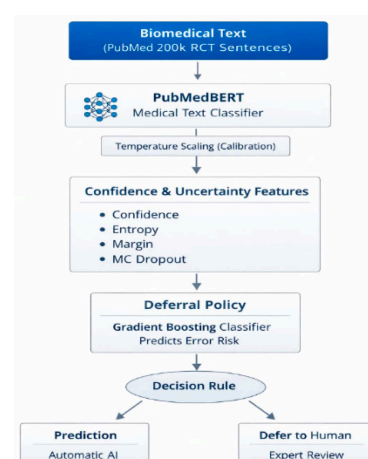


Figure 1. Architecture of the proposed selective deferral framework for medical text classification.

From the calibrated outputs, multiple uncertainty features are extracted, including confidence, entropy, margin, and Monte Carlo Dropout-based descriptors. These features are used to construct a deferral feature space that captures both predictive uncertainty and potential model error.

A gradient-boosting-based deferral policy is then trained to estimate the probability of prediction error. Based on this learned risk score, a decision rule determines whether a given input should be handled automatically or deferred to human experts. This design enables the system to prioritize high-risk cases for review, thereby improving overall reliability under constrained review budgets.

3. Results

3.1. Base Model Performance

We first evaluate the PubMedBERT classifier on the PubMed 200k RCT test set. The model achieves an overall accuracy of 0.889, with a weighted F1-score of 0.889 and a macro F1-score of 0.839, indicating strong overall performance for multi-class biomedical sentence classification.

Performance, however, is not uniform across classes. The model performs best on METHODS and RESULTS, while lower performance is observed for BACKGROUND and OBJECTIVE. This pattern suggests that some categories are inherently more difficult because of higher semantic ambiguity and weaker discriminative boundaries.

The confusion matrix further shows systematic misclassifications between semantically similar classes, particularly BACKGROUND vs. OBJECTIVE and RESULTS vs. CONCLUSIONS. These errors are therefore not random, but concentrated in linguistically overlapping categories. This observation motivates the need for uncertainty-aware decision mechanisms rather than purely accuracy-driven automation.

Figure 2 presents the confusion matrix of PubMedBERT across the five rhetorical categories. The model achieves strong performance overall, as indicated by the high values along the diagonal, particularly for METHODS and RESULTS, which are more structurally distinct and easier to classify.

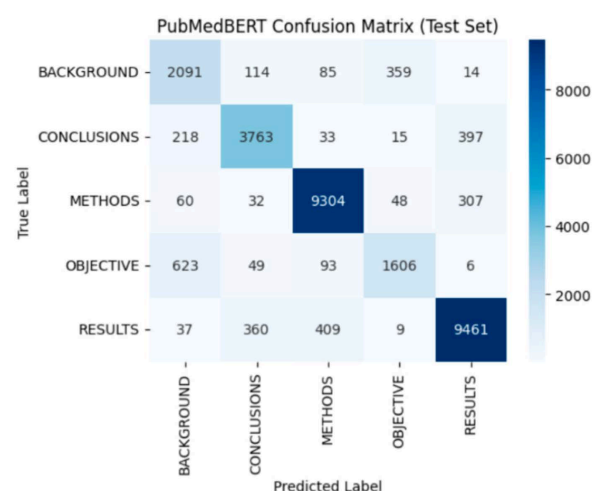


Figure 2. Confusion matrix of PubMedBERT on the PubMed 200k RCT test set, showing classification performance across five rhetorical categories.

However, several systematic misclassification patterns can be observed. Notably, there is significant confusion between BACKGROUND and OBJECTIVE, as well as between RESULTS and CONCLUSIONS. These errors reflect the semantic overlap between these categories, where distinctions are often subtle and context-dependent. For example, background statements may resemble objectives in tone, while results and conclusions frequently share similar phrasing and interpretative language.

These findings indicate that classification errors are not random but concentrated in linguistically ambiguous classes. This observation highlights the limitation of relying solely

on predictive accuracy and motivates the need for uncertainty-aware mechanisms, such as the proposed selective deferral framework, to identify and defer high-risk predictions.

