

Review

Transformer-Based Language Models for Clinical Decision Support Using Clinical Notes: A Scoping Review

Saahoon Hong¹  and Hunhui Na^{2,*}

¹ School of Social Work, Indiana University, Indianapolis, IN 46202, USA; saahong@iu.edu

² Division of Games, The University of Utah, Salt Lake City, UT 84122, USA

* Correspondence: hunhui.na@utah.edu

Highlights

What are the main findings?

- This scoping review mapped 36 empirical studies published between 2023 and 2025 by primary application purpose, care setting, transformer model approach, and reported evaluation evidence.
- Information extraction/de-identification and classification/prediction predominated, whereas summarization/generation was less commonly represented. Evidence beyond retrospective task performance was significantly limited: only a small number of studies reported an explicit direct human comparator, fairness or subgroup analyses, workflow evaluation, or clinical deployment.
- No included study evaluated a transformer-based language model application within a behavioral-health service, behavioral-health workflow, or service-level outcome.

What are the implications of the main findings?

- Key gaps include limited external generalizability, hallucination risk in generative outputs, sparse explainability, and insufficient attention to bias, equity, and real-world clinical deployment.
- Future research should prioritize prospective, multi-site evaluations, and ethically grounded explainable human–AI partnership models to ensure safe, equitable, and clinically meaningful use of transformer-based language models in behavioral healthcare.

Abstract

Background/Objectives: This scoping review examined recent evidence on the use of transformer-based language models, encompassing encoder-only architectures (e.g., BERT and its clinical variants) and generative large language models (LLMs; e.g., GPT-4 and Llama), to support clinical decision making from unstructured clinical notes, with implications for behavioral-health services where narrative documentation is central. **Methods:** Following PRISMA-ScR guidelines, PubMed, PsycINFO, and Web of Science were searched for peer-reviewed studies published between 1 January 2023, and 5 August 2025. Studies applying transformer-based language models to clinical narratives for healthcare tasks and reporting evaluative outcomes were included. We extracted data on clinical tasks, model architectures, enhancement strategies, and evaluation metrics; mapped each study by primary purpose, care setting, and primary model approach; and charted reported validation design, direct human comparison, fairness assessment, workflow evaluation, and clinical deployment. **Results:** Thirty-six studies were included. Information extraction/de-identification and classification/prediction predominated, whereas summarization/generation was less commonly represented. Model approaches appeared to



Academic Editor: Maria Carmen Carnero

Received: 19 June 2026

Revised: 16 July 2026

Accepted: 20 July 2026

Published: 22 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

align with task characteristics: encoder-only and decoder-only systems were frequently used for extraction, encoder–decoder systems for generation, and hybrid or pipeline-based approaches for classification and prediction. Standard task-specific metrics (e.g., F1 and AU-ROC) predominated, whereas evidence beyond retrospective task performance, including direct human comparison, fairness assessment, workflow evaluation, clinical deployment, and temporal or external validation, was rare. No included study evaluated a transformer-based language model application within a behavioral-health service or behavioral-health workflow. Conclusions: Transformer-based language models have been applied across diverse clinical-note tasks, but the evidence base more strongly supports retrospective task feasibility than transportability, equitable performance, workflow benefit, or safe clinical deployment. Future research should prioritize transparent reference standards, external and prospective validation, clinically meaningful human comparison, and equity-focused evaluation, including direct studies in behavioral-health services.

Keywords: transformer-based language models; generative AI; clinical decision support; clinical notes; behavioral health

1. Introduction

The continued expansion of Certified Community Behavioral Health Centers (CCBHCs), which are designed to deliver integrated, evidence-based mental health and substance use treatment, has simultaneously heightened expectations for accountability and demonstrable program effectiveness [1]. Yet, the current twenty-one CCBHC quality measures remain predominantly process-oriented, emphasizing screening practices, service provision, and timeliness of initial evaluations, with only a single indicator, remission from depression, reflecting a treatment outcome [2]. Existing accountability frameworks further fall short by relying heavily on claims data embedded within internal control and operational systems, offering limited visibility into clinician-level practices and patient-centered recovery outcomes, particularly for individuals presenting with complex co-occurring psychiatric and substance use disorders [2,3]. These challenges are compounded by the substantial operational constraints within the behavioral-health workforce faces, including high caseloads, resource deficits, and insufficient training in the nuances of evidence-based care [4,5]. This environment highlights a pressing demand for advanced decision-support tools that can augment clinical precision, support highly individualized treatment planning, and facilitate ongoing performance monitoring.

Although behavioral health provides an important motivating context because care relies heavily on longitudinal narrative documentation, the scope of this review extends across healthcare rather than behavioral health alone. Accordingly, behavioral health is presented as one clinical domain that may particularly benefit from transformer-based language models, rather than as the primary focus of the current evidence base.

