IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 April 2026, accepted 19 May 2026, date of publication 27 May 2026, date of current version 2 June 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3697415

RESEARCH ARTICLE

Automated Classification of EQ-5D Literature in PubMed Using Multi-Phase Learning and LLM-Assisted Co-Training

ZHYAR RZGAR K. ROSTAM[1,2], (Member, IEEE), MÁRTA PÉNTEK[3,4], (Member, IEEE), JÁNOS TIBOR CZERE[4], (Member, IEEE), ZSOMBOR ZRUBKA[3,4], (Member, IEEE), LÁSZLÓ GULÁCSI[3,4], (Member, IEEE), AND GÁBOR KERTÉSZ[2,5], (Senior Member, IEEE)

[1]Doctoral School of Applied Informatics and Applied Mathematics, Obuda University, 1034 Budapest, Hungary
[2]John von Neumann Faculty of Informatics, Obuda University, 1034 Budapest, Hungary
[3]Health Economics Research Center, University Research and Innovation Center, Obuda University, 1034 Budapest, Hungary
[4]Doctoral School of Innovation Management, Obuda University, 1034 Budapest, Hungary
[5]Laboratory of Parallel and Distributed Systems, Institute for Computer Science and Control (SZTAKI), Hungarian Research Network (HUN-REN), 1111 Budapest, Hungary

Corresponding author: Zhyar Rzgar K. Rostam (kwekha.rostam.zhyar@nik.uni-obuda.hu)

The work of Márta Péntek, Zsombor Zrubka, and László Gulácsi has been supported by the National Research, Development, and Innovation Fund of Hungary, financed under the TKP2021-NKTA-36 funding scheme at Obuda University, Hungary.

ABSTRACT The EQ-5D is a widely used tool for measuring health-related quality of life (HRQoL) to support clinical, economic, and policy decision making. Manually classifying the growing volume of literature reporting EQ-5D data for systematic literature reviews is a challenging, inefficient, and labor-intensive task. To address this, we propose a comprehensive classification framework utilizing pre-trained language models (PLMs), including BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT, to categorize PubMed records based on whether the article reports EQ-5D data using article metadata (titles, abstracts, and keywords). We examine three learning approaches: supervised learning, semi-supervised learning with pseudo-labeling, and a co-training strategy with and without Large Language Model (LLM) assistance (GPT and Claude) in pseudo-label generation. We introduce a confidence-based ensembling within the co-training framework to improve classification reliability and robustness. This study provides a systematic multi-phase evaluation of supervised, semi-supervised, and co-training paradigms on PubMed records using different input configurations. It investigates model performance starting with 200 labeled samples, expanded through iterative pseudo-labeling of an unlabeled dataset, while benchmarking across models. The results show that the co-training approach achieves the highest performance, with an F1-score of up to 0.85. Performance is reported using multi-seed evaluation with mean ± standard deviation and 95% confidence intervals. LLM-assisted co-training improves weaker model pairs but may degrade performance for already strong model combinations and reduce the number of high-confidence pseudo-labeled samples due to confidence thresholds. LLMs used within ensemble and co-training approaches provide an effective framework for EQ-5D literature screening under limited labeled data.

INDEX TERMS EQ-5D, health-related quality of life, large language model, PubMed, BioNLP, co-training, LLM-assisted pseudo-labeling, systematic literature screening.

I. INTRODUCTION

The EQ-5D is a standardized questionnaire commonly used in health-related research and health economics, developed by the EuroQol Group to provide a simple and general

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

measure of quality of life related to health. Although several versions of the questionnaire have been developed and validated (e.g., youth version, three- and five-level response scale, different administration methods), all comprise two main parts. The descriptive system of the questionnaire is composed of five domains along which patients' problems are self-reported, such as *mobility, self-care, usual activities, pain/discomfort, and anxiety/depression*. Responses of problem levels on the five domains provide a health profile that can be converted into a single index value that consistently has more advantages in the context of economic studies and health policies [1], [2], [3]. The second part of the questionnaire is a visual analogue scale (EQ VAS) to assess respondents' global health on a 0-100 (worst / best imaginable health) scale.

A systematic literature review (SLR) is a scientifically sound, transparent, and reproducible method to identify, select, and critically evaluate all available evidence in the literature on a specific question. In medicine, SLRs are considered the highest level of evidence and are used to develop clinical guidelines and foundations for health-related decisions. The rapid increase in academic publications has made it increasingly difficult to identify relevant studies for SLR. PubMed is one of the most widely used biomedical and life sciences literature databases, with over 39 million citations[1] and abstracts [4]. The database is free and publicly accessible; however, it does not include full-text journal articles but only main data (typically titles, abstracts, keywords, and metadata) and often links for sources where full texts are available. Classifying PubMed records for relevant studies by screening titles and abstracts is a labor-intensive task and requires substantial time and effort. Despite the involvement of multiple reviewers, the literature screening remains prone to errors and inconsistency, especially in cases involving complex inclusion criteria.

These challenges highlight the need for a reliable automated classification model to support systematic reviews [5], [6], [7]. It is important to keep a balance between reducing manual effort and maintaining the high sensitivity required to ensure comprehensive evidence capture. Recent advances in natural language processing (NLP) [8], [9] and deep learning (DL) [10], [11], [12], particularly transformer-based large language models (LLMs) [12], [13], [14], have achieved remarkable success in almost all NLP tasks [15], [16], [17], [18], [19], [20], [21].

The increasing importance of EQ-5D data in clinical decision making and health policy development demands efficient and accurate approaches for classifying studies. A recent study [3] has shown that machine learning models can effectively identify studies reporting EQ-5D data in full-text articles using PubMed metadata consisting of titles, abstracts, and keywords. In particular, transformer-based models such as BioBERT, PubMedBERT, and BioLinkBERT [22], [23], [24], specifically designed for biomedical text, have achieved better performance than general-purpose models, underscoring the potential of LLMs to automate the screening of the EQ-5D literature. However, there are challenges to address in improving model robustness, scalability, and accuracy. To overcome these challenges, this study proposes targeted enhancements in model design, training strategies, and ensemble techniques.[2] This study aims to:

- Propose different automated classification approaches to classify scientific studies in PubMed that report EQ-5D data from patients or the general public in their full-text version by only utilizing metadata (title, abstract, and keywords).
- Address the challenges in SLRs by reducing the manual labeling effort and errors in screening literature related to EQ-5D.
- Evaluate and benchmark pre-trained language models (PLMs) such as BERT [25], SciBERT [26], BioBERT [22], PubMedBERT [23], and BioLinkBERT [24] for binary classification of EQ-5D relevance.
- Enhance model performance through advanced learning strategies, including semi-supervised learning, co-training with and without LLM-assisted pseudo-labeling (GPT and Claude).
- Explore ensemble methods such as weighted voting to improve robustness and prediction accuracy.

In this study, we make the following contributions:

- Develop and evaluate a pipeline using different PLMs (BERT [25], SciBERT [26], BioBERT [22], PubMedBERT [23], and BioLinkBERT [24]) applied to EQ-5D classification tasks.
- Investigate different input strategies, incorporating titles, abstracts, keywords, and their combinations (title + abstract (TA), title + abstract + keywords (TAK)) to optimize model input representation.
- Utilize a weighted voting aggregation mechanism to enhance prediction reliability across different models and input setups.
- Implement a semi-supervised learning approach using pseudo-labeling to expand the labeled dataset, demonstrating substantial performance enhancement.
- Design a curriculum-based co-training framework where pairs of PLMs iteratively refine each other using high-confidence pseudo-labels ($\tau \in \{0.95, 0.92, 0.90\}$), and MC-Dropout uncertainty filtering across all ten pairwise combinations.
- Increase classification confidence and accuracy in the co-training framework by combining LLM-assisted pseudo-labeling (GPT and Claude) with model agreement filtering.

[1] https://pubmed.ncbi.nlm.nih.gov/

[2] The datasets, model notebooks, and sample results are available in this repository: https://github.com/ZhyarUoS/automated-multi-phase-eq-5d-classification.git

- Achieve strong classification performance, with F1-scores up to 0.85 in the best-performing co-training setups.

A. RELATED WORKS

1) NLP FOR SYSTEMATIC LITERATURE REVIEW

Wallace et al. [15] propose a semi-automated citation screening method to reduce manual work for reviews. The approach uses an ensemble of Support Vector Machines (SVMs) trained on multiple feature spaces such as titles, abstracts, MeSH terms, and UMLS concepts. An active learning approach with a novel strategy (Patient Active Learning (PAL)) is integrated, which is designed for imbalanced class datasets and introduces aggressive undersampling to prevent missing relevant studies. Evaluation on three datasets shows that this approach can reduce the manual workload up to 50% in two of the datasets and 40% in the third. The study offers a promising solution to streamline the citation screening process without compromising review quality.

Cohen et al. [27] investigate the use of automatic citation classification to reduce the expert effort in updating SR of drug class efficacy. To classify citations as either relevant or not, in the study, they develop a voting perceptron-based learning system while using data from 15 annotated systematic drug class reviews. The system is evaluated using cross-validation for its performance in precision, recall, F-measure, and work saved over sampling at a recall rate of 95%. The results showed that in most (11 out of the total 15) of the review topics, the method reduces the number of articles that need manual screening, with reductions of 50% or more in three topics. The study suggests that the use of automatic classification would be able to help maintain systematic drug reviews, although considerable extra development and integration studies are still required.

2) PLMS IN BIOMEDICAL TEXT CLASSIFICATION

Beltagy et al. [26] introduce SciBERT, a PLM designed for scientific domains in order to address the scarcity of large annotated corpora in scientific NLP tasks. SciBERT is built on BERT [25] architecture and trained on a huge multi-domain corpus of scientific literature through unsupervised learning. The model is evaluated across several downstream tasks, such as sequence tagging, sentence classification, and dependency parsing. The results showed that the model outperformed general-purpose models like BERT. The study demonstrated that the use of domain-specific pre-training and relevant vocabularies causes exceptional increases in performance; SciBERT intends to be a valuable resource for processing scientific texts.

Lee et al. [22] develop a PLM (BioBERT) specifically for biomedical text mining. BioBERT is designed based on the BERT [25] architecture, fine-tuned on domain-specific large-scale biomedical corpora, including the PubMed abstracts and full-text articles. This specialization made the model surpass the general-purpose models and the previous SOTA models significantly on different biomedical NLP tasks like named entity recognition (NER), relation extraction (RE), and question answering (QuA). Notably, performance improvements demonstrate the value of domain-specific pre-training in capturing complexity in biomedical language and terminology.

Gu et al. [23] demonstrate that the training of language models from scratch on biomedical text (such as PubMedBERT) produces better models than adapting a general model like BERT. The study shows across multiple biomedical NLP tasks that domain-specific pre-training and vocabulary are both extremely important for improving results, while mixed-domain approaches may actually hurt performance. The study also suggests that simpler modeling choices remain effective with the transformer models.

Danilov et al. [28] explore automated text classification techniques aiming at supporting the selection of scientific literature by focusing on binary classification of 630 PubMed abstracts. Researchers fine-tuned different configurations of the PubMedBERT model and evaluated the models' performance with ensemble methods combining PubMedBERT, logistic regression, random forest, and SVMs. The findings demonstrate that modern NLP methods, particularly PLMs, are highly effective for classifying short scientific texts.

Yasunaga et al. [24] propose LinkBERT, which enhances PLMs by using document links such as hyperlinks or citations, allowing the exchange of knowledge between documents. The model learns relationships between documents and achieves better performance than BERT, especially on question answering and reasoning tasks.

Landschaft et al. [29] evaluate the performance of GPT-4 as an additional reviewer in the systematic literature reviews based on the PRISMA approach. GPT-4 was integrated into the review process for the patient-trial matching, following a retrieval-augmented generation (RAG) approach using LangChain. Its performance for abstract screening and full-text parameter extraction is assessed against human reviewers. The results indicate impressive agreements with human reviewers for the screening of abstracts (Cohen's kappa $>$ 0.9), although agreement is lower in full-text tasks. The study concludes that GPT-4 is able to support or partially supplant human reviewers in the abstract screening phase, thus reducing the manual effort needed for systematic reviews.

3) SEMI-SUPERVISED AND CO-TRAINING METHODS

Blum and Mitchell [30] introduce a co-training approach, which uses two distinct and complementary views of the data to enhance learning from limited labeled examples by using large numbers of unlabeled data. Two classifiers are trained in each view separately and then iteratively teach each other by labeling new data. Experiments and results show that this approach can significantly enhance classification performance, especially in scenarios like web page classification or multi-modal learning tasks.

Gao et al. [31] suggest a reliable method that involves transfer learning and semi-supervised self-training to improve biomedical NER with limited labeled data. Leveraging models like BiLSTM-CRF and BERT, pre-training on large biomedical corpora followed by self-training with unlabeled data boosts performance. The study achieves results comparable to fully supervised algorithms, but with only a fraction of labeled examples, and works best for common biomedical entities. Transfer learning is demonstrated to be required to enable effective self-training in low-resource scenarios.