3.2. Calibration and Reliability

To improve the reliability of prediction confidence, temperature scaling is applied as a post hoc calibration method. On the test set, the Expected Calibration Error (ECE) decreases from 0.0432 to 0.0138, corresponding to an improvement of 0.0294.

This reduction indicates that predicted probabilities become better aligned with empirical outcome frequencies, thereby reducing overconfidence. Because deferral decisions rely on uncertainty estimates, this improvement in calibration is important for downstream decision quality. Figure 3 illustrates the reliability diagram of PubMedBERT before and after temperature scaling. Prior to calibration, the model exhibits noticeable miscalibration, with predictions deviating from the ideal diagonal, particularly in mid- to high-confidence regions where overconfidence is evident. After applying temperature scaling, the calibrated curve shifts closer to the perfect calibration line across most confidence bins. This improvement indicates that predicted probabilities better reflect true outcome likelihoods.

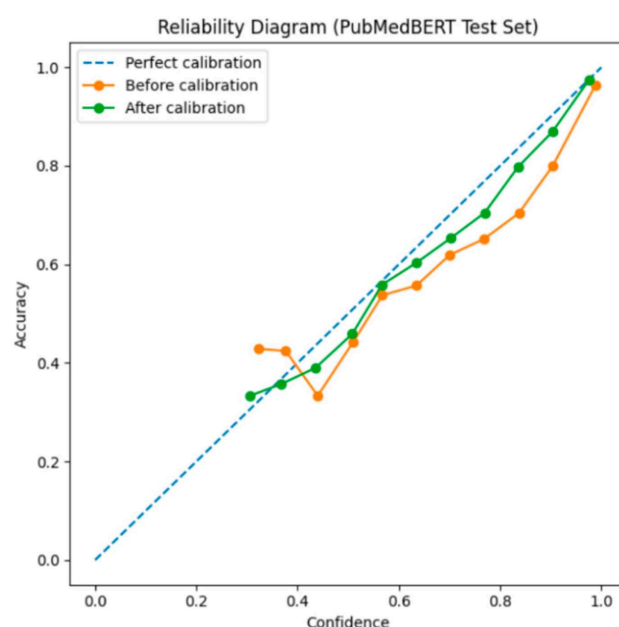


Figure 3. Reliability diagram showing the calibration performance of PubMedBERT on the PubMed 200k RCT test set before and after temperature scaling.

Importantly, improved calibration enhances the quality of uncertainty estimates used by the deferral mechanism. Since deferral decisions rely on identifying high-risk predictions, better alignment between confidence and accuracy leads to more effective prioritization of cases for human review.

3.3. Selective Prediction Performance

The performance of the learned deferral policy is summarized in Table 2.

Table 2. Selective prediction performance of the learned deferral policy (PubMedBERT).

Budget	Coverage	Selective Risk	System Risk
0.00	1.0000	0.1108	0.1108
0.05	0.9500	0.0889	0.0845
0.10	0.9000	0.0719	0.0647
0.20	0.8000	0.0450	0.0360

As shown in Table 2, increasing the deferral budget leads to a consistent reduction in both selective risk and system risk. At a 20% deferral budget, selective risk decreases to 0.0450, while system risk decreases from 0.1108 to 0.0360, corresponding to an approximately 67% reduction in system risk.

Importantly, this reduction reflects a change in decision behavior rather than an improvement in the underlying classifier. By deferring high-risk cases, the system removes a disproportionate number of likely errors from automated prediction, thereby improving reliability while maintaining substantial automated coverage.

Because the main system risk calculation assumes that deferred cases are correctly resolved by human review, we further conduct a sensitivity analysis to examine how the results change when the human reviewer is imperfect. In this analysis, reviewer accuracy is simulated at 100%, 95%, and 90%. The imperfect-review system risk is computed by adding the expected residual human errors among deferred cases to the remaining automated errors. This analysis is intended to provide a more realistic estimate of system behavior when human review is helpful but not error-free.