Transformer-based language models have emerged as a promising technology in this space, capable of synthesizing both structured and unstructured patient data to inform more timely and nuanced care decisions [6,7]. The necessity of this technology becomes clear when considering the vast amount of underutilized data in clinical records: unstructured clinical text, including physician notes, intake forms, and care summaries, constitutes over 80% of Electronic Health Record (EHR) content [2]. Transformer-based language models are particularly well suited to this context. Encoder-only transformers, such as BERT and clinical or biomedical variants, can create contextual representations that are commonly used for discriminative tasks including information extraction and classification. Decoder-only transformers and other generative large language models (LLMs) can

generate text and support prompt-specified tasks, while encoder–decoder transformers are commonly used for sequence-to-sequence generation. In addition, some evaluated systems combine language models with retrieval, structured EHR data, rules, or graph-based components [8–10].

Despite its considerable potential, the deployment of these models in behavioral healthcare introduced several critical challenges. For instance, generative systems are susceptible to hallucinations, outputs that appear coherent yet are factually incorrect, which might pose substantial clinical risks when such errors go undetected [7,11,12]. Furthermore, direct human comparison and workflow or clinical-deployment evaluation in reporting the model performance are often absent, limiting the ability to determine how systems perform relative to clinicians, across settings, and in routine care [13,14]. These considerations are particularly relevant to potential behavioral-health applications, where documentation is longitudinal and sensitive.

Although several reviews have broadly examined the use of language models in medicine [13–15], none has systematically mapped the application of transformer-based language models to unstructured clinical notes for clinical decision support by simultaneously considering the application purpose, clinical setting, primary model approach, and implementation evidence. Such a synthesis is particularly important given the breadth, heterogeneity, and rapid evolution of this literature. Unlike previous reviews, which primarily summarize clinical applications or overall model performance, the present review distinguishes technical feasibility from clinical implementation readiness by systematically characterizing validation design, direct human comparison, fairness assessment, workflow evaluation, and clinical deployment. A scoping review is therefore well suited to map the distribution of applications and implementation evidence across this emerging field rather than to pool performance estimates or determine the superiority of a particular model family. Accordingly, the purpose of this review was to examine how transformer-based language models have been applied to unstructured clinical notes for clinical decision support and to systematically characterize the current evidence regarding their applications, validation, implementation, and remaining evidence gaps. We addressed two research questions:

RQ1: For what major purposes have transformer-based language models been applied to unstructured clinical notes, and how are these applications distributed across mapped care settings and primary model approaches?

RQ2: What validation, human-comparator, fairness, workflow, and clinical-deployment evidence has been reported for these applications?

For RQ1, we map each study by primary purpose (i.e., information extraction/de-identification, classification/prediction, or summarization/generation) across care settings and primary transformer model approaches. For RQ2, we distinguish the context of reported performance values from evidence of validation beyond the test dataset: human comparison, fairness assessment, workflow evaluation, and clinical deployment. By consolidating findings across studies, the review clarifies the capabilities and limitations of current systems for analyzing clinical notes and supporting clinical decision processes. The resulting evidence map can inform future research and support practitioners and policymakers in assessing whether, and under what conditions, transformer-based language models are appropriate for clinical decision support, including future behavioral-health applications.

2. Methods

2.1. Search Strategy

This scoping review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Re-

views (PRISMA-ScR) guidelines [16]. The completed PRISMA-ScR checklist is provided in Supplementary File. Comprehensive searches were conducted in PubMed, PsycINFO, and Web of Science to identify peer-reviewed studies published between 1 January 2023 and 5 August 2025. The start date was selected to capture the rapid expansion of clinical applications following the public release of ChatGPT 3.5 rather than to review the historical evolution of transformer architectures. Consequently, this focused scope prioritizes recent applied research while acknowledging that earlier foundational studies of encoder-only transformers (e.g., BERT and its clinical variants) established many of the methodological advances on which current generative and hybrid approaches are built.

Searches were restricted to the title and abstract fields and employed a concept-based search strategy that combined terms related to clinical documentation and clinical decision support with broad concepts related to artificial intelligence and transformer-based natural language processing. Rather than enumerating individual model names (e.g., ClinicalBERT, BioBERT, Med-PaLM, or Llama) or implementation techniques (e.g., retrieval-augmented generation or instruction tuning), the search emphasized overarching concepts to remain robust to the rapidly evolving terminology of AI models. Consequently, all transformer-based architectures, including encoder-only, decoder-only, encoder–decoder, and hybrid systems, were eligible for inclusion regardless of their specific model names, provided they met the predefined eligibility criteria. The clinical-documentation search terms included “clinical note”, “clinical document”, and “clinical decision”, while the technology-related terms included “generative AI”, “generative artificial intelligence”, “natural language processing”, and “transformer”. Boolean operators were used to combine synonymous terms with OR and the two main concept groups with AND. For instance, the PubMed search strategy was as follows: (“clinical note” OR “clinical document” OR “clinical decision”) AND (“generative AI” OR “generative artificial intelligence” OR “natural language processing” OR “transformer”). The search syntax was adapted to the indexing and field conventions of each database.

The search yielded 1253 records in total: 688 from PubMed, 554 from Web of Science, and 11 from PsycINFO. Records were imported into a web-based centralized screening workspace (i.e., Covidence), deduplicated, and screened in two stages (title/abstract, then full text) against the eligibility criteria described below. Both researchers independently completed screening in Covidence using the prespecified eligibility criteria. Any disagreements were resolved through discussion until a consensus was reached.