II. MATERIALS AND METHODS

A. DATASET DESCRIPTION

In this work, we utilize a dataset prepared by Kertész et al. [3]. In brief, publications on EQ-5D in PubMed database between 1990 and 2022 were identified using the EuroQol Group's PubMed search filter,[3] resulting in a total of 15,547 discrete records. The dataset was prepared by randomly selecting 200 studies from this pool, which were manually labeled by two independent expert reviewers (based on their title/abstract/keywords, then on their full-text version) for being EQ-5D studies or not in terms of reporting EQ-5D data from patients or the general public. The classification was binary, indicating whether the studies presented data with any EQ-5D version and administration method or not (''true'' or ''false''). All abstracts from the 15,547 extracted records were in English; however, the full-text of some studies could have been in another language. Therefore, these studies were considered negative by the reviewers based on the eligibility criteria. (In our current analyses, their presence was considered as noise in terms of performance). Results from the two independent reviewers were matched, and any differences were discussed until a consensus was reached in order to categorize each study. As a result, the dataset contains 200 labeled data points (see Fig. 1), each consisting of a title, abstract, keywords, and a label (''true'' or ''false'').

1) SPLITTING DATASET

In order to evaluate model performance and accuracy, the dataset must be split into training, testing, and validation subsets. The training subset is used solely to train the models, the testing subset is used exclusively to evaluate model performance after training, and the validation subset is used for model selection and to prevent overfitting. In this study, we adopt two different splitting strategies on the labeled dataset prepared by Kertész et al. [3]: First, for the initial phase (training and fine-tuning) and the subsequent phases (inference and weight voting), the dataset is split into a training set (60%) and a temporary hold-out set (40%), which is further divided into validation (12% of the total data) and test (28%) subsets. The validation set is used only for early stopping and model selection. Second, for both the semi-supervised and the co-training approaches, the dataset is divided into a training set (50%), testing set (35%), and validation set (15%). To prevent data leakage, the test set remains strictly unseen until the final evaluation in all approaches (see (1), (2), and (3)). Where $\mathcal{D}$ is the complete labeled dataset.

$$\mathcal{D}_{\text{train}} = \mathcal{D} \setminus (\mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{test}}) \tag{1}$$

$$\mathcal{D}_{\text{val}} = \mathcal{D} \setminus (\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}) \tag{2}$$

$$\mathcal{D}_{\text{test}} = \mathcal{D} \setminus (\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}}) \tag{3}$$

[3]https://euroqol.org/information-and-support/documentation/pubmed-search/

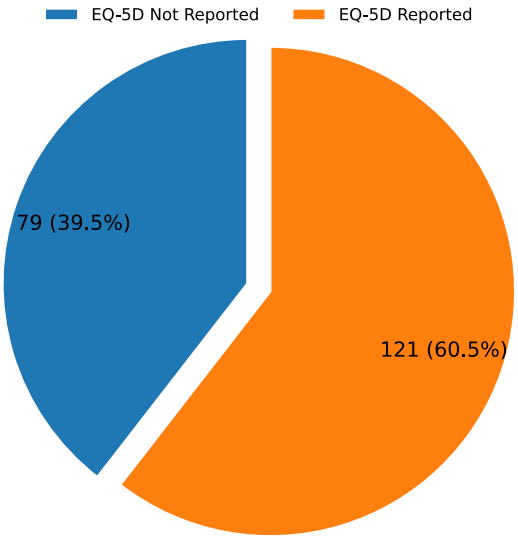

FIGURE 1. Binary label distribution in the EQ-5D dataset.

B. PREPROCESSING

We follow a systematic approach to data preparation prior to feeding data to the models. In this step, we apply data cleaning, shuffling, splitting, input configuration, tokenization, and encoding for the subsequent model training and evaluation.

1) DATA CLEANING

We normalize titles, keywords, and abstracts in the dataset by transforming texts to lowercase and stripping all non-alphanumeric characters based on a set of regular expressions. These techniques make all text parts consistent and minimize noise introduced by minor formatting inconsistencies.

2) SHUFFLING AND SPLITTING

In order to avoid ordering biases, the labeled data are randomly shuffled before feeding to the models. In addition to this, we utilize a stratified strategy, which helps preserve the class distributions across subsets. As stated in Section II-A1, two different splitting strategies are used with different phases.

3) INPUT CONFIGURATION

All inputs are treated in two different setups, individually (title, keywords, and abstract) and in a combination form (TAK), by combining all three initial parts together and separating them with a [SEP] token in between: Title [SEP] Keywords [SEP] Abstract. The specialized token [SEP] allows Transformer models to semantically differentiate between two sources of the input while keeping relative positioning.

Based on the results from the training and fine-tuning phases (as detailed in Table 3), the selected configuration for all semi-supervised and co-training tasks was the combination of title and abstract (TA), which was also investigated during the fine-tuning phase.

4) TOKENIZATION AND ENCODING

Input sequences are tokenized using pre-trained (BERT, SciBERT, BioBERT, PubMedBERT, or BioLinkBERT) tokenizer, involving truncation to the maximum sequence length of 128 tokens and padding. The tokenized inputs are converted into a set of tensors of input IDs, attention masks, and label IDs (where applicable).

This preprocessing enables a consistent and reproducible pipeline that integrates well into any Transformer-based architecture, allowing for efficient training and fine-tuning on the EQ-5D classification task.

III. SUPERVISED APPROACH SETUP

This section provides details on both training and fine-tuning strategies, hyperparameter settings, learning rate exploration, early stopping criteria, and model preservation. To ensure generalization and optimize classification performance, each of the models (BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT) is investigated through a structured training and fine-tuning regimen (see Fig. 2). Various input configurations (see Section II-B3) are evaluated during the training and fine-tuning stages. In the training phase, models are initialized from scratch using randomly initialized weights. In contrast, during fine-tuning, models are initialized from pre-trained transformer checkpoints and are further adapted through masked language model (MLM) pre-training before being optimized for the EQ-5D classification task.

A. LEARNING RATE EXPLORATION

Selecting learning rates wisely during the learning process has a direct impact on the convergence and stability of the models. We systematically evaluate nine distinct learning rates (candidate set): $\eta \in \{1\times10^{-4},\ 2\times10^{-4},\ 5\times10^{-4},\ 1\times10^{-5},\ 2\times10^{-5},\ 5\times10^{-5},\ 1\times10^{-6},\ 2\times10^{-6},\ 5\times10^{-6}\}$. Each of the learning rates mentioned is independently checked with consistent training and validation subsets. Models are monitored for performance improvements using validation micro F1 scores. This strategy assists in identifying the most effective learning rate η^* specific to each model architecture and scale of the dataset (see (4)).

$$\eta^* = \arg\max_{\eta\in\mathcal{H}} \mathrm{F1}_{\mathrm{val}}(\eta). \tag{4}$$

where $\mathcal{H}$ is the set of all possible learning rates and $\mathrm{F1}_{\mathrm{val}}(\eta)$ is the performance measure of the validation set for learning rate η.

For each candidate learning rate and each random seed, the model is reinitialized to ensure independent evaluation and avoid bias from previous optimization states.

B. EPOCH SCHEDULE

The training is conducted in two phases:

- Training phase: The labeled dataset is utilized in the full training task. All of the models are trained for up to 100 epochs.
- Fine-tuning phase: All models are fine-tuned for up to 100 epochs using the manually labeled dataset. Models are first adapted using MLM pre-training on an extra 200 randomly selected unlabeled records and then fine-tuned for the EQ-5D classification task.

C. EARLY STOPPING STRATEGY

To avoid overfitting, we apply an early stopping strategy based on validation set performance. In both the training and fine-tuning phases, the process is terminated if no further improvement is observed after ten consecutive epochs. Following early stopping, the model with the best performance on the validation set is adopted for further investigation.

D. OPTIMIZATION AND SCHEDULING

We employ the AdamW optimizer for weight decay with decoupled weights during optimizer training, which improves generalization. To smooth the early training epochs and avoid abrupt updates, a linear rate scheduler is integrated that allows for a warm-up. The number of warm-up steps is 10% of the total steps during training. This approach matches the best practices in Transformer fine-tuning and contributes to smoother convergence.

E. MULTI-SEED EVALUATION

To ensure the results are reliable and reduce variability due to random initialization, all experiments are conducted five times with different random seeds: 42, 123, 2023, 777, and 999. The reported results are the mean and standard deviation of these runs.

F. BATCH PROCESSING AND HARDWARE UTILIZATION

The models are trained with a batch size of 32, and all the experiments are done on GPU-enabled environments, ensuring efficient parallelization. Mixed precision training is enabled where feasible to minimize memory footprint and accelerate computation.

G. MODEL CHECKPOINTS AND REPRODUCIBILITY

During the learning process, we keep the best-performing model for future use. In this stage, both weights and the

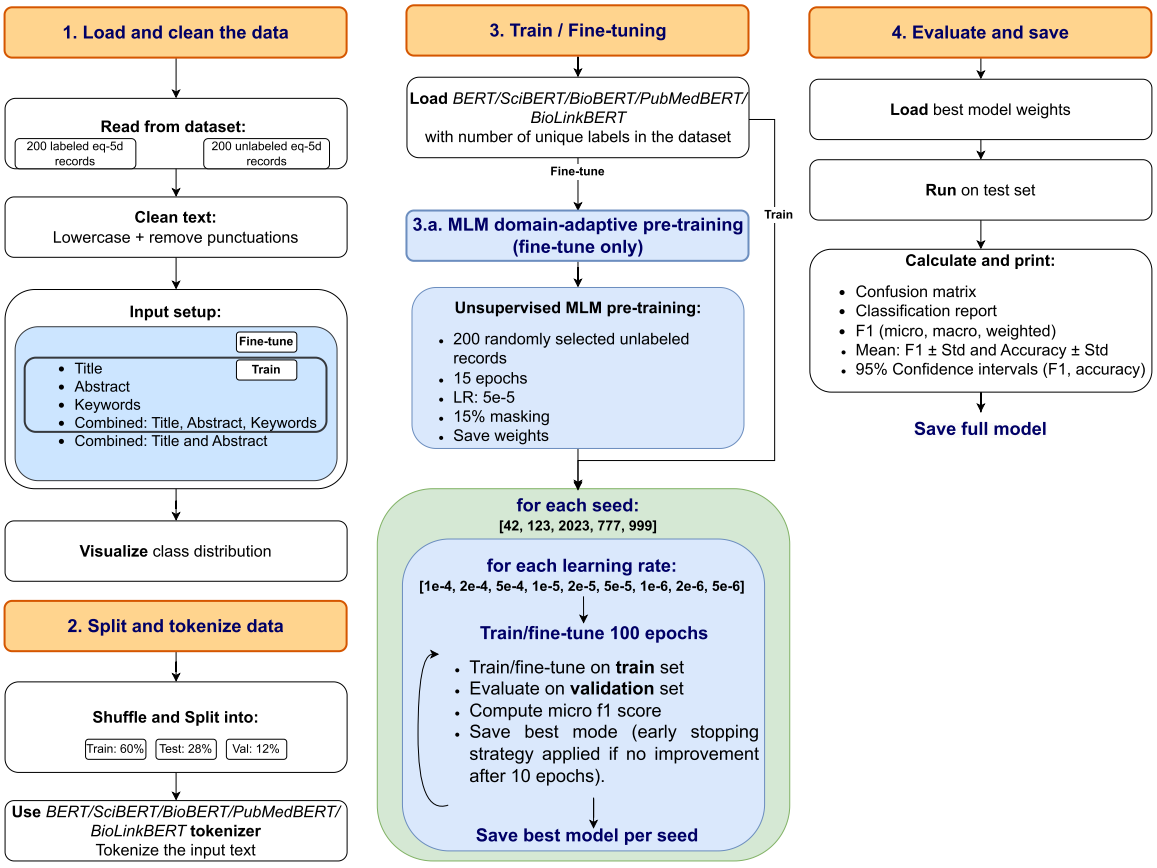

FIGURE 2. Training and fine-tuning approach.

configuration file were saved; this allows for the model to be reloaded exactly as it was during the evaluation process, inference, or future training processes.

IV. INFERENCE AND WEIGHTED VOTING AGGREGATION

In this phase, for evaluating the generalization of the model and increasing the robustness of prediction, we randomly select an extra 200 unlabeled studies from the original dataset (N = 15,547 records) collected by Kertész et al. [3], then conduct inference using an ensemble of fine-tuned models (BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT). All models used in this phase correspond to the best-performing checkpoints obtained during the fine-tuning stage. The dataset is structured with metadata fields in its records that contain title, abstract, and keywords, and this is applied across individual inputs and their combined representations to evaluate the classification performance of the model over multiple configurations.

A. MULTI-CONFIGURATION INFERENCE

Fine-tuned models (BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT) that are saved in the previous phase are applied across the four distinct input configurations: title, abstract, keywords, and combining all three. Each configuration is evaluated using fine-tuned models, resulting in a total of 20 prediction results per study: Total predictions = 5 Models × 4 input scenarios = 20 different predictions for each study. Text normalization and tokenization are conducted uniformly across all configurations. Softmax-normalized output probabilities are used to give each prediction a confidence score. (see Figs. 3, and 4).

B. WEIGHT VOTING MECHANISM

In order to achieve a single prediction across multiple predictions, the weighted voting technique is adopted. For each study, the individual class predictions from four input variants of each model are aggregated by weighting each predicted class according to its associated confidence score. The final predicted label is the class with the highest aggregated confidence from the input variants (as shown in 5):

$$\hat{y} = \arg\max_{c \in \{0,1\}} \sum_{i=1}^{4} \mathbb{I}[\hat{y}_i = c] \cdot \text{Conf}_i. \tag{5}$$

- $\hat{y}$: Final predicted label after weighted voting.