As shown in Table 3, system risk increases as reviewer accuracy decreases, which confirms that the original system risk values represent a best-case estimate. However, the deferral framework remains beneficial even under imperfect review assumptions. At the 20% budget, system risk increases from 0.0360 under perfect review to 0.0460 with 95% reviewer accuracy and 0.0560 with 90% reviewer accuracy. These values remain substantially lower than the full-automation risk of 0.1108, indicating that selective deferral can still reduce system-level risk even when human reviewers are not perfectly accurate. Nevertheless, this analysis highlights that real-world deployment should account for reviewer quality, disagreement, and operational review conditions.

Table 3. Sensitivity analysis of system risk under imperfect human review accuracy (PubMedBERT).

Budget	100% Reviewer	95% Reviewer	90% Reviewer
0.05	0.0845	0.0870	0.0895
0.10	0.0647	0.0697	0.0747
0.20	0.0360	0.0460	0.0560

Note. Values represent adjusted system risk under simulated human reviewer accuracy levels of 100%, 95%, and 90%.

3.4. Comparison with Confidence-Based Baseline

We next compare the learned deferral policy with a confidence-threshold baseline in terms of selective risk, as shown in Table 4.

The learned policy achieves lower selective risk than the confidence-threshold baseline across all evaluated budgets for PubMedBERT. Although the absolute improvements remain modest, their consistency suggests that combining multiple uncertainty signals can improve deferral ranking beyond simple confidence-based thresholding.

Table 4. Comparison between the learned deferral policy and the confidence-threshold baseline in terms of selective risk (PubMedBERT).

Budget	Learned Selective Risk	Threshold Selective Risk	Improvement
0.05	0.0889	0.0900	+0.0010
0.10	0.0719	0.0730	+0.0011
0.20	0.0450	0.0467	+0.0018

3.5. Statistical Significance, Cost-Aware Analysis, and Runtime Benchmark

To assess robustness, stratified bootstrap evaluation is performed with 10,000 resamples. Results are summarized in Table 5.

Table 5. Bootstrap confidence intervals and statistical significance analysis for selective risk (PubMedBERT).

Budget	Learned Risk	Threshold Risk	Δ Selective Risk	95% CI	Significant
0.05	0.088966	0.090041	−0.001075	[−0.002213, 0.000071]	No
0.10	0.071885	0.073011	−0.001126	[−0.002336, 0.000038]	No
0.20	0.044936	0.046781	−0.001845	[−0.003094, −0.000593]	Yes

The learned policy achieves statistically significant improvement only at the 20% deferral budget. At the 5% and 10% budgets, it consistently yields lower selective risk than the confidence-threshold baseline; however, these differences are not statistically significant, as the corresponding confidence intervals include zero. Therefore, under more restrictive and practically realistic review budgets, the learned policy should be interpreted as providing numerical but not statistically significant improvements, with statistically significant gains observed only at the 20% budget.

To complement the main reliability analysis, a cost-aware evaluation is conducted for PubMedBERT under three scenarios (low, medium, and high error cost), while assigning a fixed review cost to deferred cases, as shown in Table 6. The total normalized cost is defined as the sum of the review cost for deferred samples and the error cost for remaining automated mistakes, normalized by the test set size.

Table 6. Normalized cost comparison under different cost scenarios (PubMedBERT).

Scenario	Budget	Learned	Threshold	Random	Reduction vs. Threshold
Low error cost	0.05	0.218086	0.220934	0.260435	0.002848
Low error cost	0.10	0.229987	0.231479	0.299359	0.001492
Low error cost	0.20	0.271183	0.274777	0.377344	0.003594
Medium error cost	0.05	0.470247	0.477368	0.576069	0.007120
Medium error cost	0.10	0.424982	0.428712	0.598413	0.003730
Medium error cost	0.20	0.377988	0.386973	0.643339	0.008985
High error cost	0.05	0.890516	0.904757	1.102126	0.014241
High error cost	0.10	0.749975	0.757434	1.096837	0.007459
High error cost	0.20	0.555996	0.573967	1.086665	0.017970

Across all evaluated budgets and cost scenarios, the learned deferral policy achieves the lowest normalized cost compared with both the confidence-threshold baseline and the random baseline.

