Healthcare and Rehabilitation 2 (2026) 100100

Contents lists available at ScienceDirect

Healthcare and Rehabilitation

journal homepage: www.keaipublishing.com/en/journals/hcr/

Full length article

Automatic classification of medical literature leveraging pre-trained large language models

Chenpeng Wang[a], Zhuo Zuo[b], Chenyuyan Yang[c], Shuxia Liu[a], Yaqing Wang[a], Chendong Yang[a], Xinying An[a,*], Meimei Xu[a,*]

[a] Institute of Medical Information & Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China
[b] College of Computer Science, Sichuan University, Chengdu 610065, China
[c] TUM School of Medicine and Health, Technical University of Munich, Munich 80809, Germany

ARTICLE INFO

Keywords:
Clinical research
Medical literature
Text classification
BioBERT
Abstract reformulation
Explainable AI

ABSTRACT

Background: Identifying clinical research articles can facilitate evidence synthesis and research evaluation; however, journal categories and keyword-based approaches may not accurately capture studies at the article level.

Objective: To develop and evaluate a text-based classifier for distinguishing clinical from non-clinical research articles.

Study design: We conducted a bibliographic classification study using Web of Science records published in 2023 by authors affiliated with Chinese hospitals, with temporal evaluation based on records published in 2024 from the same source.

Methods: We assembled 135,932 records published in 2023 and developed a model using a fixed subset of 20,000 records. Bidirectional Encoder Representations from Transformers (BERT), Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT), PubMedBERT, and Biomedical Language Model (BioMedLM) were evaluated using titles and keywords alone, with original abstracts, or with DeepSeek-V3-reformulated abstracts for training and validation as well as original abstracts for testing.

Results: Adding original abstracts improved the mean performance of BERT, BioBERT, and PubMedBERT. BioBERT-D3 demonstrated the highest observed accuracy (0.926), precision (0.926), F1 score (0.919), and AUC (0.972). PubMedBERT-D3 achieved the highest recall (0.922). On the 2024 temporal set, BioBERT-D3 achieved an accuracy of 0.980, a precision of 0.981, a recall of 0.970, an F1 score of 0.976, and an AUC of 0.992.

Conclusions: The framework may support first-pass literature screening and article-level bibliometric evaluation of clinical research output. Clinically ambiguous or mixed-content articles may still require human review. Although the model performed well on records published in 2024 from the same database and institutional setting, additional validation is required using other databases, datasets from different countries, and independently annotated reference labels.

Abbreviations: AUC, area under the receiver operating characteristic curve; BERT, Bidirectional Encoder Representations from Transformers; BioBERT, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; BioMedLM, Biomedical Language Model; LoRA, low-rank adaptation; MeSH, Medical Subject Headings; NIH, National Institutes of Health; PLM, pretrained language model; SHAP, SHapley Additive exPlanations; TF-IDF, term frequency–inverse document frequency; WoS, Web of Science; CI, confidence interval; PMID, PubMed identifier; GPU, graphics processing unit; API, application programming interface; LLM, large language model; SCIE, Science Citation Index Expanded

Peer review under responsibility of Shandong University.

* Correspondence to: Institute of Medical Information & Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, No. 3 Yabao Road, Chaoyang District, Beijing 100020, China.

E-mail addresses: an.xinying@imicams.ac.cn (X. An), xu.meimei@imicams.ac.cn (M. Xu).

https://doi.org/10.1016/j.hcr.2026.100100
Received 20 July 2026; Received in revised form 11 September 2026; Accepted 11 September 2026

3050-6131/© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Clinical research plays a pivotal role in translating basic scientific discoveries into clinical applications, thereby advancing medical practice and improving public health. Consequently, strengthening clinical research infrastructure has emerged as a strategic priority in many countries. In recent years, China has actively promoted the development of clinical research capacity within hospitals through initiatives such as the establishment of clinical research innovation platforms and dedicated capacity-enhancement programs. These national efforts have been accompanied by an increasing demand for systematic evaluation of clinical research performance. Academic publications, as a major form of research output, are therefore frequently used as important indicators of such performance. For instance, the Trial Guidelines for Performance Evaluation of China's Clinical Medical Research Centers include publication output as a core evaluation metric.

Identifying which articles contain clinical research remains challenging. The rapid growth of biomedical research has led to a substantial increase in the volume of scientific literature. PubMed added approximately 1.7 million citations over the preceding year, bringing the database to more than 39 million citations in 2025.[1] This growth has increased the burden of literature screening and research evaluation. Content-based tools are therefore required to identify clinical research articles at scale.[2]

Recently, the development of deep pretrained language models (PLMs) and machine learning techniques has provided new directions for content-based automated classification.[3] Fine-tuned PLMs can capture contextual semantic information beyond simple keyword matching, potentially improving discrimination between clinical and non-clinical research articles.

This study aimed to investigate and evaluate classification methods for medical research articles using domain-specific PLMs. The goal was to develop an efficient and accurate framework for automatic classification of clinical research articles. The main contributions of this study are as follows:

- We evaluated article-level clinical/non-clinical classification in a large hospital-affiliated Web of Science (WoS) corpus.
- We evaluated three input conditions (D1—D3) using a common fixed article partition and conducted a matched D2—D3 comparison to examine performance transfer from a reformulated training and validation representation to original source abstracts.
- We characterized model behavior and clinically relevant failure modes using full-test SHapley Additive exPlanations (SHAP) analysis and clinically oriented error analysis.

1.1. Related Work

Traditional classification approaches typically rely on journal subject categories, author affiliations, or predefined database navigation tags. However, these approaches may lack the granularity required to distinguish clinical from non-clinical research at the article level. Their limited ability to capture article-level semantic information may therefore reduce the accuracy of literature retrieval and research evaluation.[4]

Haynes et al. developed one of the most influential retrieval-oriented methods, proposing a set of empirically validated search filters to identify scientifically sound clinical studies in MEDLINE.[5] These filters were developed using manually annotated gold-standard datasets[6] and optimized through large-scale evaluation of term combinations for sensitivity and specificity. The resulting tool, Clinical Queries, has been integrated into PubMed to support evidence-based literature retrieval. Montori et al. subsequently extended the method to systematic reviews, constructing similarly validated filters for high-precision retrieval.[7] However, this method functions primarily as a retrieval filter rather than an article-level classifier and depends substantially on MEDLINE-specific indexing, which may limit its portability to other databases. Moreover, it is designed primarily to retrieve specific categories of methodologically rigorous clinical studies rather than to distinguish the broader spectrum of clinical from non-clinical research.

Other studies have characterized biomedical research using citation-based or translational-research frameworks. For instance, Ke et al. employed Medical Subject Headings (MeSH) vectorization to represent research content,[8] while Weber's Biomedical Triangle Model provided a macro-level view of the translational trajectory of research outputs.[9] Although these methods have enhanced our understanding of the research translation process, they do not directly enable classification into clinical or basic research types and remain limited in their ability to handle cross-species or cross-modal studies.

Lewison et al.[10] proposed a method for classifying biomedical articles based on 121 "clinical" terms and 107 "basic" terms, categorized by research level. However, this method was developed prior to recent advances in biomedical natural language processing and does not reflect the latest developments in biomedical research. Furthermore, the classification is based solely on article titles, which frequently contain limited information. Hutchins et al.[11] developed a method to predict the translation potential of biomedical research and integrated it into the web-based iCite tool. This tool labels studies as clinical research based solely on whether their PubMed publication type matches predefined categories (e.g., clinical trials, observational studies, or guidelines). Although the tool includes indicators for clinical research, its definition is limited to clinical trials, observational studies, and clinical guidelines.[12]

In recent years, deep PLMs have enabled more context-sensitive approaches to medical literature classification.[3] Preliminary studies, such as those by Luo et al., have explored the use of Bidirectional Encoder Representations from Transformers (BERT) and generative large language models for discipline-level classification, demonstrating promising potential.[13,14] Meanwhile, Chen et al. proposed LitMC-BERT, a fine-tuned multi-label classification model based on BERT that achieved strong performance in the COVID-19 curation task by capturing label-specific features and inter-label correlations.[15] However, these efforts have primarily focused on high-level academic disciplines with limited investigation of the fundamental classification task of distinguishing clinical from non-clinical research within the medical domain.