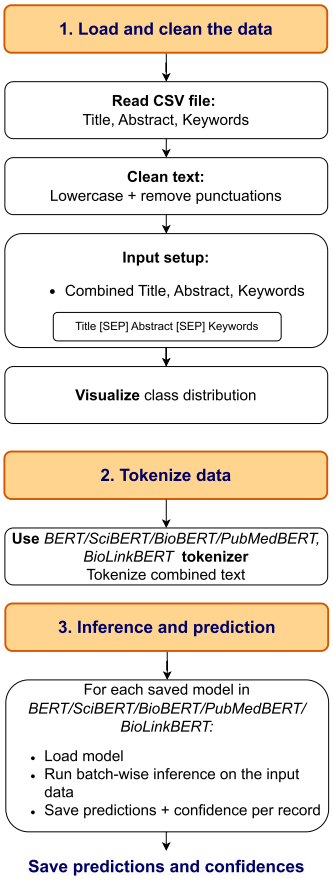

FIGURE 3. Utilize stored models for inference.

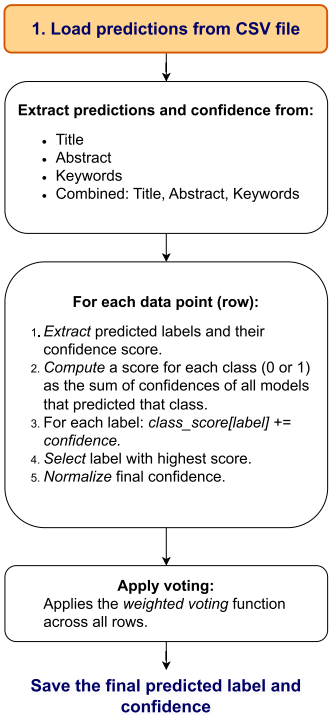

FIGURE 4. Weighted voting from multiple models.

- $\hat{y}_i$: Predicted label from the i-th input variant (title, abstract, keywords, combined).
- Conf_i: Maximum softmax probability associated with $\hat{y}_i$, i.e., $\text{Conf}_i = \max_{c\in\{0,1\}} P(c \mid x_i)$.
- $\mathbb{I}[\hat{y}_i = c]$: Indicator function that equals 1 if $\hat{y}_i = c$, otherwise 0.

The summation runs over all four input variants. The final label $\hat{y}$ is the class $c \in \{0, 1\}$ with the highest total weighted score.

The final confidence score is computed as the normalized weighted score (see (6)):

$$\text{Conf}_{\text{final}} = \frac{\sum_{i=1}^{4} \mathbb{I}[\hat{y}_i = \hat{y}] \cdot \text{Conf}_i}{\sum_{i=1}^{4} \text{Conf}_i}. \tag{6}$$

C. AGGREGATED RESULTS

The final predictions and their normalized confidence scores are stored and used for further assessment with inter-model consistency and to evaluate the relative influence of different input types.

V. SEMI-SUPERVISED LEARNING

After the inference and weight voting approach on the unlabeled dataset, the semi-supervised approach is employed to enhance the performance of the models (BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT) for text classification of detecting EQ-5D reported in the PubMed dataset. The approach uses 200 labeled records and an extra 2,000 randomly selected unlabeled records (see Section II-A, and Fig. 5) to maximize the accuracy of the model through an iterative teacher-student pseudo-labeling process.

A. INITIAL DATA PROCESSING

The initial stage of this approach involves utilizing and preparing two distinct datasets (labeled and unlabeled). The first is a manually annotated corpus of 200 studies (explained in Section II-A). For the second, unlabeled corpus, 2,000 randomly selected papers are used as the pseudo-labeling pool. Only the title and abstract (TA) fields are used as input, concatenated with a `[SEP]` token, based on the best-performing input configuration identified in the preceding supervised phases (see Section VII-A). All inputs undergo a standardized preprocessing pipeline (explained in Section II-B) before analysis. The labeled dataset is divided into three parts: a training set (50%), a testing set (35%), and a validation set (15%). This division follows the same method used in the semi-supervised and co-training approaches (see Section II-A). To prevent data leakage, the test set is kept completely separate and is not used in either dataset expansion or pseudo-label creation.

B. MODEL FINE-TUNING

The labeled dataset is used alone during the fine-tuning process. All five models (BERT, SciBERT, BioBERT,

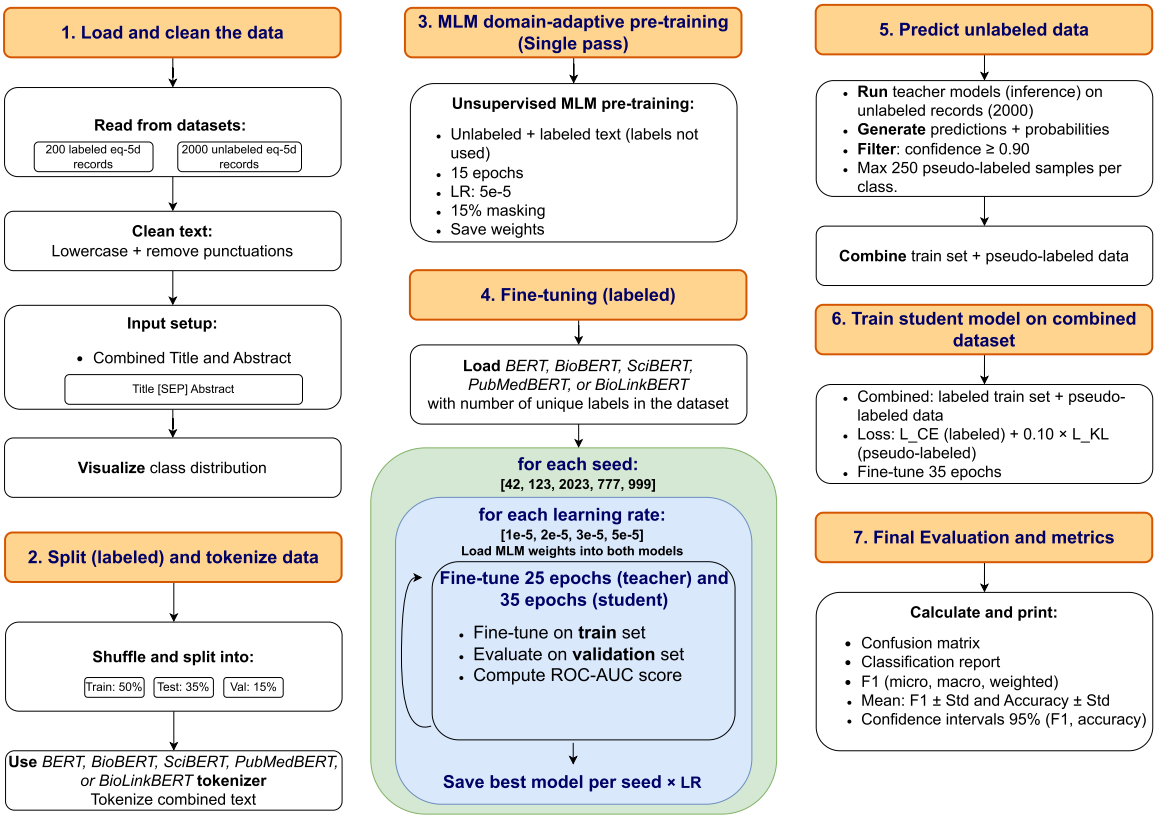

FIGURE 5. Semi-supervised learning approach.

PubMedBERT, and BioLinkBERT) are fine-tuned with four learning rates (candidate set)
$\eta \in \{1 \times 10^{-5},\ 2 \times 10^{-5},\ 3 \times 10^{-5},\ 5 \times 10^{-5}\}$ across a maximum of 25 epochs for the teacher and 35 epochs for the student model. Model selection is based on the best validation ROC-AUC score, which is adopted as the selection criterion.

C. SEMI-SUPERVISED EXTENSION THROUGH PSEUDO-LABELING

The pseudo-labeling approach is used to extend the capacity of models in a semi-supervised setting. After the initial supervised phase, the best-performing teacher model generates soft pseudo-labels for the unlabeled pool of 2,000 studies (see (7)). Each prediction is assigned a confidence score, calculated from its softmax probabilities. Only samples whose confidence exceeds a threshold $\tau = 0.90$ are retained, with a maximum of 250 pseudo-labeled samples per class (see (8)).

$$\tilde{y}_j = \arg\max_c f_\theta(x_j)_c, \quad \text{retain if } \max_c f_\theta(x_j)_c \geq \tau. \tag{7}$$

$$\begin{aligned} \mathcal{D}' &= \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{pseudo}} \\ \mathcal{D}_{\text{pseudo}} &= \{(x_j, \tilde{p}_j) \mid x_j \in \mathcal{D}_U,\ \max_c f_\theta(x_j)_c \geq \tau, \\ &\qquad |\mathcal{D}^c_{\text{pseudo}}| \leq K\}. \end{aligned} \tag{8}$$

A student model is then trained on the combined supervised and pseudo-labeled data, minimizing a joint loss (see (9)):

$$\mathcal{L}_{\text{semi}} = \mathcal{L}_{\text{CE}}(\mathcal{D}_{\text{train}};\ \theta) + \lambda\, \mathcal{L}_{\text{KL}}(\mathcal{D}_{\text{pseudo}};\ \theta). \tag{9}$$

where $\lambda = 0.10$ is the unsupervised loss weight, $f_\theta(x)$ denotes the model's predicted class probabilities (softmax output), θ represents the model parameters, $\mathcal{D}_{\text{train}}$ is the labeled training subset, $\mathcal{D}_U$ is the unlabeled pool, $\tilde{p}_j$ is the soft pseudo-label vector assigned to the unlabeled sample x_j, $\tau = 0.90$ is the confidence threshold, $\mathcal{D}^c_{\text{pseudo}}$ denotes the pseudo-labeled samples of class c, $K = 250$ is the maximum number of pseudo-labeled samples per class, and $\mathcal{D}_{\text{pseudo}}$ is the confidence-filtered pseudo-labeled subset. The performance of each model is evaluated on the fixed, held-out test set, which is excluded from all pseudo-labeling and dataset expansion steps throughout the entire semi-supervised pipeline.

VI. CO-TRAINING FRAMEWORK

In this phase, a technique of co-training, which combines both the labeled and the unlabeled datasets, utilizing all five models (BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT), is used. Our approach is to investigate all ten pairwise co-training scenarios, each with

and without LLM assistance (GPT-4o-mini and Claude Haiku). Each pairwise configuration allows two distinct models to iteratively exchange high-confidence pseudo-labels, progressively enhancing performance on unlabeled data. In this study, we use the bidirectional arrow ($\leftrightarrow$) to denote co-training between models.

A. EXPERIMENTAL SETUP AND DATA PREPROCESSING

The experiments utilize both labeled (200 manually labeled) and unlabeled (2,000 randomly selected) datasets (see Section II-A). Only the title and abstract (TA) fields are used as input, concatenated with a `[SEP]` token, based on the best-performing input configuration in the supervised phases (see Section VII-A). We follow the same strategies for preprocessing and preparing the datasets prior to feeding them to the models as stated in Section II-B. Each model undergoes domain-adaptive pre-training via MLM on the combined training and unlabeled corpus prior to fine-tuning. Each model is then fine-tuned individually on the labeled training subset, keeping the best checkpoint based on the validation ROC-AUC score. Two complementary views of the same input are constructed for each pair:

- View 1 (Model 1): Title `[SEP]` Abstract
- View 2 (Model 2): Abstract `[SEP]` Title

B. CO-TRAINING PROCEDURE (WITHOUT LLM)

In this section, we explain the co-training procedure (see Fig. 6), which consists of three iterations for each pairwise model setup, following a curriculum schedule with progressively relaxed thresholds ($\tau \in \{0.95, 0.92, 0.90\}$) and increasing pseudo-label samples ($K \in \{100, 150, 200\}$) per class, providing up to 200, 300, and 400 total pseudo-labeled samples at iterations 1, 2, and 3, respectively. Each iteration performs the following tasks:

- Prediction: Unlabeled studies receive predictions from each model per pairwise individual. At iteration t, each model $M_k^{(t)}$ ($k \in \{1, 2\}$ representing any chosen pair of PLMs) predicts class probabilities for all unlabeled samples $x_j \in \mathcal{D}_U$ using MC-Dropout over P forward passes (see (10), (11) and (12)).
- High-Confidence Filtering: Keeps only predictions where both models agree, both exceed the confidence threshold $\tau^{(t)}$, and both exhibit low predictive uncertainty (variance $< \delta$), as presented in (13).
- Cross-Labeling: Each model uses the pseudo-labeled data from the other model to augment its own training set (see (14)).
- Retraining: Each model is further trained on the expanded dataset (labeled + pseudo-labeled dataset) with a consistency regularization term (see (15)).

$$f_k^{(t)}(x_j) = \frac{1}{P}\sum_{p=1}^{P} \text{softmax}(z_{k,p}^{(t)}(x_j)), \tag{10}$$

where $z_{k,p}^{(t)}(x_j)$ are the logits produced $M_k^{(t)}$ at the dropout pass p, and P is the number of MC-dropout forward passes.