To further quantify the practical computational implications of the proposed approach, we conducted a full runtime benchmark on the complete PubMedBERT test set of 29,493 samples using an NVIDIA A100-SXM4-80GB GPU (NVIDIA Corporation, Santa Clara, CA, USA). As shown in Table 7, the benchmark compares three deployment configurations:

calibrated confidence-thresholding, a learned policy without MC Dropout descriptors, and the full learned policy with MC Dropout descriptors. Confidence-thresholding is the lowest-cost configuration because it only requires calibrated confidence scores and ranking. The learned policy without MC Dropout adds a lightweight post hoc scoring step, while the full uncertainty-aware configuration introduces the largest overhead through MC Dropout feature extraction, which requires 15 stochastic forward passes. Therefore, the proposed framework should be interpreted as deployment-dependent: confidence-thresholding is preferable for latency-sensitive settings, the non-MC learned policy provides a lightweight learned alternative, and the full MC Dropout-based policy is more suitable for offline or high-error-cost biomedical text processing.

Table 7. Runtime benchmark for PubMedBERT deferral configurations.

Configuration	MC Passes	Scoring/Ranking (s)	MC Extraction (s)
Confidence threshold	0	0.0007	0.0000
Learned policy without MC	0	0.5347	0.0000
Full learned policy with MC	15	0.3906	434.8231

Note. The benchmark was conducted on the complete PubMedBERT test set of 29,493 samples using an NVIDIA A100-SXM4-80GB GPU. One-time training costs were 1.2088 s for the learned policy without MC Dropout and 1.0955 s for the full learned policy. MC Dropout was measured on the full test set using 15 stochastic forward passes. The benchmark characterizes computational overhead and does not alter the reported predictive, calibration, or deferral results.

These runtime results help explain why the full MC Dropout-based learned policy is not intended to replace confidence-thresholding in all settings. Rather, the proposed framework provides multiple deployment configurations with different cost–reliability trade-offs. In resource-constrained or real-time environments, calibrated confidence-thresholding remains preferable because of its simplicity and minimal computational overhead. The learned policy without MC Dropout provides a lightweight post hoc alternative when a learned risk-ranking model is desired without the cost of stochastic inference. In offline biomedical text processing or high-error-cost settings, the full MC Dropout-based configuration may be more acceptable because review prioritization can be computed before human assessment. Therefore, the runtime benchmark supports a deployment-dependent interpretation of the proposed framework rather than a universal replacement of confidence-thresholding.

3.6. Robustness Across Backbones

We evaluate the framework across three biomedical transformer backbones: PubMedBERT, BioBERT, and SciBERT. Results are summarized in Table 8.

BioBERT achieves the lowest learned risk across all evaluated budgets in this experimental setting. SciBERT ranks second at the 10% and 20% budgets, while PubMedBERT yields the highest learned risk among the three backbones in this comparison. Despite these backbone-dependent differences, all three learned policies outperform the random baseline by clear margins across all evaluated budgets.

Table 8. Cross-backbone comparison of learned deferral performance under a unified uncertainty feature family.

Model	Budget	Learned Risk
BioBERT	0.05	0.083952
SciBERT	0.05	0.088631

Table 8. *Cont.*

Model	Budget	Learned Risk
PubMedBERT	0.05	0.088940
BioBERT	0.10	0.064185
SciBERT	0.10	0.067813
PubMedBERT	0.10	0.071918
BioBERT	0.20	0.035873
SciBERT	0.20	0.038891
PubMedBERT	0.20	0.044967

3.7. Error Analysis and Ablation Study

Error analysis provides additional insight into the behavior of the learned deferral policy. At a 10% deferral budget, deferred samples exhibit a high error rate, indicating that the policy concentrates difficult cases in the deferred subset, while automated predictions achieve a substantially lower error rate than the full-coverage classifier.

Deferred samples are concentrated in ambiguous classes such as BACKGROUND and OBJECTIVE, whereas clearer categories such as METHODS and RESULTS are more likely to remain automated. The persistence of some high-confidence errors among automated predictions further highlights the limitations of confidence-based heuristics and supports the use of richer uncertainty-aware deferral strategies.