2.2. Eligibility Criteria

We included empirical studies that applied transformer-based language models to clinical narratives with the purpose of encoding, extracting, or leveraging information from notes for a healthcare application. Eligible studies met all of the following criteria: (1) implemented any transformer-based language model, including encoder-only architectures (e.g., BERT and its biomedical or clinical variants), decoder-only generative large language models (LLMs) (e.g., ChatGPT and Llama), encoder–decoder transformers (e.g., T5), or hybrid systems integrating one or more transformer-based language models with complementary techniques (e.g., retrieval, knowledge graphs, multimodal fusion, or structured EHR data); all transformer-based architectures and hybrid systems were eligible regardless of their specific model names or implementation strategies; (2) used free-text clinical narratives or clinical documentation as a substantive input or output; (3) addressed a healthcare task relevant to extracting, classifying, predicting from, de-identifying, retrieving from, summarizing, or generating clinically oriented information; (4) evaluated at least one clinical task, including performance comparison to other models or human assessment; (5) were peer-reviewed journal articles or conference proceedings; (6) were written in

English; and (7) were published between 1 January 2023 and 5 August 2025, a period selected to capture the rapid expansion of transformer-based and generative AI applications following the public release of ChatGPT. We restricted the review to transformer-based language models because encoder-only, decoder-only, and encoder–decoder transformer architectures underpin contemporary clinical-note representation and generation; non-transformer approaches were therefore outside the review scope. Accordingly, studies relying exclusively on rule-based NLP or traditional machine-learning methods without a transformer-based neural language model were considered outside the scope of this review. Studies focusing solely on non-text modalities or lacking evaluative outcomes relevant to clinical decision support were also excluded. Following application of these eligibility criteria, 36 studies were retained for synthesis (Figure 1).

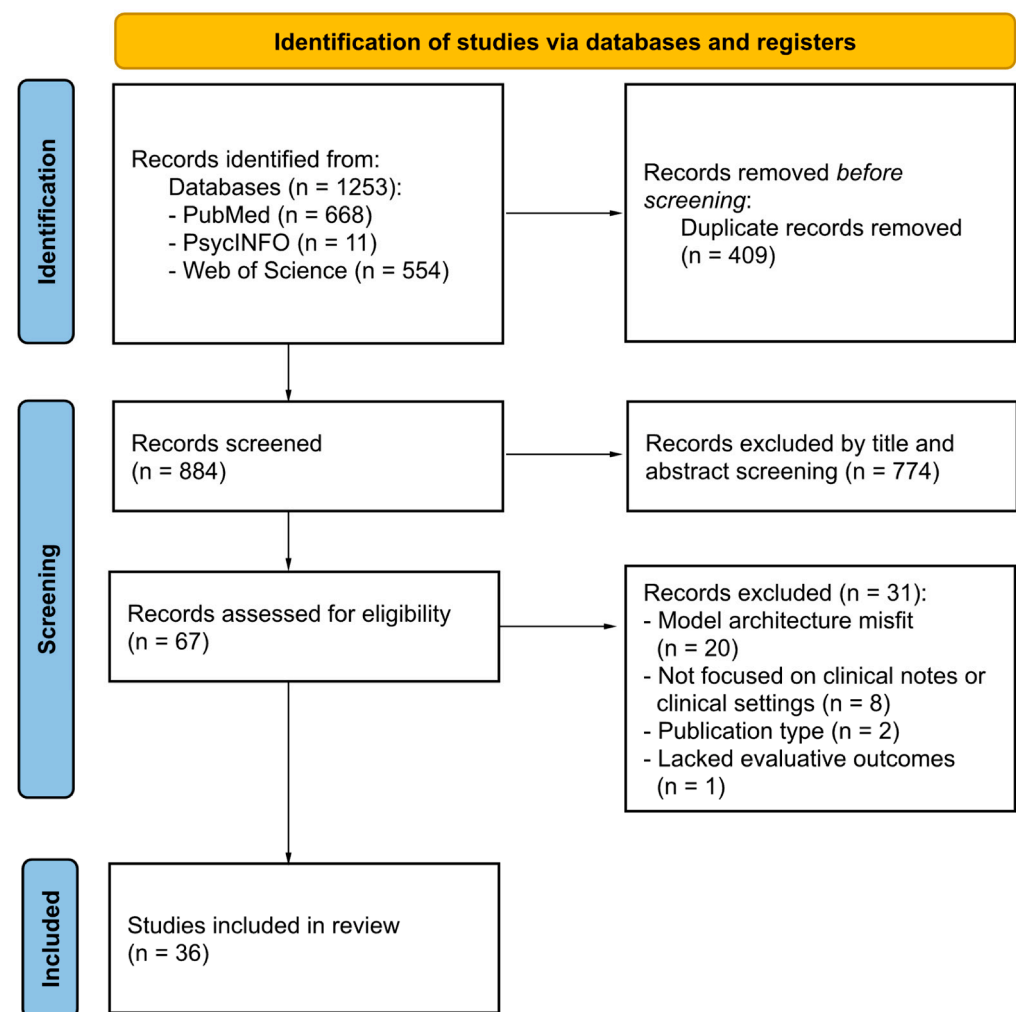


Figure 1. PRISMA chart.