The pseudo-label $\tilde{y}_{k,j}^{(t)}$ is assigned based on the highest predicted probability:

$$\tilde{y}_{k,j}^{(t)} = \arg\max_c f_k^{(t)}(x_j)_c, \tag{11}$$

with confidence

$$\text{Conf}_k^{(t)}(x_j) = \max_c f_k^{(t)}(x_j)_c. \tag{12}$$

$$\begin{aligned}
&\mathcal{P}_k^{(t)} = \{(x_j, \tilde{y}_{k,j}^{(t)}) \mid x_j \in \mathcal{D}_U, \\
&\tilde{y}_{1,j}^{(t)} = \tilde{y}_{2,j}^{(t)}, \\
&\text{Conf}_k^{(t)}(x_j) \geq \tau^{(t)}, \\
&\text{Var}_k^{(t)}(x_j) < \delta, \\
&|\mathcal{P}_k^{(t),c}| \leq K^{(t)}\}.
\end{aligned} \tag{13}$$

$$\mathcal{D}_1^{(t+1)} = \mathcal{D}_{\text{train}} \cup \mathcal{P}_2^{(t)}, \quad \mathcal{D}_2^{(t+1)} = \mathcal{D}_{\text{train}} \cup \mathcal{P}_1^{(t)}. \tag{14}$$

$$\begin{aligned}
\mathcal{L}_k^{(t+1)} = &-\frac{1}{|\mathcal{D}_k^{(t+1)}|} \sum_{(x,y)\in\mathcal{D}_k^{(t+1)}} y^\top \log f_k^{(t+1)}(x) \\
&+ \gamma\, \mathcal{L}_{\text{cons}}(x;\, \theta_k^{(t+1)}).
\end{aligned} \tag{15}$$

This process is repeated for $T = 3$ iterations, leading to progressively refined models $\{M_1^{(t)}, M_2^{(t)}\}_{t=1}^{T}$. Final predictions are obtained via an adaptive weighted ensemble, where weights are proportional to each model's F1 score at each iteration (see (16)).

$$\hat{y}_j = w_1 \cdot f_1^{(T)}(x_j) + w_2 \cdot f_2^{(T)}(x_j), \quad w_k = \frac{\text{F1}_k}{\text{F1}_1 + \text{F1}_2}. \tag{16}$$

The final decision threshold is optimized via Youden's J [32] statistic on the ROC curve.

Where $\mathcal{D}_{\text{train}}$ is the labeled training subset, $\mathcal{D}_U$ is the unlabeled pool, $\tau^{(t)}$ is the confidence threshold at iteration t, $\delta = 0.05$ is the uncertainty threshold, $\text{Var}_k^{(t)}(x_j)$ is the MC-Dropout predictive variance, $|\mathcal{P}_k^{(t),c}|$ denotes pseudo-labeled samples of class c at iteration t, $K^{(t)}$ is the per-class pseudo-label budget, and $\gamma = 0.2$ is the consistency regularization weight.

C. FINAL LABEL AGGREGATION AND MODEL CONSOLIDATION

Following the co-training iterations, predictions from each of the paired models are combined via adaptive weighted ensemble, and the final decision threshold is optimized on the ROC curve. An expanded labeled dataset is built as a result of this phase.

Final evaluations are performed on the fixed held-out test set to assess each model's capability in generalization. The results showed that the incorporation of high-confidence pseudo-labels led to robust and accurate classifiers.

D. LLM-ASSISTED CO-TRAINING

In addition to the co-training approach, we utilize LLMs with each pair to enhance the quality of pseudo-labels (see Figs. 7 and 9). Specifically, we investigate two LLMs as third

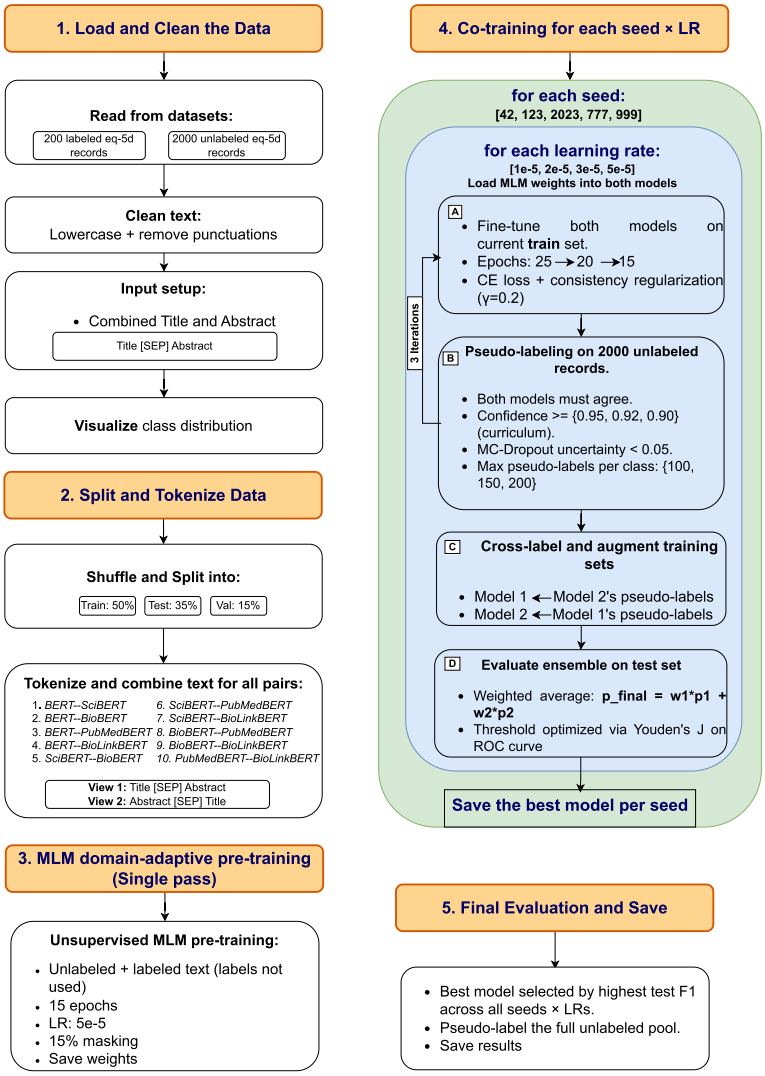

FIGURE 6. Co-training approach without LLM.

annotators: GPT-4o-mini and Claude Haiku 4.5. We pass a structured query to the LLM, injecting the title and abstract of each study and asking it to provide a prediction with a confidence level. The pseudo-label is only assigned to the study if:

- The co-training pair and the LLM achieve a majority agreement (at least 2 out of 3 votes, see (17)).
- The average confidence score across all three annotators exceeds the threshold $\tau^{(t)}$ (see (18)).

$$\hat{y}_j^{\text{LLM}} = \mathbf{1}\left[p_{1,j}^{(t)} + p_{2,j}^{(t)} + p_j^{\text{LLM}} \geq 2\right] \tag{17}$$

$$\bar{c}_j = \frac{\text{Conf}_1^{(t)}(x_j) + \text{Conf}_2^{(t)}(x_j) + c_j^{\text{LLM}}}{3}. \tag{18}$$

where $p_{k,j}^{(t)} \in \{0, 1\}$ is the hard prediction of model k, p_j^{LLM} is the LLM prediction, and c_j^{LLM} is the LLM confidence score. A sample is retained only if $\bar{c}_j \geq \tau^{(t)}$.

1) COST ANALYSIS

The total annotation cost for the 2,000 unlabeled records is presented in Table 1. GPT-4o-mini required an initial annotation cost of $0.1559 per experimental run, while Claude Haiku 4.5 required $1.1780 per experimental run. By leveraging response caching based on MD5, after the first run, all further runs across different random seeds and co-training iterations were supplied directly from the cache, thereby reducing the per-run cost to $0.0766 for GPT-4o-mini and $0.1158 for Claude Haiku 4.5, demonstrating that LLM-assisted co-training was cost-efficient compared to manual annotation.

TABLE 1. Empirical LLM annotation cost for 2,000 unlabeled records.

LLM	Avg Input Tokens/Record	Avg Output Tokens/Record	Total Tokens	First-Run Cost (USD)	Cached-Run Cost (USD)
GPT-4o-mini	510.8	2.0	1,025,600	$0.1559	$0.0766
Claude Haiku 4.5	579.0	2.0	1,162,000	$1.1780	$0.1158

E. LLM QUERY

The LLM is prompted to predict whether the full-text of a publication discusses EQ-5D based solely on its title and abstract. For each study, we run a structured prompt (stated below):

"*You're an expert in medical research. Based only on the Title, and Abstract of the following study, determine how likely it is that the full-text discusses EQ-5D. Respond with one of the following confidence levels*:

- *Definitely Yes*
- *Probably Yes*
- *Unsure*
- *Probably No*
- *Definitely No*

Title: [Title Text]
Abstract: [Abstract Text]
"

The response is parsed and mapped to a binary label and confidence score as presented in Table 2:

TABLE 2. Mapping of LLM responses to binary labels and confidence scores.

LLM Response	Binary Label	Confidence Score
Definitely Yes	1	0.99
Probably Yes	1	0.85
Unsure	0	0.50
Probably No	0	0.15
Definitely No	0	0.01

To avoid redundant API calls, LLM responses are cached using an MD5 hash of the title and abstract. The final evaluation is conducted on the fixed held-out test set, consistent with all other phases.

VII. EVALUATION AND RESULTS

In this section, we provide a comprehensive model assessment by reporting achieved results and evaluating models' performance across all phases, highlighting both classification metrics and consistency between approaches. To present the actual models' performance, different evaluation metrics are used, in addition to the models' agreement between approaches, which strengthens the reliability of predictions.

Progressive improvement in the efficiency of model representations can be observed across approaches, as one moves from supervised to semi-supervised and finally to co-training learning methodologies.

A. PERFORMANCE ACROSS SUPERVISED APPROACH

This section summarizes the classification performance of the BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT models during both training and fine-tuning (see Table 3). Metrics are presented for each input configuration (title, keywords, abstract, combined).

1) TRAINING PHASE

In this phase, all of the models achieved high recall with notably low precision, indicating a tendency to overpredict the positive class. This imbalance suggests overfitting, likely caused by the limited size of the labeled dataset, which may have led models to memorize dominant label patterns rather than learn generalization features.

BioBERT produced the most balanced performance among all models. It achieved the highest F1-score (0.68) when using title input (mean F1 $=$ 0.66 $\pm$ 0.0122, 95% CI: [0.575, 0.775]). BERT also performed well on the abstract input (F1-score is 0.63, mean F1 $=$ 0.66 $\pm$ 0.020, 95% CI: [0.600, 0.788]), whereas SciBERT achieved its best performance with title input (F1-score is 0.64, mean F1 $=$ 0.6575 $\pm$ 0.015, 95% CI: [0.575, 0.788]).

Among domain-specific models, PubMedBERT and BioLinkBERT achieved the best F1-scores of 0.62 and 0.61, respectively, both with title input (mean F1 $=$ 0.6425 $\pm$ 0.0291 and 0.6425 $\pm$ 0.0331).

2) FINE-TUNING PHASE

Fine-tuning achieved better classification balance in terms of precision and recall for all the models. PubMedBERT achieved the best overall F1-score (0.75) with title input and accuracy of 0.76 (mean F1 $=$ 0.7250 $\pm$ 0.0250, 95% CI: [0.6747, 0.8500]). BioBERT and BioLinkBERT similarly showed strong performance with title input, attaining F1-scores of 0.73 each (mean F1 $=$ 0.6750 $\pm$ 0.0500 and 0.7175 $\pm$ 0.0127 respectively), followed by SciBERT (F1-score is 0.72, mean F1 $=$ 0.6975 $\pm$ 0.0339, 95% CI: [0.6375, 0.8250]) also with title input.

BERT achieved its best result with combined input (F1-score is 0.70, mean F1 $=$ 0.6400 $\pm$ 0.0502). These gains indicated that fine-tuning assisted in reducing the overfitting impacts present in the training phase, with domain-specific models benefiting most from pre-trained biomedical knowledge.

B. INFERENCE AND AGREEMENT-BASED VOTING

In addition to assessing individual performance, we conducted inference experiments with the saved models at the fine-tuning stage. The inference and weighted voting strategy were applied to 200 unlabeled new records. The predictions were aggregated through a weighted voting strategy, prioritizing agreement among models and incorporating confidence-based selection when available. This strategy is intended to increase the overall prediction reliability by relying on an agreement across model variants.

Table 4 summarizes the agreements while utilizing inference, then a weighted voting strategy across all five models (BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT) in the fine-tuning phase. All five models agreed on 133 records (66.5%), while at least four models agreed on 173 records (86.5%). Notably, at least three models agreed on all 200 records (100%), supporting the effectiveness of the ensemble approach in ensuring reliable prediction coverage.

The agreement levels reflect a more calibrated representation of uncertainty compared to a single model prediction.