In contrast to rule-based and metadata-based approaches, contextual language models such as BERT and Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) can capture semantic information beyond explicit keyword matches, providing a methodological basis for article-level classification.

2. Materials and Methods

2.1. Research Workflow

The overall research workflow is shown in Fig. 1 and consists of the following components:

2.1.1. Data Source and Annotation Strategy

The dataset was sourced from the WoS Core Collection through the China Hospital Science and Technology Evaluation Project.[16] Annotations followed the National Institutes of Health (NIH) definition of clinical research and used a three-step procedure: database-based preliminary classification, term-based preliminary tagging, and article-level manual review of all records.

2.1.2. Text Input Strategy Design

Each model was provided with bibliographic information only, comprising the title, author keywords, and, for D2 and D3, abstract text. D1 used titles and keywords alone, whereas D2 additionally incorporated the original abstract. In D3, reformulated abstracts were used during training and validation, while the original abstracts were

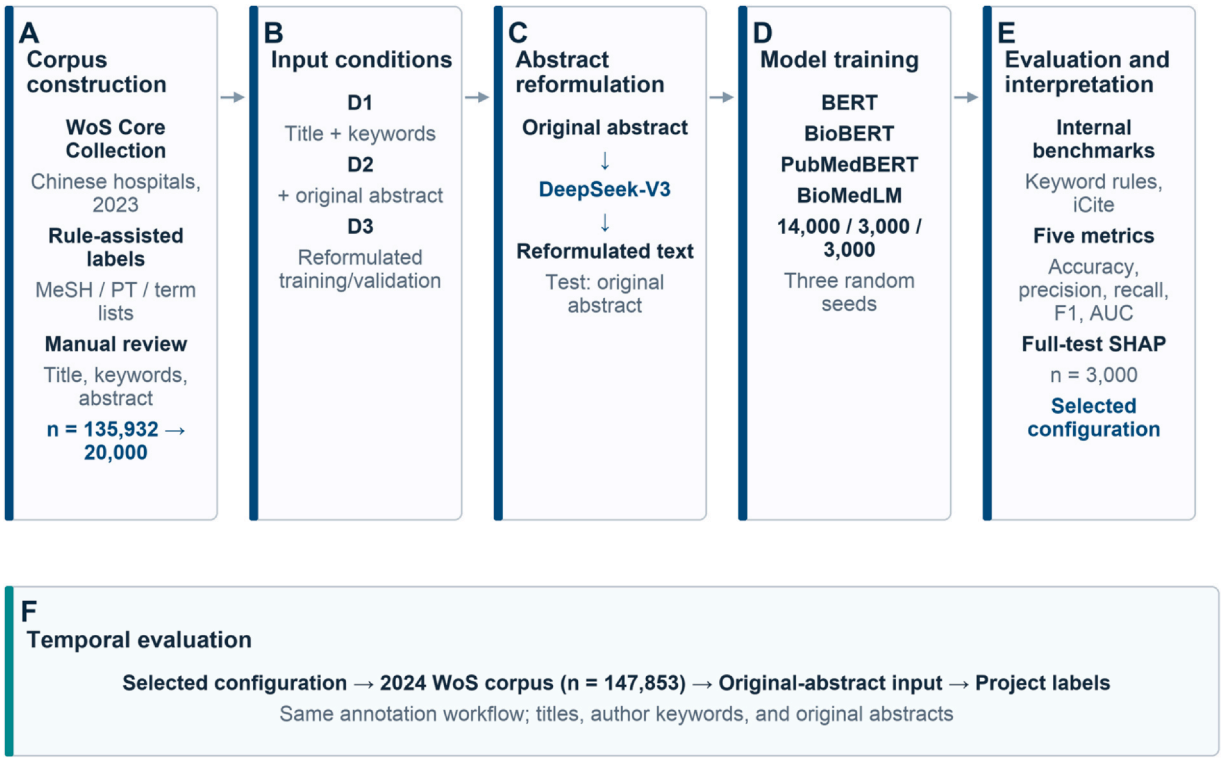


Fig. 1. Overall research workflow for the automatic classification of clinical research articles. (A) Corpus construction and manual review. (B) Matched D1–D3 input conditions. (C) DeepSeek-V3 reformulation for D3 training and validation. (D) Model training on a fixed 14,000/3000/3000 partition using three random seeds. (E) Benchmarking, performance evaluation, and full-test SHAP interpretation. (F) Temporal generalization evaluation on 2024 WoS records using the same database and annotation workflow, with titles, keywords, and original abstracts as inputs.

Abbreviations: AUC, area under the receiver operating characteristic curve; BERT, Bidirectional Encoder Representations from Transformers; BioBERT, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; BioMedLM, Biomedical Language Model; PubMedBERT, BERT model pretrained from scratch on PubMed text; F1, F1 score; MeSH, Medical Subject Headings; n, number of records; PT, publication type; SHAP, SHapley Additive exPlanations; WoS, Web of Science; D1, titles and keywords only; D2, titles, keywords, and original abstracts; D3, reformulated abstracts for training and validation and original abstracts for testing.

used for its primary test set. Therefore, the D2–D3 comparison assessed whether the representation learned from reformulated abstracts could generalize to standard source abstracts at testing, rather than directly comparing reformulated and original abstracts on the same test set. The reformulation prompt was explicitly designed to retain clinically discriminative wording, making D3 an exploratory, prompt-guided compressed representation used during training and validation.

2.1.3. Model Training and Testing

The four models represented different pretraining settings: BERT was pretrained on general-domain English text, BioBERT was further pretrained on biomedical literature, PubMedBERT was pretrained from scratch on PubMed text, and Biomedical Language Model (BioMedLM) was pretrained on biomedical text as a larger generative language model. Each model was adapted to perform the same article-level classification task: determining whether an article was classified as clinical or non-clinical research based on the available text. Performance was evaluated using accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC).

2.1.4. Results Analysis

We compared observed model performance with descriptive rule-based baselines. The BioBERT-D3 from the comparative evaluation was subsequently used for SHAP analysis to summarize how words in the held-out test set shifted model scores toward the clinical or non-clinical class.[17] The same model configuration was then applied to a separately collected corpus from the same project comprising records published in 2024.

2.2. Data Annotation

The WoS Core Collection, maintained by Clarivate, is a widely used bibliographic and citation database. Its standardized bibliographic and citation data are commonly used to assess the scientific output of countries, institutions, and researchers. However, WoS primarily classifies content at the journal level and lacks article-level subject tagging, making it unsuitable for directly identifying clinical research articles. Given its widespread use in science and technology research evaluation, this study selected WoS as the data source for developing a classification framework to distinguish clinical from non-clinical research within the medical domain.

2.2.1. Annotation Procedure

The dataset used in this study was derived from an existing corpus developed as part of the China Hospital Science and Technology Evaluation Project. The corpus comprised 135,932 articles published in 2023 by Chinese hospitals and indexed in the WoS Core Collection.

Clinical research was defined according to the NIH as research involving human participants, including epidemiologic, behavioral, health services, and clinical trial research.[18] In this study, an article was classified as clinical when the research it reported met this definition; otherwise, it was classified as non-clinical.

Multiple research personnel from the Institute of Medical Information & Medical Library, Chinese Academy of Medical Sciences, participated in annotation. The annotators had backgrounds in medicine and health information science. The quality of annotations was routinely monitored during the broader project. Historical quality-control records revealed annotation error rates of no more than 2%.