2.3. Data Extraction and Synthesis

For each study, we extracted the author/year, setting and data source, model architecture (including pretrained components), model enhancement techniques (e.g., multimodal fusion and knowledge graph integration), major tasks, and evaluation metrics (e.g., accuracy, F1-score, and AUC; expert comparison). The second author extracted the data from all included studies, and the first author independently verified the extracted information. Any discrepancies were resolved through discussion and consensus. The characteristics of the 36 included studies are summarized in Table 1 [17–52].

Table 1. Data extraction table from 36 studies.

Author (Year)	Major Purpose	Clinical Setting and Dataset	Key Model Architecture	Key Model Enhancement Techniques	Evaluation Metrics
Botero-Aguirre & García-Rivera (2025) [17]	Information Extraction	Allergy; single institution (16,176 entities)	BETO (Spanish BERT)	Fine-tuning	F1, accuracy, recall, sensitivity, specificity, and Cohen's kappa
Cutforth et al. (2023) [18]	Information Extraction	Adult stroke inpatient; multi-institutional (8067 docs)	Token-level NER using IBO tagging + PubMedBERT Ensemble	Ensembling (best performance achieved by ensembling five PubMedBERT models trained on different folds)	F1, precision, recall, and human comparator
Xie et al. (2023) [19]	Information Extraction	Neurology; single institution (55,630 notes)	Ensemble of models (Bio_ClinicalBERT + RoBERTa) + T5	Ensemble voting	F1, sensitivity, PPV, and Cohen's kappa
Laabidi et al. (2024) [20]	Information Extraction	Not specified; public dataset (4998 transcriptions)	BERT variant (BioBERT) + NegSpaCy	Negation detection based on NegEx algorithm	F1, precision, and recall
Eslami et al. (2025) [21]	Information Extraction	Not specified; MIMIC-III (30 summaries)	BERT embedding + SparkNLP NER + voting from eleven BiLSTM-based models	Performance-based voting: selecting the label from among multiple models using their F1-scores	F1
Kim et al. (2025) [22]	Information Extraction	Tertiary care; multi-institutional (5313 notes)	EILEEN (XLM-RoBERTa + Regex)	Multimodal data integration; fine-tuning	F1, AUC, and ROC
Huang et al. (2024) [23]	Information Extraction	Oncology; multi-institutional (1026 reports)	GPT-3.5	Prompt engineering (few-shot, CoT)	F1, precision, recall, accuracy, and Cohen's kappa
Zhang et al. (2025) [24]	Information Extraction	Neurology; multi-institutional (1000 notes)	RAG + Llama 3.1 8B	RAG; Prompt engineering (few-shot)	F1, precision, recall, accuracy, and BERTScore
Sivarajkumar et al. (2024) [25]	Information Extraction	Neurology; single institution (192,000 notes)	LoRA fine-tuning + Llama2	Fine-tuning (LoRA); prompt engineering (CoT)	F1, sensitivity, specificity, PPV, and AUROC
Hier et al. (2025) [26]	Information Extraction	Neurology; single institution (1618 notes)	GPT-4o + Doc2Hpo (ontology mapping)	Prompt engineering, GPT-4o for preprocessing	F1, precision, recall, and accuracy
Chi et al. (2025) [27]	Information Extraction	Adult cardiology care; multi-institutional (507 reports)	Lightweight LLM comparisons	Not specified	F1, precision, and recall
Cho et al. (2024) [28]	Information Extraction	Oncology; single institution (29,914 reports)	LoRA fine-tuning + LLM	Fine-tuning (LoRA); prompt engineering	F1, Macro F1, and accuracy
Arzideh et al. (2025) [29]	De-identification	Inpatient; single institution (10,240 German docs)	German ELECTRA (gELECTRA)	Ensemble method combining two models	F1, precision, recall, and direct human comparator

Table 1. Cont.