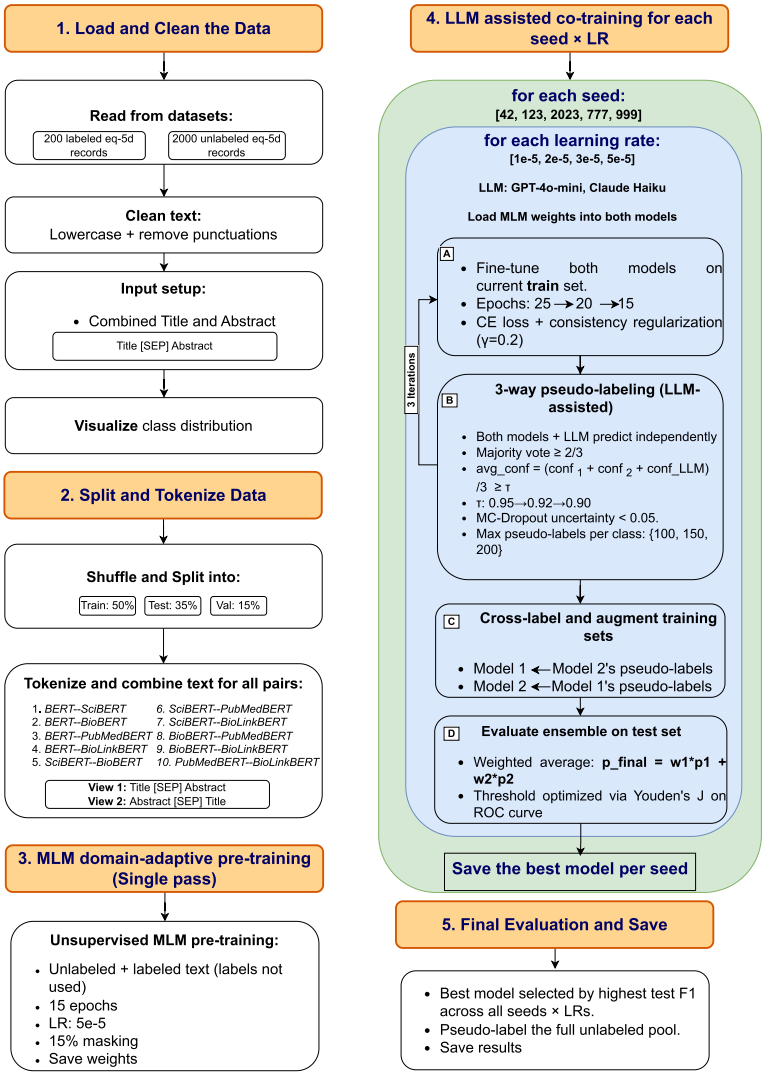

FIGURE 7. Co-training approach with LLM assistance.

Therefore, weighted voting can serve as an effective method for aggregating model outputs, offering a better balance between precision and recall than what can be achieved by a single model.

C. MODELS PERFORMANCE WITH SEMI-SUPERVISED LEARNING

By utilizing this approach, we achieved better results for all five models (BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT) in detecting EQ-5D mentioned in scientific literature by leveraging an expanded dataset through pseudo-labeling on 2,000 unlabeled records using the title and abstract (TA) input configuration.

Table 5 shows the models' results for different evaluation metrics, like accuracy, precision, recall, and F1-score. It also shows the mean F1 ± std and 95% confidence intervals across five seeds. BioLinkBERT achieved the best accuracy (0.83) and F1-score (0.82), with a mean F1 of 0.7805 ± 0.0271. PubMedBERT consistently showed a strong performance with an accuracy of 0.81 and the lowest variance (mean F1 = 0.7881 ± 0.0160). This made it the most stable model. BioBERT also achieved 0.81 accuracy and 0.81 F1-score (mean F1 = 0.7669 ± 0.0321).

SciBERT reached 0.80 accuracy and F1-score of 0.80 (mean F1 = 0.7344 ± 0.0361), while BERT achieved 0.79 accuracy and 0.77 F1-score (mean F1 = 0.7088 ± 0.0415).

Compared to the supervised fine-tuning phase, pseudo-labeling led to better generalization. The five models showed improved performance, with F1-scores ranging from 0.77 to 0.82 on the held-out test set. Using pseudo-labeling in semi-supervised learning improved both accuracy and model robustness, especially when the training data was limited and relatively low.

Table 6 and Figure 8 present the analysis of models' agreement, highlighting the robustness of the semi-supervised

TABLE 3. Model performance summary: Training and Fine-tuning.

Stage	Model	Input	Accuracy	Precision	Recall	F1-score	Mean F1 ± Std	95% CI F1	95% CI Acc
Training	BERT	abstract	0.70	0.74	0.70	0.63	0.6600 ± 0.0200	[0.600, 0.788]	[0.600, 0.800]
		keywords	0.65	0.42	0.65	0.51	0.6500 ± 0.0000	[0.550, 0.750]	[0.550, 0.750]
		title	0.65	0.42	0.65	0.51	0.6500 ± 0.0000	[0.550, 0.750]	[0.550, 0.750]
		combined	0.65	0.42	0.65	0.51	0.6500 ± 0.0000	[0.550, 0.750]	[0.550, 0.750]
	SciBERT	abstract	0.65	0.60	0.65	0.53	0.6375 ± 0.0250	[0.5497, 0.750]	[0.550, 0.750]
		keywords	0.65	0.42	0.65	0.51	0.6500 ± 0.0000	[0.550, 0.750]	[0.550, 0.750]
		title	0.69	0.67	0.69	0.64	0.6575 ± 0.0150	[0.575, 0.788]	[0.575, 0.788]
		combined	0.65	0.42	0.65	0.51	0.6500 ± 0.0000	[0.550, 0.750]	[0.550, 0.750]
	BioBERT	abstract	0.65	0.60	0.65	0.53	0.6500 ± 0.0000	[0.550, 0.750]	[0.550, 0.750]
		keywords	0.65	0.42	0.65	0.51	0.6500 ± 0.0000	[0.550, 0.750]	[0.550, 0.750]
		title	**0.68**	**0.69**	**0.68**	**0.68**	**0.6600 ± 0.0122**	**[0.575, 0.775]**	**[0.575, 0.775]**
		combined	0.65	0.42	0.65	0.51	0.6500 ± 0.0000	[0.550, 0.750]	[0.550, 0.750]
	PubMedBERT	abstract	0.66	0.66	0.66	0.56	0.6525 ± 0.0050	[0.5625, 0.7625]	[0.5625, 0.7625]
		keywords	0.65	0.42	0.65	0.51	0.6500 ± 0.0000	[0.550, 0.750]	[0.550, 0.750]
		title	0.68	0.66	0.68	0.62	0.6425 ± 0.0291	[0.5625, 0.775]	[0.575, 0.775]
		combined	0.65	0.42	0.65	0.51	0.6500 ± 0.0000	[0.550, 0.750]	[0.550, 0.750]
	BioLinkBERT	abstract	0.68	0.70	0.68	0.58	0.6525 ± 0.0122	[0.5625, 0.775]	[0.575, 0.775]
		keywords	0.65	0.42	0.65	0.51	0.6500 ± 0.0000	[0.550, 0.750]	[0.550, 0.750]
		title	0.68	0.66	0.68	0.61	0.6425 ± 0.0331	[0.575, 0.775]	[0.575, 0.763]
		combined	0.65	0.42	0.65	0.51	0.6375 ± 0.0250	[0.550, 0.750]	[0.550, 0.750]
Fine-tuning	BERT	abstract	0.71	0.71	0.71	0.68	0.6775 ± 0.0348	[0.6125, 0.8125]	[0.6125, 0.8125]
		keywords	0.66	0.63	0.66	0.62	0.5775 ± 0.0629	[0.5625, 0.7750]	[0.5625, 0.7625]
		title	0.68	0.66	0.68	0.61	0.6550 ± 0.0203	[0.5750, 0.7625]	[0.5750, 0.7750]
		combined	0.74	0.76	0.74	0.70	0.6400 ± 0.0502	[0.6375, 0.8375]	[0.6375, 0.8375]
		Title and Abstract	0.69	0.68	0.69	0.68	0.6625 ± 0.0177	[0.5875, 0.7875]	[0.5875, 0.7875]
	SciBERT	abstract	0.74	0.74	0.74	0.71	0.6825 ± 0.0281	[0.6375, 0.8250]	[0.6500, 0.8250]
		keywords	0.65	0.42	0.65	0.51	0.6100 ± 0.0406	[0.5500, 0.7500]	[0.5500, 0.7500]
		title	0.74	0.73	0.74	0.72	0.6975 ± 0.0339	[0.6375, 0.8250]	[0.6375, 0.8250]
		combined	0.68	0.66	0.68	0.61	0.6475 ± 0.0215	[0.5625, 0.7750]	[0.5750, 0.7750]
		Title and Abstract	0.68	0.66	0.68	0.65	0.6625 ± 0.0137	[0.5747, 0.775]	[0.575, 0.775]
	BioBERT	abstract	0.74	0.73	0.74	0.72	0.6925 ± 0.0341	[0.6375, 0.8250]	[0.6375, 0.8375]
		keywords	0.65	0.62	0.65	0.61	0.5975 ± 0.0578	[0.5500, 0.7500]	[0.5500, 0.7500]
		title	0.74	0.73	0.74	0.73	0.6750 ± 0.0500	[0.6500, 0.8250]	[0.6375, 0.8250]
		combined	0.66	0.64	0.66	0.58	0.6050 ± 0.0664	[0.5622, 0.7625]	[0.5625, 0.7625]
		Title and Abstract	0.74	0.73	0.74	0.73	0.6825 ± 0.0376	[0.6375, 0.8375]	[0.6500, 0.8375]
	PubMedBERT	abstract	0.74	0.74	0.74	0.71	0.6875 ± 0.0426	[0.6375, 0.8253]	[0.6375, 0.8250]
		keywords	0.65	0.42	0.65	0.51	0.5775 ± 0.0550	[0.5500, 0.7500]	[0.5500, 0.7500]
		title	**0.76**	**0.76**	**0.76**	**0.75**	**0.7250 ± 0.0250**	**[0.6747, 0.8500]**	**[0.6750, 0.8500]**
		combined	0.74	0.73	0.74	0.73	0.6875 ± 0.0326	[0.6375, 0.8250]	[0.6375, 0.8250]
		Title and Abstract	**0.75**	**0.74**	**0.75**	**0.74**	**0.7150 ± 0.0184**	**[0.6500, 0.8375]**	**[0.6500, 0.8500]**
	BioLinkBERT	abstract	0.72	0.72	0.72	0.71	0.6675 ± 0.0430	[0.6250, 0.8250]	[0.6250, 0.8250]
		keywords	0.69	0.70	0.69	0.62	0.6500 ± 0.0237	[0.5875, 0.775]	[0.5875, 0.775]
		title	0.74	0.73	0.74	0.73	0.7175 ± 0.0127	[0.6500, 0.8253]	[0.6375, 0.8375]
		combined	0.70	0.69	0.70	0.67	0.6475 ± 0.0310	[0.600, 0.7875]	[0.5875, 0.800]
		Title and Abstract	0.71	0.72	0.71	0.67	0.6775 ± 0.0330	[0.600, 0.8125]	[0.6125, 0.800]

TABLE 4. Model agreement across five models in the fine-tuning phase.

Agreement Level	Number of Records	Percentage
Full Agreement (5/5)	133	66.5%
At Least Four Agree (4/5)	173	86.5%
At Least Three Agree (3/5)	200	100%
At Least Two Agree (2/5)	200	100%

TABLE 5. Model performance: Semi-supervised learning (Title and Abstract input).

Model	Accuracy	Precision	Recall	F1	Mean F1 ± Std	95% CI F1
BERT	0.79	0.81	0.79	0.77	0.7088 ± 0.0415	[0.6702, 0.8667]
SciBERT	0.80	0.80	0.80	0.80	0.7344 ± 0.0361	[0.7007, 0.8857]
BioBERT	0.81	0.82	0.81	0.81	0.7669 ± 0.0321	[0.7100, 0.8996]
PubMedBERT	0.81	0.82	0.81	0.80	**0.7881 ± 0.0160**	[0.7056, 0.8991]
BioLinkBERT	**0.83**	**0.83**	**0.83**	**0.82**	0.7805 ± 0.0271	**[0.7336, 0.9112]**

learning approach across the 2,000 unlabeled records. All five models agreed on 1,307 records (65.35%), while at least four models agreed on 1,727 records (86.35%). Furthermore, at least three models agreed on all 2,000 records (100%), indicating strong consistency within the ensemble.

These statistics provide further evidence that the models complement each other and indicate that weighted voting based on such agreement can be used to reliably aggregate predictions and increase the confidence in classification.

TABLE 6. Model agreement summary: Semi-supervised approach (2,000 unlabeled records).

Category	Count	Percentage (%)
All 5 Models Agree	1307	65.35
At Least 4 Models Agree	1727	86.35
At Least 3 Models Agree	2000	100
At Least 2 Models Agree	2000	100

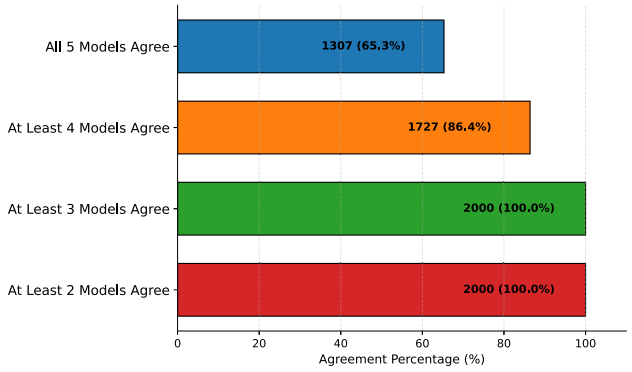

FIGURE 8. Semi-supervised model agreements.