Database information and term-based rules were first used to generate preliminary classifications, followed by a manual review of all

records using the title, author keywords, and abstract to determine the final clinical or non-clinical label. The annotation procedure included the following steps:

(1) Index-Based Filtering

Because academic evaluation primarily focuses on original research, the classification target for this study was original research articles. Science Citation Index Expanded (SCIE)-indexed publications are frequently used in academic evaluation; however, the "article" document type in the WoS database may include both original research and review articles. Accordingly, for this original-research classification task, review articles were operationally classified as non-clinical.

Given the overlap between WoS and PubMed in indexed literature, we used PubMed's MeSH indexing for supplementary classification. Articles tagged with clinical study-related publication types and involving humans only (excluding animals) were preliminarily identified as clinical research. On the contrary, articles tagged with in vitro research-related MeSH terms and classified under non-clinical publication types were preliminarily labeled as non-clinical research. The complete PubMed search strategies are provided in Appendix A (Table A.1).

(2) Term-Based Tagging

Articles with exclusionary terms in their titles or abstracts, such as meta-analysis, systematic review, systematic evaluation, or similar terms, were preliminarily tagged as non-clinical. Articles mentioning animal-related terms (e.g., mouse, dog, rabbit, primate, monkey, baboon, chimpanzee, Xenopus, zebrafish, Drosophila, wax moth), but making no reference to humans, were preliminarily tagged as non-clinical.

A list of terms indicative of clinical research was developed by reviewing existing clinical studies and the terms proposed by Lewison et al.[10] Articles with clinical-specific terms in their titles or the methods and results sections of abstracts as well as human-related terminology were preliminarily tagged as clinical.

(3) Manual Review

All records subsequently underwent manual review at the article level. Using the title, author keywords, and abstract, annotators verified or revised the preliminary classification according to the definition of clinical research described above. Articles reporting epidemiologic, behavioral, health services, or clinical trial research involving human participants were classified as clinical. In contrast, articles that did not meet the operational criteria were classified as non-clinical. In cases where an article contained both clinical and basic-research components, it was classified as clinical.

2.3. Final Dataset Statistics

After annotation, the full corpus consisted of 135,932 records published in 2023, including 58,270 clinical and 77,662 non-clinical articles. For model development, we used a randomly sampled 20,000-record subset of the full corpus, containing 9034 clinical and 10,966 non-clinical articles. It was divided into fixed 14,000/3000/3000 training, validation, and test sets. The subset size was selected for computational feasibility while retaining a sample larger than those used in many published BERT fine-tuning studies of scientific or biomedical text classification and literature screening.[19,20]

2.4. Model Selection

The four models were selected to provide a focused comparison across distinct model families and biomedical pretraining strategies, rather than an exhaustive benchmark of all available biomedical language models. The comparison involved a general-domain encoder (BERT), two biomedical-literature encoders with different pretraining strategies (BioBERT and PubMedBERT[21]), and a larger BioMedLM.

ClinicalBERT-style models were excluded as additional encoder baselines because they were pretrained primarily on electronic health record notes rather than scientific literature,[22] whereas the current task uses article titles, author keywords, and abstracts. The resulting model set was therefore designed to investigate the relevance of pretraining domain and model family to this article-level classification task. The four selected models, their pretraining backgrounds, and the rationale for their inclusion are summarized in Table 1.

2.5. Data Preprocessing

For each condition, the relevant text fields were combined in a consistent order and formatted for the model. Long inputs were limited to the model's maximum input length; these processing steps did not change the article labels.

2.5.1. Input Representations and Dataset Splitting

To systematically evaluate the effect of text representation, we constructed three input conditions using a common fixed 14,000/3000/3000 training–validation–test partition. The exact input used at each stage is specified in Table 2.

D1 used titles and keywords for training, validation, and testing. D2 used titles, keywords, and the original abstracts for training, validation, and testing. D3 used titles, keywords, and DeepSeek-V3 reformulations for training and validation, whereas titles, keywords, and original abstracts were used for testing.

No titles, author keywords, or abstracts were missing in the 20,000-record model-development subset. Before sampling, records with missing or whitespace-only values for titles, author keywords, abstracts, or reference labels were excluded. No text was imputed. All titles, author keywords, and abstracts in the model-development subset were in English.

D2 and D3 were tested using the same original abstracts. The comparison therefore assessed the transferability of the representation

Table 1
Four Pretrained Language Models Selected for Comparison.

Model	Brief Description	Reason for Inclusion
BERT[23]	A general-purpose bidirectional Transformer encoder pretrained on large-scale English corpora; widely adopted as a strong NLP baseline	Provides a domain-agnostic performance reference
BioBERT[24]	Shares BERT's architecture but is further pretrained on PubMed abstracts and PMC full texts, giving it enhanced sensitivity to biomedical terminology and context	Represents domain-specific pretraining and tests its added value for medical-text classification
PubMedBERT[21]	A BERT-style encoder pretrained from scratch on PubMed abstracts using a domain-specific vocabulary	Represents in-domain pretraining from scratch and tests a literature-oriented biomedical encoder against continued-pretraining BioBERT
BioMedLM[25]	A GPT-style generative model (2.7 billion parameters) trained on PubMed data, capable of both language understanding and generation	Evaluates the potential and limitations of large generative language models on discriminative tasks

Abbreviations: BERT, Bidirectional Encoder Representations from Transformers; BioBERT, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; BioMedLM, Biomedical Language Model; PubMedBERT, BERT model pretrained from scratch on PubMed text; GPT, generative pre-trained transformer; NLP, natural language processing; PMC, PubMed Central.

Table 2
Definition of the D1–D3 input conditions. All conditions used the same fixed article-level partition. Original abstract refers to the source abstract. Reformulated abstract refers to the DeepSeek-V3 reformulation used only in D3 training and validation.

Condition	Training input	Validation input	Test input
D1	Title + keywords	Title + keywords	Title + keywords
D2	Title + keywords + original abstract	Title + keywords + original abstract	Title + keywords + original abstract
D3	Title + keywords + reformulated abstract	Title + keywords + reformulated abstract	Title + keywords + original abstract

Abbreviations: D1, titles and keywords only; D2, titles, keywords, and original abstracts; D3, reformulated abstracts for training and validation and original abstracts for testing.

learned from reformulated abstracts during D3 training, rather than comparing original and reformulated text at test time.

DeepSeek-V3 was used to produce the reformulations for D3. Its purpose was to standardize training and validation representations, not to assign article labels.[26,27] DeepSeek-V3 was prompted to generate an English reformulation of the source abstract. The full prompt is provided in Appendix C.[28]

2.5.2. Reformulation Fidelity

Reformulation fidelity was evaluated using paired original (D2) and reformulated (D3) abstracts from the training and validation partitions (n = 17,000). Automatic similarity measures included token-overlap F1, term frequency–inverse document frequency (TF-IDF) cosine similarity, and numeral recall. A stratified random sample of 100 pairs (40 clinical and 40 non-clinical training records; 10 clinical and 10 non-clinical validation records) was reviewed at the document level by one rater. Each pair was rated on a scale of 1–5 for faithfulness and completeness.[27]

2.5.3. Tokenizer Configuration and Text Concatenation

Each input was converted into tokens and limited by the model's maximum input length. BERT, BioBERT, and PubMedBERT all used their native encoder format, while BioMedLM used its native generative-model format.[29,30] These computational steps affected only how text was packaged for the models, not the clinical versus non-clinical labels.

2.5.4. Adapting the Models for Classification

To assess whether model type was relevant, BERT, BioBERT, PubMedBERT, and BioMedLM were trained under the same D1–D3 input conditions. The main configuration choices are listed in Appendix B (Table A.2).