Author (Year)	Major Purpose	Clinical Setting and Dataset	Key Model Architecture	Key Model Enhancement Techniques	Evaluation Metrics
Alba et al. (2024) [30]	Classification /Prediction	Primary Care; Veterans Affairs Documents (4B + filtered)	ClinicalBERT	Fine-tuning	F1, precision, recall, and AUC
Floyd et al. (2025) [31]	Classification /Prediction	Orthopedic; single institution (868 notes)	BERT variant (Bio-ClinicalBERT)	Fine-tuning	F1, specificity, sensitivity, precision, AUC, and ROC
Bowles et al. (2025) [32]	Information Extraction	Not specified; Veterans Affairs Documents (1000 titles)	BERT variant (ClinicalTrans-formerNER)	Prompt engineering	F1, precision, and recall
Comito et al. (2025) [33]	Classification /Prediction	Inpatient; single institution (1008 Italian notes for training; 42,335 notes for evaluation)	Class balancing + embedding + ML or BERT	Bayesian hyperparameter optimization; SMOTE and random oversampling	F1, precision, recall, accuracy, and workflow evaluation
Laurentiev et al. (2024) [34]	Information Extraction	Primary care; multi-institutional (10 k key-term-filtered sentences)	BERT variant (BioClinical)	Fine-tuning	AUPRC and AUROC
Ameli et al. (2024) [35]	Classification /Prediction	Dental; single institution (309 charts with notes)	BERT + LSTM	Hybrid architecture	F1, precision, and recall
Jaiswal et al. (2023) [36]	Classification /Prediction	Primary care; multi-institutional (2749 notes)	Sampling/labeling + Blue-BERT/BERTCNN multi-label classification	Down-sampling, unsupervised semantic textual similarity (TSDAE, SimCSE)	AUROC
Holmes et al. (2024) [37]	Classification /Prediction	Inpatient; MIMIC-IV (5857 notes)	MedCAT NER + Sentence Transformer + UMAP + K-means	Guided UMAP; K-means clustering	F1, accuracy, and Silhouette score
Kennedy et al. (2023) [38]	Classification /Prediction	Post-acute care; multi-institutional (594 notes)	Amazon AutoGluon's TextPredictor (transformer-based, pretrained NLP model)	Fine-tuning	F1, precision, recall, and AUROC
Jaiswal et al. (2023) [36]	Classification /Prediction	Primary care; multi-institutional (2749 notes)	Sampling/labeling + Blue-BERT/BERTCNN multi-label classification	Down-sampling, unsupervised semantic textual similarity (TSDAE, SimCSE)	AUROC
Kennedy et al. (2023) [38]	Classification /Prediction	Post-acute care; multi-institutional (594 notes)	Amazon AutoGluon's TextPredictor (transformer-based, pretrained NLP model)	Fine-tuning	F1, precision, recall, and AUROC

Table 1. Cont.

Author (Year)	Major Purpose	Clinical Setting and Dataset	Key Model Architecture	Key Model Enhancement Techniques	Evaluation Metrics
Miller et al. (2025) [39]	Classification /Prediction	ICU; MIMIC-III (4161 cases) + single institution (3532 cases)	Dynamic few-shot prompting + LLM	Dynamic few-shot prompt engineering	F1, precision, and recall
Nguyen Phan et al. (2024) [40]	Classification /Prediction	Inpatient; MIMIC-III + single Institution (1.4 M+)	Multimodal Embedding Extraction + Disentangled Transformer Fusion	Multimodal data fusion; Masked Lab-Test Modeling (MLTM)	F1, precision, recall, and accuracy
De Hond et al. (2024) [41]	Classification /Prediction	Oncology; multi-institutional (16,159 patients)	Data fusion + BERT variant (DistilBERT) or LASSO logistics regression	Multimodal data fusion	F1, AUC, calibration, and fairness/subgroup assessment
Li et al. (2023) [42]	Classification /Prediction	Oncology; multi-institutional (EHR from 1463 patients)	Fusion framework (ML + BERT)	Multimodal fusion; interpretability uses SHAP (Shapley Additive exPlanations)	F1, precision, recall, accuracy, AUC, and human comparator
Kolukula et al. (2024) [43]	Classification /Prediction	ICU; MIMIC-III (7499 patients)	FABiLSTM (PubMedBERT + BiLSTM)	feedback attention; K-means, a billiards-inspired optimization algorithm (BOA)	F1, precision, recall, accuracy, and AUC-ROC
Abdullahi et al. (2024) [44]	Classification /Prediction	Not specified; MIMIC-III + PubMed (33M abstracts)	BM25 + transformer-based MedCPT + ranking by TF-IDF scores	Hybrid architecture; RAG	Mean Reciprocal Rank (MRR)
Yang et al. (2023) [45]	Classification /Prediction	ICU; MIMIC-III (7493) + MIMIC-IV (10,000) + eICU (9408 patients)	Time-Transformer + Personalized Knowledge Graph (PKG) + Graph Attention Network (GAN) + ChatGPT	Knowledge graph integration with graph attention and edge weight computation	F1, recall, and AUC
Gao et al. (2025) [46]	Summarization/ Generation	ICU; MIMIC-III (1005 progress note) + private 4815 notes	UMLS-based knowledge graph + Concept embedding with SapBERT + LLM	Multi-hop path generation on knowledge graph, prompt engineering	F1, precision, recall, and ROUGE
Daphne et al. (2025) [47]	Classification /Prediction	Critical care; MIMIC-IV (42,671 cases)	PKGNN (Patient Knowledge Graph Neural Network)—BERT variants + a two-layer Graph Convolutional Network (GCN)	Ensemble approach with graph network; BERT variants are individually fine-tuned to encode medical notes	Precision, recall, accuracy, UROC, AUPRC, and R@P80
Lyu et al. (2025) [48]	Classification /Prediction	ICU; MIMIC-III (7346 patients)	PDSNet (Graph Convolutional Network (GCN) + Transformer)	2-layer ontology-based GCN; masking-based data augmentation	F1, precision, recall, AUC, and fairness/subgroup assessment
Chien et al. (2024) [49]	Summarization/ Generation	Neuroradiology; 137 reports + 100 PubMed cases	LLM comparisons	Not specified	F1, ROUGE, and BERTScore

Table 1. Cont.