To better understand the contribution of different components in the proposed framework, two complementary analyses are conducted on PubMedBERT: an ablation study and a feature-group permutation importance analysis. The ablation study evaluates several feature configurations under fixed deferral budgets, as shown in Table 9, while the permutation analysis measures the sensitivity of system risk and selective risk to each feature group.

Table 9. Ablation study of the learned deferral policy on PubMedBERT.

Variant	Features	5% System Risk	10% System Risk	20% System Risk
Deterministic only	confidence, entropy, margin	0.085681	0.065439	0.037297
Deterministic + calibration	deterministic features + calibration-aware signals	0.085715	0.066151	0.037127
Deterministic + MC Dropout	deterministic features + epistemic uncertainty features	0.084664	0.064829	0.036754
Full without class-aware	full feature set without predicted-class indicators	0.084732	0.064931	0.036754
Full without blending	full feature set, model score only	0.084562	0.064592	0.035975
Full proposed	full feature set + blended score	0.084359	0.064592	0.036042

Lower system risk is better. The ablation results show that deterministic uncertainty features alone provide a useful but relatively limited deferral signal, whereas Monte Carlo Dropout-based features contribute a clearer improvement across all budgets. The strongest

overall performance is obtained by combining multiple complementary uncertainty and error-aware signals.

To further examine whether the learned deferral policy benefits from richer uncertainty representations beyond confidence-based signals, a feature-group permutation importance analysis was performed. In this analysis, each feature group was permuted while keeping the remaining features unchanged, and the resulting increase in system risk and selective risk was measured. Larger increases indicate that the corresponding feature group contributes more strongly to the learned risk-ranking behavior.

As shown in Table 10, MC Dropout-based features contribute the largest increase in both system risk and selective risk when permuted, with a mean increase of 0.032573 in system risk and a maximum increase of 0.051043. This indicates that epistemic uncertainty descriptors provide substantial information for identifying high-risk predictions beyond deterministic confidence-based features. In contrast, class-aware, deterministic uncertainty, engineered confidence, and calibration-aware features show smaller but positive contributions. These results support the use of a multi-signal deferral policy and help justify the additional computational overhead associated with MC Dropout in settings where offline processing or high error costs make richer uncertainty estimation acceptable.

Table 10. Feature-group permutation importance analysis for the learned deferral policy on PubMedBERT.

Feature Group	Features (n)	Mean Δ System Risk	Max Δ System Risk	Mean Δ Selective Risk
MC Dropout	6	0.032573	0.051043	0.038241
Class-aware	5	0.001014	0.001163	0.001152
Deterministic uncertainty	3	0.000414	0.000432	0.000471
Engineered confidence	2	0.000405	0.000822	0.000490
Calibration-aware	3	0.000096	0.000229	0.000099

Note. Features (n) indicates the number of features included in each feature group.

4. Discussion

The findings of this study show that reliability in medical text classification can be improved substantially through selective deferral under constrained review budgets. The most important result is that the largest gain comes from the deferral mechanism itself. For PubMedBERT, deferring 20% of the highest-risk cases reduces system risk from 0.1108 to 0.0360, demonstrating that a meaningful portion of harmful automated errors can be removed while maintaining substantial automated coverage. While the learned policy consistently improves selective risk relative to the confidence-threshold baseline, these gains remain modest in absolute terms. This indicates that the primary benefit arises from the deferral mechanism itself, while the learned policy provides incremental refinement over strong baseline strategies. This highlights a trade-off between model complexity and performance gain, suggesting that the usefulness of the learned policy depends on the operational context, particularly the cost of errors and the availability of review resources.

From the perspective of prior work, these findings are consistent with research on selective prediction, reject-option learning, and human–AI collaboration, which shows that abstention improves reliability under uncertainty. The present study extends this line of work by integrating transformer-based biomedical classification, calibration, Monte Carlo Dropout-based uncertainty estimation, class-aware engineered features, and explicit budget-constrained evaluation within a unified pipeline. This addresses a practical gap in prior studies that often consider these components in isolation.