Fine-tuning adapts a pretrained model to the clinical/non-clinical decision using labeled training records.[31] For the larger BioMedLM, low-rank adaptation (LoRA)[32] updates a small set of added parameters rather than the entire model; mixed-precision training was used to reduce memory usage. These choices did not change the texts, labels, or classification task. Each model assigned clinical and non-clinical probabilities, with the highest probability determining the prediction. Epoch-wise training and validation losses for BioMedLM on D2 and D3 are provided in Appendix B (Figure B.1 A, D2; Figure B.1 B, D3). For BioMedLM, a fixed LoRA configuration was used, and LoRA hyperparameters were not systematically optimized in the present study.

2.5.5. Implementation Details

All experiments were implemented using the Hugging Face Transformers library (version 4.32.0) and PyTorch 2.5.1 on a graphics processing unit (GPU)-enabled virtual machine.

Appendix B (Table A.2) reports the training settings used for reproducibility. The substantive comparison in this study concerned the input conditions and the observed performance of the selected models, rather than optimizing across every possible hyperparameter configuration.

To ensure reproducibility, the training, validation, and test partitions were fixed across all configurations. Each configuration was trained three times using prespecified random seeds 42, 123, and 2025 for model initialization and training-data shuffling, while the validation and held-out test records remained unchanged. PyTorch and NumPy were initialized with the same seed within each run, and mean metrics across the three runs were reported.

2.6. Evaluation Metrics and Practical Interpretation

Model performance was evaluated using accuracy, precision, recall, F1 score, and AUC. In the current analysis, accuracy is defined as the proportion of correctly classified test articles; precision refers to the proportion of articles predicted to be clinical that were labeled as clinical; recall (sensitivity) is the proportion of labeled clinical articles that the model identifies; the F1 score combines precision and recall in a single measure; finally, AUC measures how well the model ranks clinical articles above non-clinical articles across classification thresholds.

For first-pass screening, recall can be prioritized to reduce the number of clinical studies missed, with false positives removed by human reviewers. False negatives may prevent relevant clinical articles from subsequent review, whereas false positives increase the burden of manual verification and inflate estimates of clinical-research output. When the aim is to estimate clinical-research output or review capacity is limited, precision may be more important. Accuracy, F1 score, and AUC are complementary measures that do not replace inspection of both error types.

Following model development, the BioBERT-D3 was applied to 147,853 unique 2024 WoS records from Chinese hospital-affiliated institutions in the same evaluation project. The labels used the same NIH definition and title–keyword–abstract review as in Section 2.2.1 (61,353 clinical; 86,500 non-clinical).

2.7. Clinically Oriented Error Analysis

To investigate the clinical significance of model errors, we conducted a qualitative review of false-positive and false-negative predictions made by the BioBERT-D3 on the held-out test set. Two false-positive and two false-negative records were selected to represent four distinct error contexts: an animal-only study, a review article, a randomized phase III clinical-trial subgroup analysis, and a mixed clinical-experimental study. For each record, we examined the apparent study design, the involvement of human participants or specimens, the clinical outcomes or research aims relevant to the operational definition, and the most likely source of the discrepancy. This analysis was intended to identify clinically meaningful boundary conditions for model use and was not designed to estimate corpus-level error frequencies.

3. Results

3.1. Model Comparison

BERT, BioBERT, PubMedBERT, and BioMedLM were evaluated using the D1, D2, and D3 input conditions, with reported values representing mean test-set metrics from three prespecified random-seed runs (42, 123, and 2025; Fig. 2). Results for each seed are provided in Appendix B (Table A.3). Under D1, PubMedBERT showed the highest mean accuracy, F1 score, and AUC among the four models. Adding

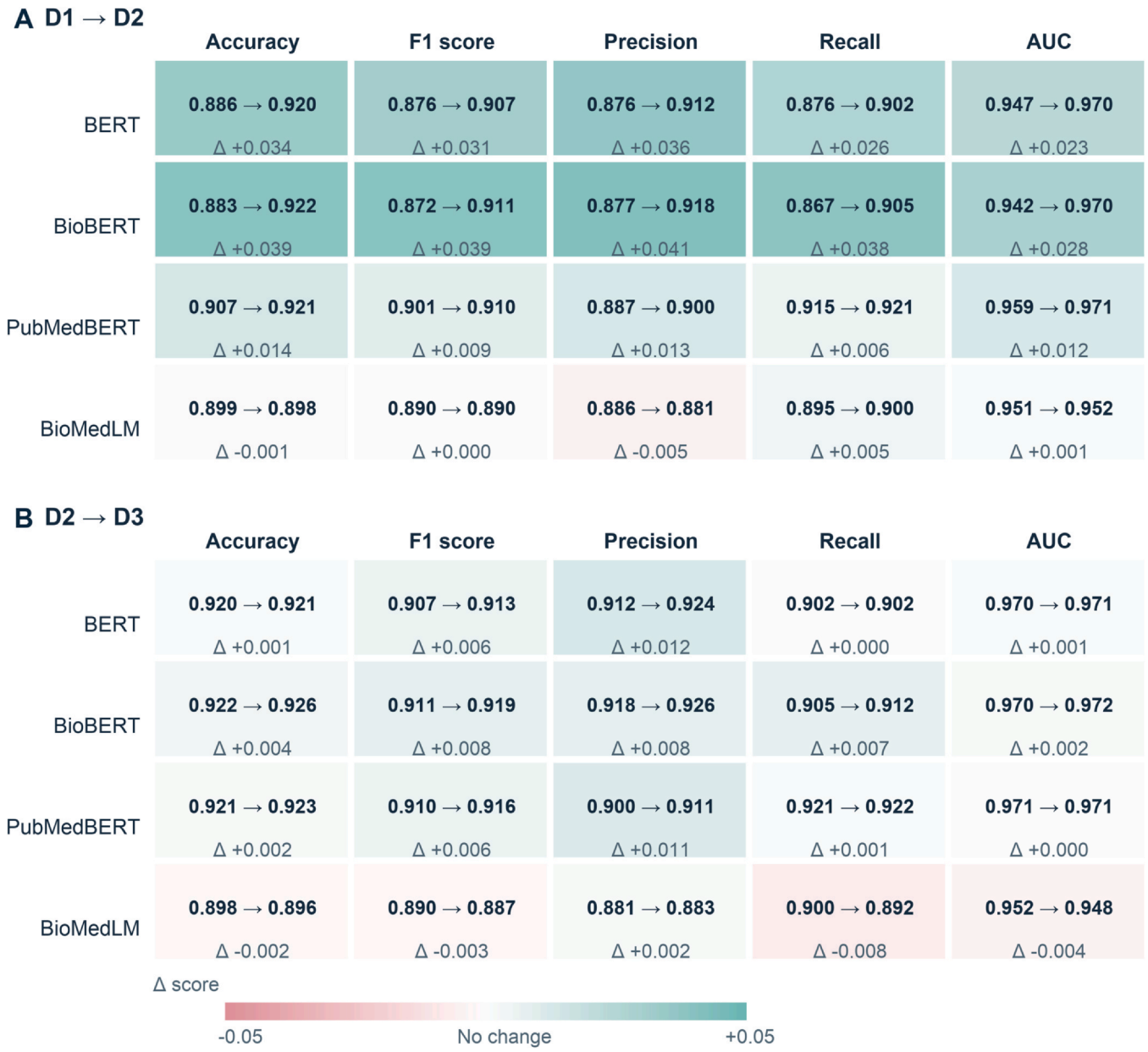


Fig. 2. Performance comparison of BERT, BioBERT, PubMedBERT, and BioMedLM. (A) Change from D1 to D2. (B) Change from D2 to D3. Cell fill encodes the signed score change; printed values show the mean source and target scores across three prespecified random-seed runs. D2 and D3 were evaluated on the same original-abstract test set. Abbreviations: AUC, area under the receiver operating characteristic curve; BERT, Bidirectional Encoder Representations from Transformers; BioBERT, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; BioMedLM, Biomedical Language Model; PubMedBERT, BERT model pretrained from scratch on PubMed text; F1, F1 score; Δ, change in score; D1, titles and keywords only; D2, titles, keywords, and original abstracts; D3, reformulated abstracts for training and validation and original abstracts for testing.

original abstracts in D2 was associated with higher performance for BERT, BioBERT, and PubMedBERT, whereas BioMedLM showed minimal overall improvement.