Author (Year)	Major Purpose	Clinical Setting and Dataset	Key Model Architecture	Key Model Enhancement Techniques	Evaluation Metrics
Chuang et al. (2024) [50]	Summarization/ Generation	Radiology; MIMIC-CXR (91,544 radiology reports)	SPeC (Soft Prompt-Based Calibration)	Soft prompt-based calibration	ROUGE and F1
Eghbali et al. (2025) [51]	Summarization/ Generation	Orthopedic inpatient; single institution (1000 notes)	T5 with two-stage (unsupervised denoising pretraining + supervised fine-tuning) for summarization	Two-step fine-tuning approach	ROUGE, cosine similarity, and human comparator
Afshar et al. (2025) [52]	Summarization/ Generation	Inpatient; multi-institutional (8527 docs)	Abridge AI + GPT-4o	Prompt engineering	F1, precision, recall, Pearson's r, human comparator, workflow evaluation, and clinical deployment

To answer RQ1, each record received one primary-purpose code: (1) information extraction/de-identification; (2) classification/prediction; or (3) summarization/generation, including clinically oriented summary, report, or note generation. Each study was also mapped to one of five care-setting categories: (1) acute care, (2) ambulatory or outpatient care, (3) diagnostic services, (4) long-term or residential care, or (5) mixed/health-system-wide care. Primary model approach was coded as (1) encoder-only transformer, (2) decoder-only transformer/LLM, (3) encoder-decoder transformer, or (4) hybrid and pipeline-based approach. When a study used multiple architectures or components, the primary model code represented the approach central to its principal clinical task and reported outcome. Fine-tuning, prompting, retrieval augmentation, multimodal fusion, and other adaptation methods were therefore not treated as model families.

To answer RQ2, we performed a structured methodological appraisal. We classified the strongest supported validation design and recorded four categories: (1) human comparison, (2) fairness assessment, (3) workflow evaluation, and (4) clinical deployment. Human comparison referred to an explicit model-versus-human performance comparison on matched cases; creation of a reference standard or human rating of model outputs alone did not meet this definition. Fairness assessment referred to an explicit analysis of performance, calibration, or errors across demographic or other prespecified subgroups. Workflow evaluation referred to a qualified assessment in a simulated or live clinical workflow, whereas clinical deployment required implementation in a live clinical service. A record could contribute to more than one indicator.

For RQ1 and RQ2 results reporting, we used descriptive counts and narrative synthesis. In addition to the descriptive mappings, we characterized recurring application patterns for RQ1 and recurring evaluation and implementation patterns for RQ2. No meta-analysis or cross-study performance ranking was conducted because task definitions, source populations, reference standards, test denominators, and evaluation designs were not sufficiently comparable.

3. Results

3.1. RQ1: Major Purposes, Clinical Distribution, and Model Families

3.1.1. Overall Distribution

Across the 36 records, 15 (41.7%) addressed information extraction or de-identification, 16 (44.4%) classification or prediction, and 5 (13.9%) summarization or generation (Figure 2). The mapped care-setting distribution was mixed/health-system-wide in 18 records (50.0%), diagnostic services in 6 (16.7%), ambulatory/outpatient care in 6 (16.7%), acute care in 5 (13.9%), and long-term/residential care in 1 (2.8%).