The comparison with the confidence-threshold baseline is particularly informative. Although the learned policy consistently outperforms calibrated confidence-based ranking, the absolute gains remain modest. Importantly, this should not be interpreted as indicating that the learned policy universally replaces confidence thresholding. Instead, the framework is intended for risk-aware decision-making under constrained review budgets, where even modest improvements in prioritizing high-risk cases can be meaningful. The results also indicate that calibrated confidence is already a strong ranking signal, making large additional gains difficult to achieve. At the same time, bootstrap analysis shows that the learned policy becomes more advantageous as review capacity increases, with statistically significant improvement observed at the 20% budget.

The backbone analysis further demonstrates that the proposed framework generalizes across PubMedBERT, BioBERT, and SciBERT under a shared uncertainty feature representation. However, the magnitude of improvement varies across models, suggesting that deferral performance depends on both the uncertainty features and the calibration behavior of the underlying classifier. From a broader perspective, the framework can also be interpreted within a cost-sensitive decision-making paradigm, where trade-offs between automation, human review, and prediction error are explicitly considered.

The ablation and feature-group permutation importance analyses indicate that the observed gains arise from multiple complementary signals, with Monte Carlo Dropout-based descriptors providing the strongest contribution. When MC Dropout features are permuted, system risk increases substantially more than when other feature groups are permuted, suggesting that epistemic uncertainty provides useful information not captured by confidence alone. At the same time, deterministic uncertainty measures, calibration-aware features, engineered confidence features, and class-aware indicators show smaller but positive contributions. The use of a gradient-boosting model enables the capture of nonlinear interactions among these heterogeneous features. These findings also clarify the computational trade-off reported in the runtime benchmark: MC Dropout introduces the largest overhead, but it also provides the strongest contribution to risk ranking. However, alternative formulations, including simpler linear models or end-to-end approaches, may offer different trade-offs between interpretability and performance and remain important directions for future research.

Several limitations should be acknowledged. First, human review is simulated using ground-truth labels rather than real expert annotations. Although the sensitivity analysis under 95% and 90% reviewer accuracy shows that the proposed framework remains beneficial when human review is imperfect, this analysis remains simulated and does not capture real reviewer disagreement, fatigue, latency, domain expertise, or actual human decision costs. Second, the experiments are limited to a single benchmark dataset (PubMed 200k RCT), which restricts claims about generalizability across biomedical domains. Third, although the learned policy provides consistent improvements over strong baselines, these gains remain modest and are dependent on the available review budget. Fourth, the calibration analysis indicates that the mid-confidence region (approximately 0.4–0.6) remains challenging even after temperature scaling. While the learned policy partially mitigates this issue by combining multiple uncertainty signals, improving reliability in this region remains an open challenge. Finally, the use of the validation set for both temperature scaling and deferral model training introduces a potential source of dependency, and future work should consider stricter designs such as separate validation splits or nested validation.

Future work should explore real human-in-the-loop evaluation, cost-sensitive learning, adaptive budget allocation, and evaluation across diverse biomedical NLP tasks and domains to further improve reliability and deployment readiness.

5. Conclusions

This work presents a learned selective deferral framework for improving reliability in medical text classification under constrained review budgets. Instead of enforcing fully automated predictions for every input, the framework allows uncertain or high-risk cases to be deferred, supporting a more reliability-oriented human–AI collaboration setting. Using the PubMed 200k RCT dataset, the study demonstrates that strong biomedical language models can still produce non-negligible errors despite high predictive accuracy. To address this limitation, the proposed framework combines transformer-based classification with uncertainty estimation, calibration, and a learned deferral policy leveraging confidence, entropy, calibration-aware features, and Monte Carlo Dropout-based uncertainty descriptors to estimate prediction risk.

The empirical results show that budget-constrained deferral substantially improves system-level reliability. In the primary PubMedBERT analysis, deferring 20% of the highest-risk cases reduces system risk from 0.1108 to 0.0360. Relative to the confidence-threshold baseline, the learned deferral policy provides modest but generally favorable additional gains, with statistically significant improvement observed only at the 20% budget. Overall, these findings support the usefulness of learned uncertainty-aware deferral for improving reliability in biomedical text classification under constrained review settings. From a practical perspective, the framework should be considered deployment-dependent: confidence-thresholding remains a strong and efficient baseline for latency-sensitive settings, while the learned policy without MC Dropout offers a lightweight alternative. The full configuration, including MC Dropout-based features, is more suitable for offline or high-risk scenarios where richer uncertainty estimation justifies additional computational cost.