D. MODELS PERFORMANCE WITH CO-TRAINING LEARNING

1) CO-TRAINING APPROACHES WITHOUT LLM

Table 7 presents the summary of the performance of co-training setups after three iterations for pseudo-labeling cycles that involved all ten pairwise combinations of BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT. Performance across all settings appeared enhanced systematically as the iterations progressed, demonstrating the effectiveness of co-training in leveraging the unlabeled data to enhance model robustness and accuracy.

The SciBERT ↔ BioLinkBERT achieved the best performance, reaching an F1-score of 0.85 by the third iteration, which is an improvement from 0.76 in the first iteration. The mean F1 is 0.8366 ± 0.0210 (95% CI: [0.8075, 0.8657]). PubMedBERT ↔ BioLinkBERT had the most consistent performance, with a mean F1 of 0.8375 ± 0.0029 (95% CI: [0.8334, 0.8416]), reaching an F1-score of 0.84 at iteration three. BERT ↔ BioLinkBERT and BioBERT ↔ BioLinkBERT both showed consistent improvement across iterations, attaining an F1-score of 0.84 each by iteration three, with mean F1 of 0.8055 ± 0.0455 and 0.8170 ± 0.0260, respectively. Additionally, pairs with models from the same domain, such as BERT ↔ SciBERT (mean F1 = 0.7487 ± 0.0650) and BioBERT ↔ PubMedBERT (mean F1 = 0.7474 ± 0.0631) showed modest gains. This suggests that model diversity is a key driver of co-training effectiveness.

2) CO-TRAINING APPROACHES WITH GPT

GPT-assisted pseudo-labeling is extended with co-training across all ten pairs (see Table 8). The best performing GPT-assisted setup is GPT-BERT ↔ BioLinkBERT, achieving an F1-score of 0.85 at iteration three with a mean F1 of 0.8217 ± 0.0420 (95% CI: [0.7635, 0.8800]), followed by GPT-BERT ↔ PubMedBERT (F1-score 0.84, mean F1 = 0.8200 ± 0.0250) and GPT-BERT ↔ BioBERT (F1-score 0.83, mean F1 = 0.8114 ± 0.0205). GPT assistance most benefited weaker pairs: GPT-BERT ↔ SciBERT improved from F1-score 0.73 at i1 to 0.82 at i3 (mean F1 = 0.8055 ± 0.0191), and GPT-BioBERT ↔ PubMedBERT from 0.75 to 0.83 (mean F1 = 0.8056 ± 0.0287). However, for pairs already performing strongly without LLM, GPT involvement resulted in slightly lower performance, SciBERT ↔ BioLinkBERT dropped to mean F1 = 0.8013 ± 0.0288 and PubMedBERT ↔ BioLinkBERT to 0.8089 ± 0.0337, compared to 0.8366 ± 0.0210 and 0.8375 ± 0.0029 without LLM, suggesting that high-quality PLM pairs do not benefit from external annotation.

3) CO-TRAINING APPROACHES WITH CLAUDE

Claude Haiku assistance produced comparable results to GPT across most pairs (see Table 9). The best Claude-assisted setup was Claude-BERT ↔ BioBERT, reaching F1-score 0.84 at iteration three with a mean F1 of 0.8216 ± 0.0228 (95% CI: [0.7900, 0.8531]). Claude-BERT ↔ BioLinkBERT and Claude-BioBERT ↔ BioLinkBERT also reached F1-score 0.84 at iteration three. Claude's assistance showed the largest gains for Claude-BERT ↔ BioBERT (from 0.78 at i1 to 0.84 at i3) and Claude-BioBERT ↔ BioLinkBERT (from 0.75 at i1 to 0.84 at i3).

Similar to GPT, Claude's assistance slightly degraded performance for BioLinkBERT-containing pairs that were already strong without LLM involvement. Notably, Claude-SciBERT ↔ BioBERT showed a non-monotonic pattern, peaking at i1 (F1-score 0.81) before dropping at i2 (F1-score 0.73) and partially recovering at i3 (F1-score 0.78), with the lowest variance among all Claude pairs (mean F1 = 0.7743 ± 0.0062, 95% CI: [0.7656, 0.7829]), suggesting instability when Claude annotations conflicted with strong PLM pair agreement.

E. MODEL AGREEMENT AND CONFIDENCE-WEIGHTED VOTING

With the aim of refining pseudo-labeled examples, agreements between co-trained classifier pairs (without LLM, with GPT-4o-mini, and with Claude Haiku) were analyzed across the 2,000 unlabeled records. For each record, the agreement count reflected the number of pairs out of ten that were assigned the same pseudo-label, where the pseudo-label of each pair was derived from the weighted ensemble described in Section VI-B (see (16)). Agreement levels range from full agreement (all 10 pairs) to at least 5 pairs agreeing.

Table 10 summarizes the levels of agreement across all three co-training configurations on the 2,000 unlabeled records. Without LLM involvement, all ten pairs agreed on 1,285 records (64.25%), with at least five pairs agreeing on all 2,000 records (100%). With GPT-4o-mini assistance, full agreement across all ten pairs was observed for

TABLE 7. Co-training performance after 3 iterations (without LLM).

Pair	Iter	Acc	Prec	Rec	F1	Mean F1 ± Std	95% CI F1	95% CI Acc	Mean Acc ± Std
BERT ↔ SciBERT	i1	0.80	0.81	0.80	0.79	0.7487 ± 0.0650	[0.6587, 0.8388]	[0.6520, 0.8480]	0.7500 ± 0.0707
	i2	0.80	0.82	0.80	0.79				
	i3	0.80	0.80	0.80	0.79				
BERT ↔ BioBERT	i1	0.73	0.73	0.73	0.73	0.8040 ± 0.0169	[0.7806, 0.8274]	[0.7863, 0.8423]	0.8143 ± 0.0202
	i2	0.80	0.80	0.80	0.79				
	i3	0.83	0.87	0.83	0.82				
BERT ↔ PubMedBERT	i1	0.76	0.76	0.76	0.76	0.8012 ± 0.0130	[0.7832, 0.8192]	[0.7931, 0.8211]	0.8071 ± 0.0101
	i2	0.79	0.79	0.79	0.78				
	i3	0.81	0.82	0.81	0.81				
BERT ↔ BioLinkBERT	i1	0.81	0.82	0.81	0.81	0.8055 ± 0.0455	[0.7424, 0.8686]	[0.7371, 0.8771]	0.8071 ± 0.0505
	i2	0.84	0.85	0.84	0.84				
	i3	0.84	0.85	0.84	0.84				
SciBERT ↔ BioBERT	i1	0.73	0.73	0.73	0.73	0.7626 ± 0.0454	[0.6996, 0.8255]	[0.6943, 0.8343]	0.7643 ± 0.0505
	i2	0.79	0.82	0.79	0.77				
	i3	0.80	0.80	0.80	0.79				
SciBERT ↔ PubMedBERT	i1	0.74	0.74	0.74	0.73	0.7877 ± 0.0252	[0.7527, 0.8226]	[0.7509, 0.8349]	0.7929 ± 0.0303
	i2	0.81	0.83	0.81	0.81				
	i3	0.81	0.83	0.81	0.81				
SciBERT ↔ BioLinkBERT	i1	0.76	0.76	0.76	0.76	**0.8366 ± 0.0210**	**[0.8075, 0.8657]**	**[0.8149, 0.8709]**	**0.8429 ± 0.0202**
	i2	0.77	0.80	0.77	0.77				
	i3	**0.86**	**0.87**	**0.86**	**0.85**				
BioBERT ↔ PubMedBERT	i1	0.74	0.76	0.74	0.75	0.7474 ± 0.0631	[0.6599, 0.8349]	[0.6520, 0.8480]	0.7500 ± 0.0707
	i2	0.77	0.77	0.77	0.77				
	i3	0.80	0.80	0.80	0.79				
BioBERT ↔ BioLinkBERT	i1	0.79	0.78	0.79	0.78	0.8170 ± 0.0260	[0.7810, 0.8531]	[0.7794, 0.8634]	0.8214 ± 0.0303
	i2	0.83	0.83	0.83	0.82				
	i3	0.84	0.86	0.84	0.84				
PubMedBERT ↔ BioLinkBERT	i1	0.79	0.81	0.79	0.79	**0.8375 ± 0.0029**	**[0.8334, 0.8416]**	**[0.8429, 0.8429]**	**0.8429 ± 0.0000**
	i2	0.81	0.82	0.81	0.82				
	i3	0.84	0.85	0.84	0.84				

1,280 records (64%), and with Claude Haiku for 1,264 records (63.2%). Across all three configurations, at least five pairs agreed on every record (100%), and at least six pairs agreed on 97.6%, 97.35%, and 98% of records for the no-LLM, GPT, and Claude configurations, respectively. Such high levels of agreement support the effectiveness of the confidence-weighted voting scheme in selecting high-quality pseudo-labels for co-training.

VIII. DISCUSSION

Supervised learning results indicated an overfitting of the limited training data. However, the fine-tuning process assisted models in generalizing better, with domain-specific models (PubMedBERT and BioLinkBERT) showing the strongest overall performance. Agreement-based weight voting enhances prediction reliability, especially after fine-tuning. Prediction reliability improved with agreement-based weight voting, particularly after fine-tuning. All five models achieved F1-scores between 0.77 and 0.82, indicating a considerable improvement in overall model performance with semi-supervised learning via pseudo-labeling. Co-training without LLM achieved improvements across all model pairs, with BioLinkBERT-containing pairs attaining the highest performance. Co-training with LLM (GPT and Claude) helped weaker model pairs but slightly degraded performance for already strong combinations, indicating that pseudo-label quality is sensitive to the strength of the underlying PLM pair. Increasing the number of records in the unlabeled pool to 2,000 improved co-training performance, demonstrating the method's ability to scale and its effectiveness.

A. SUPERVISED LEARNING RESULTS

The results of the supervised learning approach revealed a consistent pattern across models: high recall along with relatively low precision during the training phase, which indicated overfitting due to the small size of the labeled data. This overprediction for the positive class also indicated that the models learned dominated labeling patterns rather than general features.

BioBERT demonstrates the most consistent performance among the five PLMs (BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT), attaining the highest F1-score of 0.68 when given title input. BERT also shows encouraging results with abstract input, reaching an F1-score of 0.63, whereas SciBERT's best performance is seen with title input (F1-score 0.64). Furthermore, the domain-specific models, PubMedBERT and BioLinkBERT, achieved the

TABLE 8. Co-training performance after 3 iterations (with GPT-4o-mini).

Pair	Iter	Acc	Prec	Rec	F1	Mean F1 ± Std	95% CI F1	95% CI Acc	Mean Acc ± Std
BERT ↔ SciBERT	i1	0.74	0.74	0.74	0.73	0.8055 ± 0.0191	[0.7791, 0.8320]	[0.7863, 0.8423]	0.8143 ± 0.0202
	i2	0.77	0.77	0.77	0.77				
	i3	0.83	0.85	0.83	0.82				
BERT ↔ BioBERT	i1	0.77	0.77	0.77	0.77	0.8114 ± 0.0205	[0.7829, 0.8398]	[0.7863, 0.8423]	0.8143 ± 0.0202
	i2	0.80	0.80	0.80	0.80				
	i3	0.83	0.83	0.83	0.83				
BERT ↔ PubMedBERT	i1	0.81	0.82	0.81	0.81	0.8200 ± 0.0250	[0.7854, 0.8546]	[0.8006, 0.8566]	0.8286 ± 0.0202
	i2	0.83	0.84	0.83	0.82				
	i3	0.84	0.85	0.84	0.84				
BERT ↔ BioLinkBERT	i1	0.79	0.79	0.79	0.78	**0.8217 ± 0.0420**	**[0.7635, 0.8800]**	**[0.7726, 0.8846]**	**0.8286 ± 0.0404**
	i2	0.81	0.81	0.81	0.81				
	i3	**0.86**	**0.87**	**0.86**	**0.85**				
SciBERT ↔ BioBERT	i1	0.76	0.77	0.76	0.76	0.7880 ± 0.0285	[0.7485, 0.8275]	[0.7509, 0.8349]	0.7929 ± 0.0303
	i2	0.76	0.76	0.76	0.76				
	i3	0.81	0.82	0.81	0.81				
SciBERT ↔ PubMedBERT	i1	0.80	0.82	0.80	0.79	**0.7972 ± 0.0073**	**[0.7871, 0.8073]**	**[0.7931, 0.8211]**	**0.8071 ± 0.0101**
	i2	0.83	0.87	0.83	0.82				
	i3	0.81	0.84	0.81	0.80				
SciBERT ↔ BioLinkBERT	i1	0.79	0.79	0.79	0.78	0.8013 ± 0.0288	[0.7614, 0.8413]	[0.7863, 0.8423]	0.8143 ± 0.0202
	i2	0.81	0.82	0.81	0.82				
	i3	0.83	0.84	0.83	0.82				
BioBERT ↔ PubMedBERT	i1	0.76	0.75	0.76	0.75	0.8056 ± 0.0287	[0.7657, 0.8454]	[0.7863, 0.8423]	0.8143 ± 0.0202
	i2	0.81	0.84	0.81	0.80				
	i3	0.83	0.83	0.83	0.83				
BioBERT ↔ BioLinkBERT	i1	0.81	0.81	0.81	0.81	0.8113 ± 0.0179	[0.7864, 0.8362]	[0.7863, 0.8423]	0.8143 ± 0.0202
	i2	0.80	0.82	0.80	0.80				
	i3	0.83	0.83	0.83	0.82				
PubMedBERT ↔ BioLinkBERT	i1	0.76	0.76	0.76	0.76	0.8089 ± 0.0337	[0.7621, 0.8557]	[0.7583, 0.8703]	0.8143 ± 0.0404
	i2	0.79	0.79	0.79	0.79				
	i3	0.84	0.88	0.84	0.83				

highest training F1-scores of 0.62 and 0.61, respectively, both utilizing title input (see Table 3).