Evidence map of the 36 included records

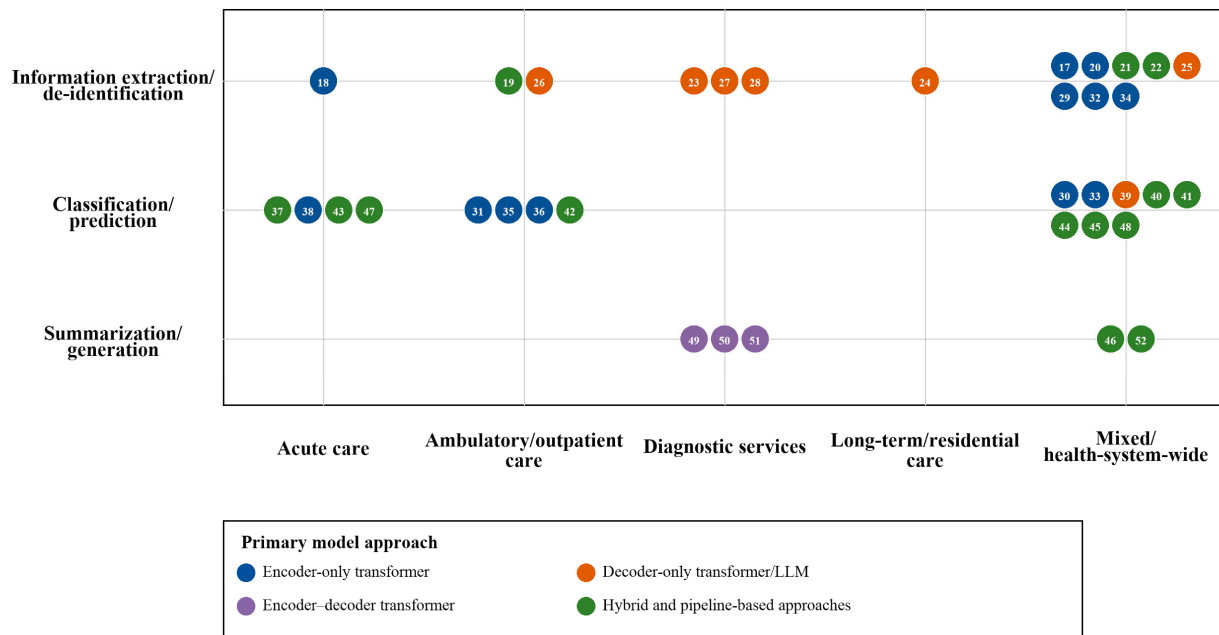


Figure 2. Evidence map by primary purpose, care setting, and primary model approach.

The distribution of major purposes differed across mapped care settings. Information extraction/de-identification and classification/prediction were each most frequently mapped to mixed/health-system-wide sources (8/15 and 8/16 records, respectively). In contrast, summarization/generation was concentrated in diagnostic services (3/5 records), principally radiology-oriented applications. Within the mixed/health-system-wide category, eight records addressed information extraction/de-identification, eight addressed classification/prediction, and two addressed summarization/generation.

Regarding primary model approach, 12 records (33.3%) used encoder-only transformers, 7 (19.4%) used decoder-only transformer/LLM systems, 3 (8.3%) used encoder-decoder transformers, and 14 (38.9%) used hybrid and pipeline-based approaches. Model approaches also differed by primary purpose: six of the seven decoder-only transformer/LLM records addressed information extraction/de-identification; all three encoder-decoder transformer records addressed summarization/generation; and nine of the fourteen hybrid and pipeline-based records addressed classification/prediction. Encoder-only transformers were distributed evenly between information extraction/de-identification and classification/prediction (six records each). Within the mixed/health-system-wide category, hybrid and pipeline-based approaches accounted for nine records, encoder-only transformers for seven, and decoder-only transformer/LLM systems for two.

No record evaluated a transformer-based language model application within a behavioral-health service. Alcohol-use extraction [22], psychiatric symptom profiling

from hospital and emergency-department documentation [37], and depression-risk modeling in cancer care [41] addressed behaviorally or mentally relevant content, but none evaluated a behavioral-health clinic, treatment workflow, or service-level outcome.

3.1.2. Patterns in Information Extraction and De-Identification

The 15 information-extraction records covered medication-allergy entities [17], thrombolysis-relevant concepts [18], epilepsy outcomes [19], symptoms and assertions [20,21], alcohol-use patterns [22], pathology attributes and HPO concepts [23,26,28], dementia-agitation information [24], sleep concepts [25], echocardiographic findings [27], protected health information [29], note-title ontology elements [32], and functional-status evidence [34]. These studies spanned all five mapped settings, with diagnostic-report services represented by pathology and echocardiography applications [23,27,28], and one long-term/residential-care study focused on nursing-progress documentation [24].

Encoder-only transformers were commonly applied to entity recognition, de-identification, document-title normalization, and functional-status extraction. Several studies showed that fine-tuning BERT models effectively identified key entities. For example, the reviewed studies used language-specific or biomedical BERT variants for allergy extraction, stroke-related concept identification, symptom mapping, and protected-health-information detection [17,18,20,29]. Extending these developments, Eslami et al. [21] proposed a hybrid voting framework (BioBERT and BiLSTM-CNN-Char) that classified clinical assertions. Kim et al. [22] used a hybrid method to extract alcohol-consumption patterns from clinical notes, combining semantic embeddings from a fine-tuned multilingual transformer (mBERT and XLM-RoBERTa) with keyword and regex-based extraction of explicit details.

Building on such advances, recent research has increasingly turned to generative LLMs (e.g., GPT-4o), as these models can perform contextual reasoning and information extraction through prompt-based learning, reducing dependence on large annotated datasets. Huang et al. [23] used GPT3davinci to extract medical facts via prompt engineering, outperforming traditional NLP without problem-specific annotations. Expanding on this line of work, Zhang et al. [24] applied a retrieval-augmented generation (RAG) framework with Llama 3.1 for dementia care. Sivarajkumar et al. [25] compared rule-based and generative LLM approaches for extracting sleep-related information from Alzheimer's patient records, finding that rule-based systems can remain competitive when training data are scarce. In parallel, Hier et al. [26] showed that preprocessing notes with GPT-4o significantly improved extraction accuracy. To address practicality, researchers explored lightweight open-source language models. Chi et al. [27] and Cho et al. [28] demonstrated that lightweight open-source language models (e.g., Llama3-8B and Mistral-7B) achieved competitive performance in extracting structured information from echocardiography and pathology reports, suggesting their potential for selected domain-specific clinical documentation tasks. Regarding protecting patients' privacy, Arzideh et al. [29] developed a transformer-based pipeline for de-identifying German clinical documents, surpassing human annotators.