Under D3, BioBERT achieved the highest mean accuracy (0.926), precision (0.926), F1 score (0.919), and AUC (0.972), whereas PubMedBERT achieved the highest recall (0.922). Compared to D2, BioBERT showed higher mean values across all five metrics. PubMedBERT showed higher accuracy, precision, F1 score, and recall and an unchanged AUC. BioMedLM did not show a similar improvement. Formal inferential comparisons were not conducted; these differences are therefore interpreted descriptively. For BioMedLM, Figure B.1 in Appendix B shows the training and validation loss trajectories for random seed 42, with panel A corresponding to D2 and panel B to D3.

BioBERT-D3 was selected for subsequent error analysis, SHAP interpretation, internal benchmarking, and temporal evaluation because it had the highest observed mean accuracy, precision, F1 score, and AUC on the held-out test set. This selection was descriptive and did not establish overall superiority; PubMedBERT-D3 had the highest observed mean recall. Table 3 provides illustrative classification results.

3.2. Benchmark Comparison

BioBERT-D3 was compared descriptively with a keyword-based baseline and an iCite-derived baseline on the held-out test set. Error bars shown in Fig. 3 are Wilson 95% confidence intervals (CIs) for accuracy, precision, and recall, reflecting the sampling uncertainty of these estimates on the 3,000-record held-out test set. The keyword-based method yielded an accuracy of

Table 3
Illustrative Classification Results.

Title	Label	P(1)	P(0)
Mimicking Antioxidases and Hyaluronan Synthase: A Zwitterionic Nanozyme for Photothermal Therapy of Osteoarthritis	0	0.80%	99.20%
Quercetin influences intestinal dysbacteriosis and delays alveolar epithelial cell senescence by regulating PTEN/PI3K/AKT signaling in pulmonary fibrosis	0	0.98%	99.02%
Effects and mechanisms of trifluridine alone or in combination with cryptotanshinone in inhibiting malignant biological behavior of gastric cancer	0	0.87%	99.13%
Clinical outcomes of unilateral biportal endoscopic lumbar interbody fusion (ULIF) compared with conventional posterior lumbar interbody fusion (PLIF)	1	99.20%	0.80%
Effect of susceptibility-weighted imaging on preoperative evaluation of microvascular decompression in patients with trigeminal neuralgia	1	99.02%	0.98%

Note: Label 1 indicates clinical research and label 0 indicates non-clinical research. P(1) and P(0) denote the predicted probabilities of the clinical and non-clinical classes, respectively.

0.464 (95% CI, 0.446–0.482) and an F1 score of 0.600. The iCite-derived method achieved a recall of 1.000 but a substantially lower precision of 0.459 (95% CI, 0.441–0.477), with an F1 score of 0.629. BioBERT-D3 achieved an accuracy of 0.926 (95% CI, 0.916–0.935), precision of 0.926 (95% CI, 0.911–0.939), recall of 0.912 (95% CI, 0.896–0.926), an F1 score of 0.919, and an AUC of 0.972.

These comparisons are descriptive internal benchmarks because the methods used unequal input information and the preliminary labels incorporated information related to the comparator methods. They should therefore not be interpreted as independent external validation or evidence of statistical superiority.

3.3. Clinically Oriented Error Analysis

Table 4 lists two false-positive and two false-negative predictions selected to span distinct clinical contexts and methodological boundary conditions. The aim was not to estimate error frequencies but to identify situations in which article-level screening may require additional human review.

The false-positive cases illustrate how strong clinical terminology can lead to a clinical prediction even when the study does not meet the operational definition of clinical research. For example, the canine stent-retrieval study was conducted exclusively in animals, but its use of terms related to stent implantation, retrieval, angiography, vascular injury, and follow-up may have contributed to its clinical prediction. Similarly, the pulmonary sarcoidosis article used extensive diagnostic, treatment, and follow-up terminology, but it was a review that synthesized previously published evidence rather than reporting new research involving human participants.

The false-negative cases show different ways in which clinically relevant studies may be missed. The CheckMate 649 subgroup analysis explicitly reported randomization, treatment groups, survival outcomes, and safety assessment; however, the model assigned a very low clinical probability. The Semaphorin 3 A study combined patient blood and patient-derived cells with extensive mechanistic and experimental content; these competing signals may have contributed to its non-clinical prediction.

Taken together, these representative cases suggest that errors may arise when clinical terminology appears in non-clinical study designs, such as animal experiments or review articles, and when clinical information is embedded within mixed clinical-experimental research. Such cases may therefore merit targeted human review when the model is used for first-pass literature screening or bibliometric assessment.

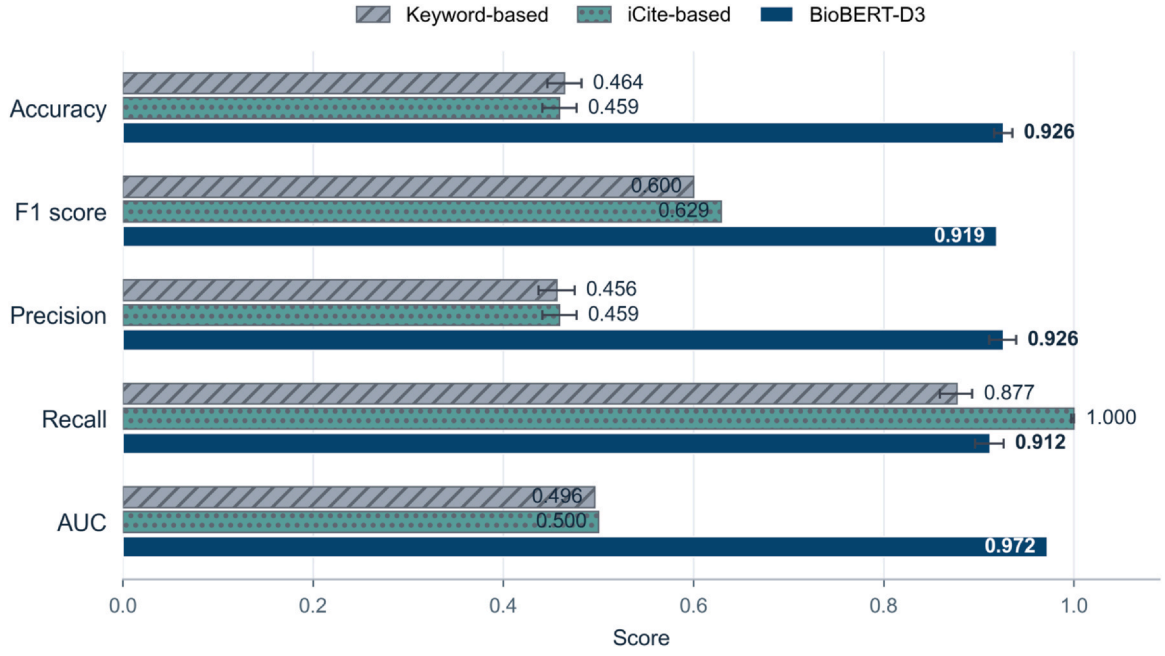


Fig. 3. Descriptive internal-test comparison of BioBERT-D3 with keyword-based and iCite-derived baselines. Error bars show 95% confidence intervals for accuracy, precision, and recall on the 3,000-record held-out test.
Abbreviations: AUC, area under the receiver operating characteristic curve; BioBERT, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; D3, reformulated abstracts for training and validation and original abstracts for testing; F1, F1 score.