Fine-tuning led to an improvement in the performance of all models, particularly in achieving a more precise balance between precision and recall. PubMedBERT dominated with the best overall F1-score (0.75, mean F1 = 0.7250 ± 0.0250) on title input, followed by BioBERT and BioLinkBERT (both F1-score 0.73 with title input), and SciBERT (F1-score 0.72). BERT achieved its best result with combined input (F1-score 0.70).

Fine-tuning effectively reduced overfitting and improved the generalization ability of all models, with domain-specific models benefiting most from their pre-trained biomedical knowledge. The title and abstract (TA) input configuration also demonstrates strong and consistent performance across all models during fine-tuning, justifying its adoption as the primary input configuration for all subsequent phases (see Table 3).

B. INFERENCE AND AGREEMENT-BASED WEIGHT VOTING

Inference with agreement-based voting demonstrated calibrated uncertainty among the five fine-tuned models across 200 unlabeled studies. All five models achieved full agreement in 66.5% of cases, while a minimum of four models agreed on 86.5% of predictions. Moreover, at least three models agreed on all 200 records (100%), confirming the reliability of the ensemble (see Table 4).

The results indicate enhanced uncertainty calibration relative to a singular model prediction and illustrate the feasibility of weighted voting as an aggregation method that enhances prediction confidence while maintaining a balance between precision and recall.

C. SEMI-SUPERVISED LEARNING RESULTS

Semi-supervised learning with pseudo-labeling enhanced model performance and generalization across all five models using the title and abstract (TA) input configuration on the 2,000 unlabeled records. For instance, BioLinkBERT achieved the best performance in terms of accuracy (0.83) and F1-score (0.82), while PubMedBERT showed the most stable results with the lowest variance among all models (mean F1 = 0.7881 ± 0.0160).

All five models achieved F1-scores between 0.77 and 0.82, which is a substantial improvement over the supervised fine-tuning phase.

Agreement analysis confirmed strong ensemble coherence, with all five models agreeing on 65.35% of the 2,000 unlabeled recordings, and at least three models agreeing

TABLE 9. Co-training performance after 3 iterations (with Claude Haiku).

Pair	Iter	Acc	Prec	Rec	F1	Mean F1 ± Std	95% CI F1	95% CI Acc	Mean Acc ± Std
BERT ↔ SciBERT	i1	0.83	0.83	0.83	0.83	0.7756 ± 0.0188	[0.7496, 0.8017]	[0.7577, 0.8137]	0.7857 ± 0.0202
	i2	0.79	0.79	0.79	0.79				
	i3	0.80	0.82	0.80	0.79				
BERT ↔ BioBERT	i1	0.79	0.80	0.79	0.78	**0.8216 ± 0.0228**	**[0.7900, 0.8531]**	**[0.8006, 0.8566]**	**0.8286 ± 0.0202**
	i2	0.81	0.82	0.81	0.81				
	i3	**0.84**	**0.85**	**0.84**	**0.84**				
BERT ↔ PubMedBERT	i1	0.81	0.82	0.81	0.81	0.8090 ± 0.0239	[0.7758, 0.8421]	[0.7863, 0.8423]	0.8143 ± 0.0202
	i2	0.81	0.82	0.81	0.81				
	i3	0.83	0.83	0.83	0.83				
BERT ↔ BioLinkBERT	i1	0.83	0.84	0.83	0.82	0.7956 ± 0.0428	[0.7363, 0.8549]	[0.7440, 0.8560]	0.8000 ± 0.0404
	i2	0.84	0.84	0.84	0.84				
	i3	0.84	0.84	0.84	0.84				
SciBERT ↔ BioBERT	i1	0.81	0.83	0.81	0.81	**0.7743 ± 0.0062**	**[0.7656, 0.7829]**	**[0.7646, 0.7926]**	**0.7786 ± 0.0101**
	i2	0.73	0.75	0.73	0.73				
	i3	0.79	0.79	0.79	0.78				
SciBERT ↔ PubMedBERT	i1	0.79	0.79	0.79	0.78	0.8001 ± 0.0077	[0.7895, 0.8107]	[0.7931, 0.8211]	0.8071 ± 0.0101
	i2	0.83	0.87	0.83	0.82				
	i3	0.81	0.83	0.81	0.81				
SciBERT ↔ BioLinkBERT	i1	0.77	0.77	0.77	0.77	0.8120 ± 0.0170	[0.7885, 0.8355]	[0.7863, 0.8423]	0.8143 ± 0.0202
	i2	0.80	0.80	0.80	0.80				
	i3	0.83	0.83	0.83	0.82				
BioBERT ↔ PubMedBERT	i1	0.77	0.81	0.77	0.75	0.7996 ± 0.0152	[0.7786, 0.8207]	[0.7931, 0.8211]	0.8071 ± 0.0101
	i2	0.81	0.82	0.81	0.81				
	i3	0.81	0.82	0.81	0.81				
BioBERT ↔ BioLinkBERT	i1	0.76	0.75	0.76	0.75	0.7854 ± 0.0765	[0.6794, 0.8915]	[0.6737, 0.8977]	0.7857 ± 0.0808
	i2	0.81	0.81	0.81	0.81				
	i3	0.84	0.85	0.84	0.84				
PubMedBERT ↔ BioLinkBERT	i1	0.76	0.76	0.76	0.76	0.8093 ± 0.0176	[0.7849, 0.8337]	[0.7863, 0.8423]	0.8143 ± 0.0202
	i2	0.80	0.81	0.80	0.80				
	i3	0.83	0.84	0.83	0.82				

TABLE 10. Model agreement summary across co-training configurations (2,000 unlabeled records, 10 pairs).

(a) Co-training without LLM

Agreement	Count	%
All 10 agree	1285	64.25
At least 9 agree	1563	78.15
At least 8 agree	1713	85.65
At least 7 agree	1848	92.40
At least 6 agree	1952	97.60
At least 5 agree	2000	100

(b) Co-training with GPT-4o-mini

Agreement	Count	%
All 10 agree	1280	64.00
At least 9 agree	1548	77.40
At least 8 agree	1707	85.35
At least 7 agree	1831	91.55
At least 6 agree	1947	97.35
At least 5 agree	2000	100

(c) Co-training with Claude Haiku

Agreement	Count	%
All 10 agree	1264	63.20
At least 9 agree	1550	77.50
At least 8 agree	1700	85.00
At least 7 agree	1844	92.20
At least 6 agree	1960	98.00
At least 5 agree	2000	100

on every record (100%). These findings demonstrated that pseudo-labeling significantly improved classification performance and reliability (see Table 5 and Fig. 8).

D. MODELS PERFORMANCE WITH CO-TRAINING LEARNING

1) EFFECTIVENESS OF CO-TRAINING WITHOUT LLM

In order to ensure mutual gains for models, we utilized all ten pairwise combinations across BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT through iterative pseudo-labeling. Pairs with BioLinkBERT consistently achieved better performance, with SciBERT ↔ BioLinkBERT reaching F1 = 0.85 (mean F1 = 0.8366 ± 0.0210) and PubMedBERT ↔ BioLinkBERT achieving the most stable results (mean F1 = 0.8375 ± 0.0029).

Pairs such as BERT ↔ SciBERT (mean F1 = 0.7487 ± 0.0650) achieved moderate results, suggesting that model diversity is a key driver of co-training effectiveness. These findings present the effectiveness of the co-training approach that utilizes unlabeled data (see Table 7).

2) CO-TRAINING WITH LLM ASSISTANCE

LLM-assisted co-training (GPT-4o-mini and Claude Haiku) achieved improved performance for weaker model pairs across all configurations (see Fig. 9). GPT assistance most benefited BERT ↔ SciBERT (mean F1 improved from 0.7487 to 0.8055) and BioBERT ↔ PubMedBERT (from

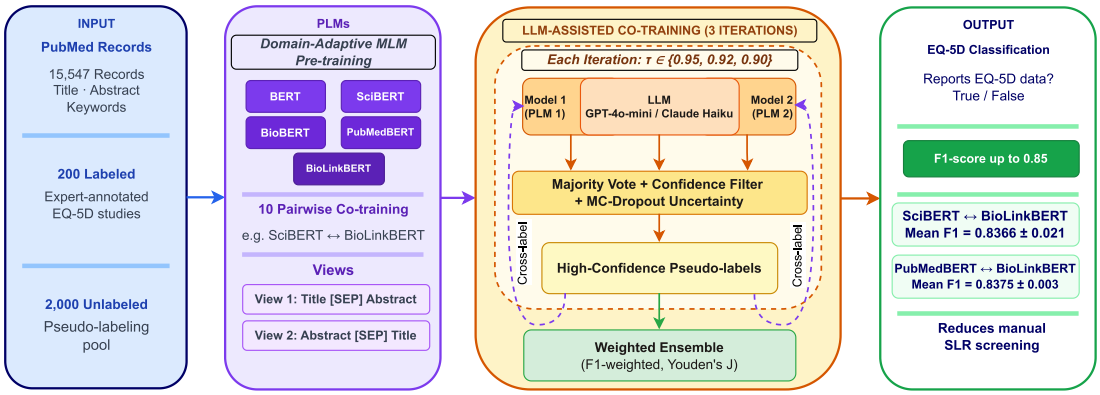

FIGURE 9. Overview of the co-training approach with LLM assistance.

0.7474 to 0.8056). Claude Haiku similarly improved Claude-BERT ↔ BioBERT (mean F1 = 0.8216 ± 0.0228). However, for pairs already achieving strong performance without LLM (e.g. SciBERT ↔ BioLinkBERT and PubMedBERT ↔ BioLinkBERT), both GPT and Claude involvement resulted in slightly lower performance, indicating that high-quality PLM pairs generate reliable pseudo-labels without requiring external annotation. Agreement analysis across the 2,000 unlabeled records showed consistently high consensus, with at least five pairs agreeing on all records in all three configurations (no-LLM, GPT, and Claude), confirming the robustness of the confidence-weighted voting scheme regardless of the annotator used (see Tables 8, 9, and 10).

IX. CONCLUSION AND FUTURE DIRECTIONS

In this study, we developed and evaluated robust and scalable approaches for the automated classification of the EQ-5D-related literature in PubMed. Utilizing PLMs such as BERT, SciBERT, BioBERT, PubMedBERT, and BioLinkBERT, we presented the efficacy of different training approaches: supervised, semi-supervised, and co-training. Also, we investigated different voting strategies such as weighted voting during the inference phase and ensemble confidence-weighted voting to enhance classification accuracy and reliability. Notably, co-training approaches, with and without LLM assistance (GPT-4o-mini and Claude Haiku), enhanced the quality of pseudo-labeling and confidence. The best co-training results were achieved by SciBERT ↔ BioLinkBERT and PubMedBERT ↔ BioLinkBERT (F1 up to 0.85, mean F1 = 0.8375 ± 0.0029). LLM assistance improved weaker model pairs but did not benefit already strong combinations, due to the high confidence threshold applied in the pseudo-labeling process, which limited coverage of the unlabeled dataset. These results suggest that LLMs, when used in conjunction with iterative pseudo-labeling and ensemble learning approaches, are able to reduce the manual effort of systematic literature evaluation in healthcare research, more specifically for the EQ-5D domain. To further enhance scalability, adaptability, and reliability, the following directions are recommended:

- Contrastive learning: Utilizing contrastive learning can assist models in developing semantic relationships between relevant and irrelevant studies, especially in low-resource or noisy data environments.
- Multi-view curriculum learning: Implementing multi-view curriculum learning, progressively feeding inputs (title, abstract, and keywords) to the same model in different shots to enhance model robustness and generalization, especially in co-training.
- Active and few-shot learning: Active learning or few-shot prompting can be incorporated into an LLM to reduce the reliance on large labeled datasets to increase label efficiency.
- Model distillation and efficiency: Investigate model distillation or quantization to reduce computational overhead and allow deployments in real-world resource-constrained setups.
- Generalization: Extending the approach to additional literature sources (e.g., Web of Science, Scopus, EMBASE) to validate robustness across databases.
- Examining adaptive confidence mapping for LLM annotators (GPT and Claude) to enhance the alignment of their static confidence scores with the difficulty distribution of domain-specific relevance tasks.