Several directions can extend this work. First, future studies should incorporate real human-in-the-loop evaluation, including expert disagreement, annotation latency, and review cost, to better reflect clinical deployment conditions. Second, the framework should be evaluated on additional biomedical and clinical datasets (e.g., clinical notes or multilingual corpora) to assess generalizability under domain shift. Third, more advanced uncertainty estimation techniques, such as deep ensembles or Bayesian transformer variants, may further improve risk estimation and deferral quality. Fourth, adaptive or dynamic budget allocation strategies could be explored, allowing the system to adjust deferral rates based on input complexity or operational constraints rather than fixed budgets. Finally, integrating cost-sensitive learning and decision-theoretic optimization may provide a more principled trade-off between automation, human effort, and error risk, while recent advances in large language models suggest that verbalized uncertainty and self-reflection may offer complementary approaches for reliability estimation.

Author Contributions: Conceptualization, T.A. and A.A.; Methodology, T.A.; Investigation, T.A.; Data curation, T.A.; Writing—original draft, T.A.; Writing—review & editing, A.A.; Supervision, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available on the Kaggle platform: <https://www.kaggle.com/datasets/matthewjansen/pubmed-200k-rtc> (accessed on 7 May 2026).

Acknowledgments: We would like to thank the Deanship of Scientific Research at Shaqra University for supporting this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hughes, M.; Li, I.; Kotoulas, S.; Suzumura, T. Medical Text Classification Using Convolutional Neural Networks. In *Informatics for Health: Connected Citizen-Led Wellness and Population Health*; Studies in Health Technology and Informatics; IOS Press: Amsterdam, The Netherlands, 2017; Volume 235, pp. 246–250.
2. Peluso, A.; Danciu, I.; Yoon, H.-J.; Yusof, J.M.; Bhattacharya, T.; Spannaus, A.T.; Schaefferkoetter, N.T.; Durbin, E.B.; Wu, X.-C.; Stroup, A.; et al. Deep Learning Uncertainty Quantification for Clinical Text Classification. *J. Biomed. Inform.* **2024**, *149*, 104576. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Strong, J.; Men, Q.; Noble, J.A. Trustworthy and Practical AI for Healthcare: A Guided Deferral System with Large Language Models. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 28413–28421. [\[CrossRef\]](#)
4. Dernoncourt, F.; Lee, J.Y. PubMed 200k RCT: A Dataset for Sequential Sentence Classification in Medical Abstracts. In *Proceedings of the 8th International Joint Conference on Natural Language Processing, (Vol. 2: Short Papers), Taipei, Taiwan, 27 November–1 December 2017*; Asian Federation of Natural Language Processing: Taipei, Taiwan, 2017; pp. 308–313.
5. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2020**, *36*, 1234–1240. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Wu, X.; Xiao, L.; Sun, Y.; Zhang, J.; Ma, T.; He, L. A Survey of Human-in-the-Loop for Machine Learning. *Future Gener. Comput. Syst.* **2021**, *135*, 364–381. [\[CrossRef\]](#)
8. Zhu, F.; Zhang, X.-Y.; Cheng, Z.; Liu, C.-L. Revisiting Confidence Estimation: Towards Reliable Failure Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 3370–3387. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Hemmer, P.; Thede, L.; Vossing, M.; Jakubik, J.; Köhl, N. Learning to Defer with Limited Expert Predictions. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 6002–6011. [\[CrossRef\]](#)
10. Hu, Y.; Chen, Y.; Xu, H. Towards More Generalizable and Accurate Sentence Classification in Medical Abstracts with Less Data. *J. Healthc. Inform. Res.* **2023**, *7*, 542–556. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Hu, Y.; Chen, Y.; Xu, H. Improving Sentence Classification in Abstracts of Randomized Controlled Trial Using Prompt Learning. In *Proceedings of the 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), Rochester, MN, USA, 11–14 June 2022*; pp. 606–607.
12. Jin, D.; Szolovits, P. Hierarchical Neural Networks for Sequential Sentence Classification in Medical Scientific Abstracts. *arXiv* **2018**, arXiv:1808.06161. [\[CrossRef\]](#)
13. Mehrtens, H.A.; Kurz, A.; Bucher, T.-C.; Brinker, T.J. Benchmarking Common Uncertainty Estimation Methods with Histopathological Images under Domain Shift and Label Noise. *Med. Image Anal.* **2023**, *89*, 102914. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
15. Theunissen, L.; Mortier, T.; Saey, Y.; Waegeman, W. Uncertainty-Aware Single-Cell Annotation with a Hierarchical Reject Option. *Bioinformatics* **2024**, *40*, btae128. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Hüllermeier, E.; Waegeman, W. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Mach. Learn.* **2021**, *110*, 457–506. [\[CrossRef\]](#)
17. Lu, Y.; Chen, T.; Hao, N.; Van Rechem, C.; Chen, J.; Fu, T. Uncertainty Quantification and Interpretability for Clinical Trial Approval Prediction. *Health Data Sci.* **2024**, *4*, 0126. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Hendrickx, K.; Perini, L.; Van der Plas, D.; Meert, W.; Davis, J. Machine Learning with a Reject Option: A Survey. *Mach. Learn.* **2024**, *113*, 3073–3110. [\[CrossRef\]](#)
19. Franc, V.; Prusa, D.; Voracek, V. Optimal Strategies for Reject Option Classifiers. *J. Mach. Learn. Res.* **2023**, *24*, 1–49.
20. Geifman, Y.; El-Yaniv, R. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*; Proceedings of Machine Learning Research; PMLR: Cambridge, MA, USA, 2019; Volume 97, pp. 2151–2159.
21. García-Galindo, A.; López-De-Castro, M.; Armañanzas, R. Multi-class Classification with Reject Option and Performance Guarantees Using Conformal Prediction. In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*; Proceedings of Machine Learning Research; PMLR: Cambridge, MA, USA, 2024; Volume 230, pp. 295–314.
22. Geifman, G.; El-Yaniv, R. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
23. Liu, Z.; Wang, Z.; Liang, P.P.; Salakhutdinov, R.; Morency, L.-P.; Ueda, M. Deep Gamblers: Learning to Abstain with Portfolio Theory. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 10622–10632.