Table 4
Illustrative BioBERT-D3 misclassifications by clinical study context and plausible source of error.

Error direction and clinical context	Representative article	P(clinical)	Clinical interpretation and plausible source of error
False positive; animal-only study	Preliminary studies on intimal injury related to stent retrieval in a canine model (PMID: 37290284)	0.509	Clinical procedural terminology in an animal-only study. The abstract contains terms related to stent implantation, retrieval, angiography, vascular injury, and follow-up, although the experiment was conducted exclusively in canines.
False positive; review article	Pulmonary Function in Pulmonary Sarcoidosis (PMID: 37959167)	0.896	Clinical-content/document-type mismatch. The abstract contains extensive diagnosis-, treatment-, and follow-up-related terminology, although the article synthesizes previously published evidence rather than reporting new research involving human participants.
False negative; randomized phase III clinical trial subgroup	First-line nivolumab plus chemotherapy vs chemotherapy in patients with advanced gastric, gastroesophageal junction and esophageal adenocarcinoma: CheckMate 649 Chinese subgroup analysis (PMID: 36121651)	0.026	Missed explicit clinical design. Randomization, treatment groups, survival outcomes, and safety assessment were present; however, the model assigned a low clinical probability.
False negative; mixed clinical–experimental study	Semaphorin 3 A restores the ability of type 1 regulatory T cells to suppress food allergy (PMID: 37999823)	0.233	Competing clinical and experimental signals. Patient blood and patient-derived cells support the clinical label, whereas mechanistic experimental framing may have dominated the prediction.

Abbreviations: BioBERT, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; D3, reformulated abstracts for training and validation and original abstracts for testing; PMID, PubMed identifier; P(clinical), predicted probability of the clinical class.

3.4. Full-Test SHAP Analysis

For a frozen BioBERT-D3 checkpoint, SHAP values were computed for all 3000 records in the principal D3 test condition, which used title, keywords, and abstract. Tokens representing the same reconstructed word were combined, and the resulting signed contributions were averaged across occurrences. A positive value indicates that the word shifted the score toward the clinical class, whereas a negative value indicates a shift toward the non-clinical class. For a readable aggregate display, Fig. 4 shows recurrent content words occurring at least 100 times with the largest positive or negative mean contribution, after exclusion of punctuation, special tokens, numerals, common function words, and non-lexical boundary artifacts.

Fig. 4 shows word-level SHAP contributions across all 3000 records in the principal D3 test condition. Patient- and study-design-related wording generally shifted scores toward the clinical class: the most positively contributing recurrent words included “participants,” “retrospectively,” “subjects,” “enrolled,” “cohort,” “case,” “retrospective,” “underwent,” and “patients.” In contrast, review- and animal-related wording shifted scores toward the non-clinical class: "review" and "literature" showed the largest negative average contributions, followed by "meta," "mouse/mice," "rat/rats," and "Cochrane."

These aggregate patterns are consistent with the clinical/non-clinical distinction used in this corpus and supported the clinically oriented error analysis in Table 4. They should be interpreted as context-

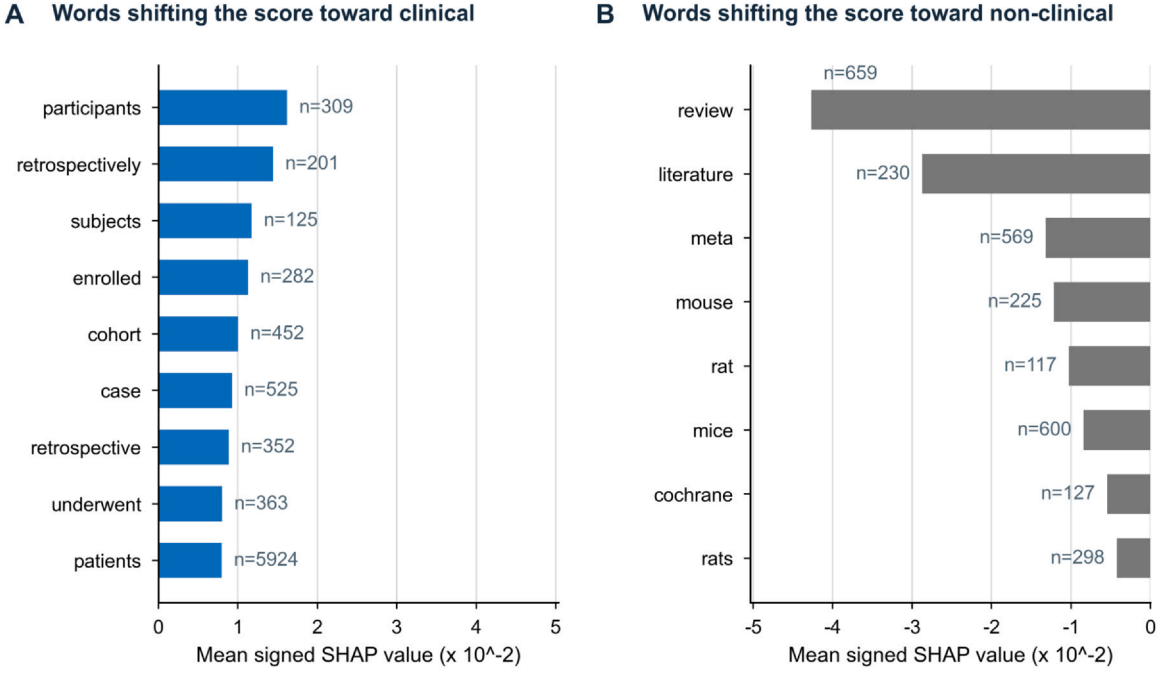


Fig. 4. Full-test SHAP summary for BioBERT-D3 across 3000 D3 test records. (A) Words shifting scores toward the clinical class. (B) Words shifting scores toward the non-clinical class. Both panels use the same scale. Shown are recurrent content words (≥100 occurrences) with the largest mean contributions; “n” indicates word occurrences. Abbreviations: BioBERT, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; D3, reformulated abstracts for training and validation and original abstracts for testing; SHAP, SHapley Additive exPlanations; n, number of word occurrences.

dependent descriptions of how BioBERT-D3 used textual cues, rather than fixed decision rules for single words.

Because SHAP was calculated for BioBERT-D3 rather than as a paired D2–D3 comparison, these results characterize the features used by the D3 model but do not demonstrate that reformulation increased sensitivity to clinical features.

3.5. Reformulation Fidelity

Across 17,000 pairs, the mean TF-IDF cosine similarity was 0.912 (median, 1.000), and the mean token-overlap F1 score was 0.781. Compressed abstracts were shorter than their matched originals, with a mean compressed-to-original length ratio of 0.734. Among pairs containing numerals in the source abstract (n = 15,272), the mean numeral recall was 0.882. Overall, the D3 abstracts retained substantial textual overlap and most numerical information from the source abstracts while being considerably shorter.

In the 100-pair review, the reformulated and source abstracts were evaluated to convey the same meaning in nearly all cases. The median faithfulness score was 5 (73 pairs scored 5, 18 scored 4, 8 scored 3, and 1 scored 2). All 100 pairs received a completeness score of 5. A contradiction with the source abstract was identified in one pair.

3.6. Temporal Generalization Within the Same Database and Annotation Workflow

Table 5 reports the performance of the trained BioBERT-D3 model on records published in 2024. Accuracy was 0.980, precision was 0.981, recall was 0.970, F1 score was 0.976, and AUC was 0.992. These values are not directly comparable with the mean BioBERT-D3 performance across three random-seed runs on the 3,000-record test set from 2023 (accuracy, 0.926; precision, 0.926; recall, 0.912; F1 score, 0.919; AUC, 0.972) because the temporal evaluation used a single trained checkpoint and a full-year corpus from a later publication year. The higher observed scores in 2024 may reflect differences in corpus composition or classification difficulty between the two evaluation sets; these explanations were not directly tested. The shared database and annotation workflow may also have limited distribution shift across years, so this analysis should be interpreted as a temporal generalization check within that setting.