X. LIMITATIONS

Despite the strong performance achieved by the different approaches, the study has several limitations that should be acknowledged:

- Small labeled dataset: The labeled dataset only contains 200 manually labeled samples, which may compromise the models' ability to generalize.
- Input data: Classification relied only on titles, abstracts, and keywords. Full-text was not utilized during the classification process.

- Language scope: The dataset used for this study was limited to English language metadata.
- Computational demand: Co-training and LLM-assisted pseudo-labeling (GPT-4o-mini and Claude Haiku) are computationally expensive tasks, requiring GPU infrastructure and high memory. It is challenging to reproduce in low-resource settings.
- LLM annotator sensitivity: The performance of LLM assistance degrades for robust PLM pairs (e.g. SciBERT ↔ BioLinkBERT, PubMedBERT ↔ BioLinkBERT), indicating that fixed confidence mapping may not be ideal for all pair configurations.

LIST OF ABBREVIATIONS

AdamW	Adaptive Moment Estimation with Weight Decay.
BERT	Bidirectional Encoder Representations from Transformers.
BioBERT	Biomedical Bidirectional Encoder Representations from Transformers.
BioLinkBERT	Biomedical Link Bidirectional Encoder Representations from Transformers.
BiLSTM-CRF	Bidirectional Long Short-Term Memory with Conditional Random Fields.
CI	Confidence Interval.
DL	Deep Learning.
EQ-5D	EuroQol 5-Dimensions.
FPR	False Positive Rate.
GPT	Generative Pre-trained Transformer.
GPU	Graphics Processing Unit.
HRQoL	Health-Related Quality of Life.
KL	Kullback–Leibler Divergence.
LM	Language Model.
LLM	Large Language Model.
MC	Monte Carlo.
MeSH	Medical Subject Headings.
MLM	Masked Language Modeling.
MSE	Mean Squared Error.
NER	Named Entity Recognition.
NLP	Natural Language Processing.
PAL	Patient Active Learning.
PLM	Pre-trained Language Model.
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
PubMedBERT	PubMed Bidirectional Encoder Representations from Transformers.
QuA	Question Answering.
RAG	Retrieval Augmented Generation.
RE	Relation Extraction.
ROC	Receiver Operating Characteristic.
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve.
SciBERT	Scientific Bidirectional Encoder Representations from Transformers.
SLR	Systematic Literature Review.
SOTA	State of the Art.
SR	Systematic Review.
SVM	Support Vector Machine.
TPR	True Positive Rate.
UMLS	Unified Medical Language System.

ACKNOWLEDGMENT

The authors are thankful to the members of the Applied Machine Learning Research Group at Obuda University's John von Neumann Faculty of Informatics for their valuable comments and suggestions. They also wish to mention the support provided by the Doctoral School of Applied Informatics and Applied Mathematics at Obuda University. Furthermore, on behalf of the "distributed deep learning in parallel and distributed environment" project they are grateful for the possibility to use HUN-REN Cloud (see Héder et al. [33] 2022; https://science-cloud.hu/), which helped them achieve the results published in this article. Disclosure: Márta Péntek is a member of the EuroQol Group. Additionally, they used an AI-based tool for minor text editing and grammar enhancement of the manuscript.

REFERENCES

[1] EuroQol Group, "EuroQol—A new facility for the measurement of health-related quality of life," *Health Policy*, vol. 16, no. 3, pp. 199–208, Dec. 1990.

[2] A. Szende, B. Janssen, and J. Cabases, *Self-Reported Population Health: An International Perspective Based on EQ-5D*. Dordrecht, The Netherlands: Springer, 2014.

[3] G. Kertész, J. T. Czere, Z. Zrubka, L. Gulácsi, and M. Péntek, "Towards automating the selection of articles reporting EQ-5D data for systematic literature reviews using large language models," Social Science Research Network (SSRN), Rochester, NY, USA, Tech. Rep. SSRN ID 4876024, Oct. 2025, doi: 10.2139/ssrn.4876024. [Online]. Available: https://ssrn.com/abstract=4876024

[4] J. White, "PubMed 2.0," *Med. Ref. Services Quart.*, vol. 39, no. 4, pp. 382–387, Oct. 2020.

[5] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou, "Using text mining for study identification in systematic reviews: A systematic review of current approaches," *Systematic Rev.*, vol. 4, no. 1, pp. 1–22, Jan. 2015, doi: 10.1186/2046-4053-4-5.

[6] Z. R. K. Rostam and G. Kertész, "Advancing scientific text classification: Fine-tuned models with dataset expansion and hard-voting," in *Proc. IEEE 23rd World Symp. Appl. Mach. Intell. Informat. (SAMI)*, Jan. 2025, pp. 000527–000532.

[7] R. Alfaro, H. Allende-Cid, and H. Allende, "Multilabel text classification with label-dependent representation," *Appl. Sci.*, vol. 13, no. 6, p. 3594, Mar. 2023.

[8] J. Peng and K. Han, "Survey of pre-trained models for natural language processing," in *Proc. Int. Conf. Electron. Commun., Internet Things Big Data (ICEIB)*, Dec. 2021, pp. 277–280.

[9] U. Naseem, A. G. Dunn, M. Khushi, and J. Kim, "Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT," *BMC Bioinf.*, vol. 23, no. 1, pp. 1–15, Apr. 2022, doi: 10.1186/s12859-022-04688-w.

[10] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: A comprehensive review," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, Apr. 2021.

[11] Q. Jiao, "A brief survey of text classification methods," in *Proc. IEEE 3rd Int. Conf. Inf. Technol., Big Data Artif. Intell. (ICIBA)*, vol. 3, May 2023, pp. 1384–1389.

[12] I. Alimova, E. Tutubalina, and S. I. Nikolenko, "Cross-domain limitations of neural models on biomedical relation classification," *IEEE Access*, vol. 10, pp. 1432–1439, 2022.

[13] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang, "Text classification via large language models," 2023, *arXiv:2305.08377*.
[14] W. X. Zhao, "A survey of large language models," 2023, *arXiv:2303.18223*.
[15] B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, and C. H. Schmid, "Semi-automated screening of biomedical citations for systematic reviews," *BMC Bioinf.*, vol. 11, no. 1, pp. 1–11, Jan. 2010, doi: 10.1186/1471-2105-11-55.
[16] J. Fields, K. Chovanec, and P. Madiraju, "A survey of text classification with transformers: How wide? How large? How long? How accurate? How expensive? How safe?" *IEEE Access*, vol. 12, pp. 6518–6531, 2024.
[17] Z. R. K. Rostam and G. Kertész, "Advances in pre-trained language models for domain-specific text classification: A systematic review," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 6, pp. 1–41, Oct. 2025.
[18] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," 2019, *arXiv:1908.10063*.
[19] L. J. Laki and Z. G. Yang, "Sentiment analysis with neural models for Hungarian," *Acta Polytechnica Hungarica*, vol. 20, no. 5, pp. 109–128, 2023.
[20] P. Kherwa and P. Bansal, "Topic modeling: A comprehensive review," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 7, no. 24, pp. 1–16, Jul. 2019, doi: 10.4108/eai.13-7-2018.159623.
[21] A. L. Lezama-Sánchez, M. Tovar Vidal, and J. A. Reyes-Ortiz, "Integrating text classification into topic discovery using semantic embedding models," *Appl. Sci.*, vol. 13, no. 17, p. 9857, Aug. 2023.
[22] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.
[23] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, pp. 1–23, Oct. 2021.
[24] M. Yasunaga, J. Leskovec, and P. Liang, "LinkBERT: Pretraining language models with document links," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, Dublin, Ireland, May 2022, pp. 8003–8016.
[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.
[26] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," 2019, *arXiv:1903.10676*.
[27] A. M. Cohen, W. R. Hersh, K. Peterson, and P.-Y. Yen, "Reducing workload in systematic review preparation using automated citation classification," *J. Amer. Med. Inform. Assoc.*, vol. 13, no. 2, pp. 206–219, Mar. 2006.
[28] G. Danilov, T. Ishankulov, K. Kotik, Y. Orlov, M. Shifrin, and P. Aa, "The classification of short scientific texts using pretrained BERT model," *Public Health Informat.*, vol. 281, pp. 83–87, May 2021.
[29] A. Landschaft, D. Antweiler, S. Mackay, S. Kugler, S. Rüping, S. Wrobel, T. Höres, and H. Allende-Cid, "Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews," *Int. J. Med. Informat.*, vol. 189, Sep. 2024, Art. no. 105531.
[30] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory*, Madison, WI, USA, Jul. 1998, pp. 92–100.
[31] S. Gao, O. Kotevska, A. Sorokine, and J. B. Christian, "A pre-training and self-training approach for biomedical named entity recognition," *PLoS ONE*, vol. 16, no. 2, Feb. 2021, Art. no. e0246310.
[32] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
[33] M. Héder, E. Rigó, D. Medgyesi, R. Lovas, S. Tenczer, F. Török, A. Farkas, M. Emődi, J. Kadlecsik, G. Mező, Á. Pintér, and P. Kacsuk, "The past, present and future of the ELKH cloud," *Információs Társadalom*, vol. 22, no. 2, pp. 128–137, Aug. 2022.

ZHYAR RZGAR K. ROSTAM (Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from the University of Sulaimani, KRG-Iraq, in 2013 and 2019, respectively. He is currently pursuing the Ph.D. degree in information science and technology with Obuda University, Budapest, Hungary. His current research interests include large language models, scientific text classification, deep learning techniques, machine learning algorithms, and artificial intelligence.

MÁRTA PÉNTEK (Member, IEEE) received the Ph.D. degree in theoretical medical science from Semmelweis University, Budapest, Hungary, in 2008, the Habilitation degree in health sciences from the University of Pécs, Pécs, Hungary, in 2013, and the D.Sc. degree in health sciences from the Hungarian Academy of Sciences, Budapest, in 2022.

She is currently a Professor with the Health Economics Research Center, Obuda University, Budapest, and a part-time Senior Consultant Rheumatologist with the Department of Rheumatology, Flór Ferenc Hospital, Kistarcsa, Hungary. She has coordinated several national and EU-funded international research projects (FIKP, EuroVaQ, and BURQOL-RD), and participated in the HealthPros Marie Skłodowska-Curie Innovative Training Network.

JÁNOS TIBOR CZERE (Member, IEEE) is currently pursuing the Ph.D. degree with the Doctoral School of Innovation Management, Obuda University, Budapest, Hungary.

He has held teaching and IT leadership positions at Kálmán Kando Technical College, Olympus Hungary Ltd., and Danubiana Bt., and oversees IT operations for the Benelux countries, Scandinavia, and Hungary at PSI CRO Hungary LLC. His research interests include artificial intelligence-based virtual assistants in clinical trials, large language models, natural language processing, and prompt engineering in healthcare. He is a member of the Hungarian Economic Society's Section of Health and Health Economics and the Hungarian Artificial Intelligence Coalition. He led the eLearning development team of the EU "Leonardo da Vinci" project "Training of Trainers for Market Manpower Training I–II," whose work received second place at the 2025 European Network for Artificial Intelligence Safety Field Building Excellence Awards.

ZSOMBOR ZRUBKA (Member, IEEE) received the Ph.D. degree in business and management from the Corvinus University of Budapest, Budapest, Hungary, in 2019. He is a Medical Doctor with 17 years of experience in the pharmaceutical industry, including ten years of international experience at Pfizer and Egis Pharmaceuticals in various commercial and medical roles. He is currently an Associate Professor and the Head of the Health Economics Research Center and the University Research and Innovation Center, Obuda University, Budapest.

LÁSZLÓ GULÁCSI (Member, IEEE) received the Ph.D. degrees from the University of Amsterdam, Corvinus University of Budapest, Semmelweis Medical University Budapest, and the Debrecen Medical University. He works as a Vice Rector for Research and a Professor at Health Economics. By profession he is a physician (Debrecen Medical University), having university degrees of programming mathematics (Kossuth Lajos University of Arts and Sciences, Debrecen), mathematical economics and sociology (Corvinus University of Budapest) and health economics (University of York). He is qualified physician in social medicine. Habilitated at the University of Pecs. He is a Doctor of the Hungarian Academy of Science (V. Department of Medical Sciences). The Bulkley-Barry-Cooper Professorship was awarded to him by the London Institute for Healthcare Engineering, King's Health Partners in 2025. He is a Top 1% researcher in Social Sciences and Humanities. To date, he has co-authored 814 scientific articles in peer-reviewed journals, published 16 books, and 78 book chapters. His IF: 446.134; Google Scholar citations: 8347; Hirsch index: 53.

GÁBOR KERTÉSZ (Senior Member, IEEE) received the Ph.D. degree in information science and technology, in 2019.

He is currently an Associate Professor and the Vice-Dean for Research with Obuda University John von Neumann Faculty of Informatics, Budapest, Hungary, and a part-time Research Fellow with the HUN-REN SZTAKI (Institute for Computer Science and Control). He is the Leader of the Applied Machine Learning Research Group, John von Neumann Faculty of Informatics. His main areas of research are computer vision, parallel processing, and deep machine learning. His current research interests include distributed deep learning, metric learning, and applied machine intelligence.

Dr. Kertész is the Founding President of the High Performance Computing Division, John von Neumann Computer Society, and the President of the IEEE Computational Intelligence Hungary Chapter.