24. Cortes, C.; DeSalvo, G.; Mohri, M. Naturally Constrained Reject Option Classification. In *Advances in Neural Information Processing Systems 29 (NeurIPS 2016)*; Curran Associates, Inc.: Red Hook, NY, USA, 2016.
25. Rabanser, S.; Thudi, A.; Hamidieh, K.; Dziedzic, A.; Bahceci, I.; Bin Sediq, A.; Sokun, H.; Papernot, N. Selective Prediction via Training Dynamics. 2025. Available online: <https://openreview.net/forum?id=niHMkXwPxf> (accessed on 7 May 2026).
26. de Carvalho, S.G.T.; de Moraes, R.M.; Ludermit, T.B. Selective Prediction with Long Short-Term Memory Using Ensemble Confidence Measures. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, 19–24 July 2020.
27. Atf, Z.; Mahjoub Far, A.; Lewis, P.R. From Confidence to Care: Rule-Based Escalation for Trustworthy Clinical AI. In *Proceedings of the 2025 IEEE International Conference on Collaborative Advances in Software and Computing (CASCON)*, Toronto, ON, Canada, 10–13 November 2025; pp. 587–588. [[CrossRef](#)]
28. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 3–7 November 2019; pp. 3615–3620.
29. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*; *Proceedings of Machine Learning Research*; PMLR: Cambridge, MA, USA, 2017; Volume 70, pp. 1321–1330.
30. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*; *Proceedings of Machine Learning Research*; PMLR: Cambridge, MA, USA, 2016; Volume 48, pp. 1050–1059.
31. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, NY, USA, 1993.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.