4. Discussion

4.1. Principal Findings

In this corpus, adding abstracts was associated with better observed performance for BERT, BioBERT, and PubMedBERT than titles and keywords alone. BioBERT-D3 achieved the highest observed mean accuracy, precision, F1 score, and AUC. PubMedBERT-D3 achieved the highest recall. The higher observed recall of PubMedBERT-D3 may be advantageous for first-pass screening, whereas the higher observed precision of BioBERT-D3 may be advantageous when false positives would inflate estimates of clinical research output. For BioBERT, the D2–D3 comparison is consistent with a possible transfer benefit from the reformulated representation used during D3 training and validation. However, the observed differences are descriptive and do not demonstrate a statistically significant effect.

The practical value of the framework is as a first-stage aid rather than a replacement for human review. [33–35] In literature screening, it may help prioritize records for subsequent review; in bibliometric evaluation, it can provide article-level estimates of clinical research output when journal-level categories are too coarse. The error analysis identified contexts that may require targeted review, including animal-only studies with strong clinical terminology, review articles, clinical-trial reports with atypical textual signals, and mixed clinical–experimental studies.

The BioMedLM loss curves provide additional diagnostic context for the lower D3 than D2 accuracy. Both conditions relied on the same original-abstract test set. Under D2, validation loss decreased until epoch 7, whereas under D3, training loss continued to decrease but validation loss increased after epoch 1. This pattern, together with the corresponding test accuracies (0.905 for D2 and 0.891 for D3), is consistent with a possible mismatch between the reformulated training representation and the original-abstract test representation. The 100-pair fidelity review showed generally high agreement between reformulated and original abstracts, suggesting that the observed difference is unlikely to be explained solely by loss of source meaning.

4.2. Comparison With Prior Work

The present results are broadly consistent with prior studies showing the value of biomedical pretraining for text classification.[36,37] In this study, biomedical language models were assessed for article-level identification of clinical research in a hospital-affiliated bibliometric corpus, together with a large language model (LLM)-based reformulated training representation and model-interpretation analyses. The full-test SHAP analysis showed that participant- and study-design-related terms generally shifted predictions toward the clinical class, whereas review- and animal-related terms shifted predictions toward the non-clinical class. The error analysis further showed that mixed clinical-experimental articles could present competing textual signals. In the descriptive internal benchmark comparison, BioBERT-D3 showed higher observed accuracy, precision, F1 score, and AUC than the keyword-based and iCite-derived comparators. However, these comparisons involved unequal input information and should not be interpreted as evidence of statistical superiority.

4.3. Limitations

Several limitations should be noted. The 20,000-record subset was selected based on practical feasibility, without formal power or saturation assessment. Because preliminary candidate labels were informed by database fields and term-based rules, the keyword-based and iCite comparisons should be viewed as internal benchmarks within this corpus rather than independent external validation. Although all records underwent article-level manual review, the reference labels were not established through independent blinded duplicate annotation. Inter-rater agreement statistics are therefore unavailable, and residual annotation error cannot be excluded. The binary formulation was designed to determine whether an article met the operational definition of clinical research; it does not distinguish finer-grained study types, disease areas, or methodological categories. Such distinctions could be addressed in future multi-label extensions. BioMedLM was evaluated using LoRA-based adaptation because full-parameter fine-tuning exceeded the available hardware capacity. The LoRA hyperparameters were not systematically optimized; therefore, the observed BioMedLM

Table 5
Temporal generalization of BioBERT-D3 on 2024 WoS records within the same database and annotation workflow.

Item	Value
Source	WoS Core Collection, Chinese hospital-affiliated institutions
Publication year	2024
Records	147,853
Clinical / non-clinical labels	61,353 / 86,500
Input	Title + keywords + original abstract
Accuracy	0.980
Precision	0.981
Recall	0.970
F1 score	0.976
AUC	0.992

Abbreviations: AUC, area under the receiver operating characteristic curve; BioBERT, Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; D3, reformulated abstracts for training and validation and original abstracts for testing; F1, F1 score; WoS, Web of Science.

performance should be interpreted as specific to the configuration evaluated here rather than as representing the optimal performance of BioMedLM.

Compression was applied to the training and validation abstracts, reducing the proportion of training inputs exceeding 512 tokens from 41.1% to 0.19%. Although the fidelity analyses indicated substantial overall agreement with the source abstracts, some source details were not always retained, and rare semantic distortions cannot be excluded. The document-level fidelity review was conducted by a single rater on 100 abstract pairs; therefore, inter-rater reliability was not assessed. In addition, the D3 design did not include a manually extracted methods-and-results control, a length-matched non-LLM control, or reformulations generated by alternative LLMs. Because the prompt explicitly prioritized retention of clinically discriminative wording, the observed D2–D3 differences cannot isolate the contributions of task-guided content retention, semantic reformulation, and reduced truncation. Only DeepSeek-V3 was used, and reliance on a third-party model service may limit reproducibility because model versions, application programming interface (API) behavior, availability, and cost may change over time.

4.4. Generalizability and Future Validation

The evaluation of records published in 2024 supports temporal generalization within the same database, institutional setting, and annotation workflow. Because the classifiers use article-level textual inputs rather than database-specific classification fields, transfer beyond the present dataset may be possible. However, performance may still be affected by differences across databases, countries, and institutional settings in corpus composition, study-type distributions, language use, and text-field availability. All titles, author keywords, and abstracts in the present model-development subset were in English. Therefore, this study does not establish model performance for non-English or multilingual literature. Direct validation using records from PubMed and Embase, as well as independently annotated datasets from other countries and languages, is required before broader generalizability can be established.

5. Conclusions

This study developed and evaluated an applied framework for classifying clinical and non-clinical research articles in a hospital-affiliated bibliometric corpus. Within the present corpus, BioBERT-D3 produced the highest observed mean accuracy, precision, F1 score, and AUC, whereas PubMedBERT-D3 produced the observed highest mean recall. These descriptive comparisons do not establish overall model superiority. Model choice should therefore depend on the intended use. PubMedBERT-D3 may be more suitable when maximizing recall is important for first-pass screening, whereas BioBERT-D3 may be more suitable when higher precision is preferred for bibliometric estimation of clinical research output. D3 should be interpreted as an exploratory, prompt-guided compressed representation used during training and validation, and the observed differences do not establish that reformulation itself improves classification performance. The framework may support first-pass literature screening and article-level assessment of clinical research output, but these potential uses require future validation against independently annotated reference labels. Clinically ambiguous or mixed-content articles require targeted human review. The 2024 evaluation supports temporal generalization within the same database and annotation workflow. Further validation is required across other databases, countries, institutional settings, and languages.

CRediT authorship contribution statement

Shuxia Liu: Data curation. **Chenyuyan Yang:** Data curation. **Zhuo Zuo:** Methodology, Investigation. **Chenpeng Wang:** Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Meimei Xu:** Writing – review & editing, Methodology, Investigation. **Xinying An:** Writing – review & editing, Funding acquisition. **Chendong Yang:** Data curation. **Yaqing Wang:** Data curation. All authors read and approved the final manuscript.

Ethics statement

This study used bibliographic metadata and did not involve human participants, patient-level data, or personally identifiable information; institutional ethical approval and informed consent were therefore not applicable.

Funding

This study was supported by the Non-communicable Chronic Diseases – National Science and Technology Major Project (2023ZD509703–04, 2023ZD509705–03); the Fundamental Research Funds for the Central Public Welfare Research Institutes (2024-ZHCH630–02); and the CAMS Initiative for Innovative Medicine (2021-I2M-1–033). The funders had no role in study design, data collection, data analysis, interpretation of results, manuscript preparation, or the decision to submit the manuscript for publication.

Data availability

The source records were obtained from the Web of Science Core Collection under institutional access and cannot be publicly redistributed because of database licensing restrictions. De-identified derived data and analysis code may be made available by the corresponding authors upon reasonable request, subject to institutional and licensing restrictions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of Generative AI and AI-assisted technologies in the writing process

During research data processing, the authors used DeepSeek-V3 to generate reformulations of complete source abstracts for use as D3 model inputs; this use formed part of the study methodology. During manuscript preparation, the authors used OpenAI Codex for language editing, consistency checking, formatting, and assistance in organizing revisions. The authors reviewed and edited all outputs, verified the manuscript against the underlying data and analysis, and take full responsibility for the content of the published article.

Acknowledgements

We thank the research personnel from the Institute of Medical Information & Medical Library, Chinese Academy of Medical Sciences, for their contributions to article-level annotation and annotation quality control.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi: 10.1016/j.hcr.2026.100100.

References

1. Sayers EW, Bolton EE, Fine AM, et al. Database resources of the National Center for Biotechnology Information in 2026. *Nucleic Acids Res.* 2026;54(D1):D20–D27. https://doi.org/10.1093/nar/gkaf1060
2. Liu CL, Chen S. A review on the clinical impact identification and prediction of academic papers from the perspective of scientometrics. *Libr Inf Serv.* 2024;68(5):132–148. https://doi.org/10.13266/j.issn.0252-3116.2024.05.013
3. Wang BY, Xie QQ, Pei JH, et al. Pre-trained language models in biomedical domain: a systematic survey. *ACM Comput Surv.* 2024;56(3):1–52. https://doi.org/10.1145/3611651
4. Li XY, Liu Y, Li AH, Yang XM, Tang XL. Research on the subject mapping methods optimization of multi-sources biomedical literatures. *Digit Libr Forum.* 2023;19(3):1–9. https://doi.org/10.3772/j.issn.1673-2286.2023.03.001
5. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR. Hedges Team. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. *BMJ.* 2005;330(7501):1179. https://doi.org/10.1136/bmj.38446.498542.8F
6. Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. *Stud Health Technol Inform.* 2001;84(pt 1):390–393. https://doi.org/10.3233/978-1-60750-928-8-390
7. Montori VM, Wilczynski NL, Morgan D, Haynes RB. Hedges Team. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. *BMJ.* 2005;330(7482):68. https://doi.org/10.1136/bmj.38336.804167.47
8. Ke Q. Identifying translational science through embeddings of controlled vocabularies. *J Am Med Inform Assoc.* 2019;26(6):516–523. https://doi.org/10.1093/jamia/ocy177
9. Weber GM. Identifying translational science within the triangle of biomedicine. *J Transl Med.* 2013;11:126. https://doi.org/10.1186/1479-5876-11-126
10. Lewison G, Paraje G. The classification of biomedical journals by research level. *Scientometrics.* 2004;60(2):145–157. https://doi.org/10.1023/B:SCIE.0000027677.79173.b8
11. Hutchins BI, Yuan X, Anderson JM, Santangelo GM. Relative Citation Ratio (RCR): a new metric that uses citation rates to measure influence at the article level. *PLoS Biol.* 2016;14(9):e1002541. https://doi.org/10.1371/journal.pbio.1002541
12. Hutchins BI, Davis MT, Meseroll RA, Santangelo GM. Predicting translational progress in biomedical research. *PLoS Biol.* 2019;17(10):e3000416. https://doi.org/10.1371/journal.pbio.3000416
13. Luo PC, Wang JM, Nie L. Research on automatic classification of literature resources based on generative large language model. *Inf Stud:Theory Appl.* 2024;47(12):174–182.
14. Luo PC, Wang YB, Wang JM. Automatic discipline classification for scientific papers based on a deep pre-training language model. *J China Soc Sci Tech Inf.* 2020;39(10):1046–1059.
15. Chen Q, Du J, Allot A, Lu Z. LitMC-BERT: transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation. *IEEE/ACM Trans Comput Biol Bioinform.* 2022;19(5):2584–2595. https://doi.org/10.1109/TCBB.2022.3173562
16. Institute of Medical Information, Chinese Academy Medical Sciences. Science Technology Evaluation Metrics (STEM). Published July 19 2024. Accessed April 29, 2026. https://stem.pumc.edu.cn/2024/index.htm.
17. Zhou F, Saha A, Afzal M, et al. Understanding transformer-based classifications of medical text using a large language model for the attribution of feature importance: proof-of-concept algorithm development and validation study. *JMIR Med Inform.* 2026;14:e81644. https://doi.org/10.2196/81644
18. National Institutes of Health. NIH Definitions. Accessed April 27, 2025. https://orwh.od.nih.gov/toolkit/nih-policies-inclusion/definitions.
19. van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nat Mach Intell.* 2021;3(2):125–133. https://doi.org/10.1038/s42256-020-00287-7
20. Khadhraoui M, Bellaaj H, Ammar MB, Hamam H, Jmaiel M. Survey of BERT-base models for scientific text classification: COVID-19 case study. *Appl Sci (Basel).* 2022;12(6):2891. https://doi.org/10.3390/app12062891
21. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc.* 2022;3(1):1–23. https://doi.org/10.1145/3458754
22. Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical BERT embeddings. *In: Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2019:72–78. https://doi.org/10.18653/v1/W19-1909
23. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1(Long and Short Papers)*. Association for Computational Linguistics; 2019:4171–4186. https://doi.org/10.18653/v1/N19-1423
24. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* 2020;36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682
25. Bolton E, Venigalla A, Yasunaga M, et al. BioMedLM: a 2.7B parameter language model trained on biomedical text. arXiv. Preprint posted online March 27, 2024. https://doi.org/10.48550/arXiv.2403.18421
26. Liu A, Feng B, Xue B, et al. DeepSeek-V3 technical report. arXiv. Preprint posted online December 27, 2024. https://doi.org/10.48550/arXiv.2412.19437
27. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med.* 2024;30(4):1134–1142. https://doi.org/10.1038/s41591-024-02855-5
28. DeepSeek Prompt Libr | DeepSeek API Docs. Accessed April 27, 2025.
29. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. *In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics; 2020:38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6
30. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. *Language models are unsupervised multitask learners.* OpenAI; 2019. Accessed April 27, 2026. https://cdn.openai.com/better-language-models/language-models.pdf.
31. Tinn R, Cheng H, Gu Y, et al. Fine-tuning large neural language models for biomedical natural language processing. *Patterns (N Y).* 2023;4(4):100729. https://doi.org/10.1016/j.patter.2023.100729
32. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations (ICLR 2022). OpenReview.net; 2022. Accessed April 27, 2026. https://openreview.net/forum?id=nZeVKeeFYf9
33. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res Synth Methods.* 2024;15(4):616–626. https://doi.org/10.1002/jrsm.1715
34. Lieberum JL, Toews M, Metzendorf MI, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review. *J Clin Epidemiol.* 2025;181:111746. https://doi.org/10.1016/j.jclinepi.2025.111746
35. Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. *Syst Rev.* 2024;13(1):158. https://doi.org/10.1186/s13643-024-02575-4
36. Emilova Doneva S, de Viragh S, Hubarava H, Schandelmaier S, Briel M, Ineichen BV. StudyTypeTeller-large language models to automatically classify research study types for systematic reviews. *Res Synth Methods.* 2025;16(6):1005–1024. https://doi.org/10.1017/rsm.2025.10031
37. Tran H, Yang Z, Yao Z, Yu H. BioInstruct: instruction tuning of large language models for biomedical natural language processing. *J Am Med Inform Assoc.* 2024;31(9):1821–1832. https://doi.org/10.1093/jamia/ocae122