3.1.3. Patterns in Classification and Predictive Modeling

The 16 classification and prediction studies applied transformer-based language models to diverse clinical classification tasks using unstructured clinical notes. These tasks included identifying tobacco-use status [30], predicting orthopedic follow-up outcomes [31], detecting chronic pain [33], classifying periodontitis stage and grade [35], identifying lower-back-pain risk factors [36], stratifying psychiatric symptom profiles [37], classifying patient preferences [38], and categorizing clinical-note sections [39]. A second cluster addressed multi-disease, depression-risk, liver-metastasis, mortality/readmission,

diagnostic-retrieval, or cardiovascular-prediction tasks [40–45,47,48]. Eight records mapped to mixed/health-system-wide sources, four to ambulatory/outpatient care, and four to acute care.

Fine-tuned encoder-only transformers were commonly used for this task, and while they often yield superior performance, this advantage is not consistent across contexts. For instance, fine-tuned BERT models (e.g., ClinicalBERT and Bio-ClinicalBERT) substantially outperformed ML baselines in classification tasks [30,31]. Conversely, several studies have reported that BERT models do not uniformly outperform traditional ML methods, particularly when clinical documentation is concise, containing many abbreviations, or limited in narrative context. Comito et al. [33] found that XGBoost and Random Forest models outperformed BERT variants in classifying chronic pain and dementia impairments from brief, keyword-driven texts.

A growing body of work integrates fine-tuned BERT with other deep learning models or clustering techniques to enhance classification performance. For instance, Ameli et al. [35] developed a hybrid BERT-LSTM architecture to classify periodontitis stage and grade from dental clinical notes, recognizing that BERT's standard context window limitation becomes problematic when capturing complex clinical information distributed across lengthy notes. Similarly, Jaiswal et al. [36] used a BERT-CNN multi-task model to classify multiple lower-back-pain risk factors simultaneously, yielding a 15% improvement over TF-IDF + logistic regression. Holmes et al. [37] combined sentence-transformer embeddings with guided UMAP and k-means clustering to stratify psychiatric symptom profiles based on clinically relevant terms extracted from notes.

Several studies highlight the practical benefits of using lightweight or commercial language models for classification tasks in clinical settings. For instance, Kennedy et al. [38] used Amazon's AutoGluon TextPredictor to classify negative patient preferences in notes from hospitalized older adults, achieving an F1 of 0.92 and outperforming 13 traditional ML models. Miller et al. [39] implemented a dynamic few-shot prompting technique and evaluated lightweight LLMs such as Llama and Mistral for classifying note sections (e.g., subjective or objective).

To enhance predictive accuracy and interpretability, researchers have increasingly adopted advanced methods for predictive modeling, especially multimodal integration and knowledge graphs. In multimodal approaches, researchers combine structured data (e.g., laboratory values and demographics) with unstructured data (e.g., clinical notes), often using fusion or hybrid model architectures. MEDFuse [40] fused Medical-Llama3 text embeddings with a transformer for lab data, achieving high performance (F1 = 0.92). De Hond et al. [41] similarly showed that adding patient and clinician notes could improve depression-risk prediction. Li et al. [42] integrated BERT-based text embeddings with outputs from traditional ML classifiers in a fusion model for liver metastasis prediction. FABiLSTM [43] combined preprocessed clinical text with TF-IDF-based K-means clustering, feeding these into a feedback-attention BiLSTM to improve clinical prediction. Abdullahi et al. [44] introduced ClinIQIR, a retrieval-augmented pipeline that uses BM25 and MedCPT to search 33 million PubMed abstracts and generate differential diagnoses, showing strong performance for rare conditions.

Several studies have enabled models to reference medical knowledge and relational structures by utilizing knowledge graphs and graph-based architectures. For instance, Yang et al. [45] combined admission note histories with a medical knowledge graph to enhance diagnosis prediction. Gao et al. [46] likewise strengthened a transformer with graph-based medical knowledge to better identify complex conditions for summarization and generation purpose. Daphne et al. [47] used Graph Convolutional Networks to fuse ClinicalBERT, BioBERT, and BlueBERT embeddings into patient graphs, improving predictions such as

ICU mortality and readmission. Lyu et al. [48] developed PDSNet, which constructs patient and disease similarity graphs processed through Graph Convolutional Networks (GCNs) before a transformer predicts heart failure risk.

3.1.4. Patterns in Summarization and Text Generation

Five records addressed generation: active problem or diagnosis-list generation from inpatient progress notes [46], longitudinal neuroradiology-report summarization [49], findings-to-impression generation for chest radiology [50], knee-MRI history generation from orthopedic notes [51], and ambient draft-note generation from encounter audio [52]. Three records were situated in diagnostic services [49–51], and two were mapped to mixed/health-system-wide documentation [46,52].

Given the inherent limitation of encoder-only models such as BERT in text generation, recent studies have adopted encoder–decoder or decoder-only architectures (e.g., LongT5, Flan-T5, and GPT-based models) for clinical-note summarization and generation. Chien et al. [49] compared multiple encoder–decoder and decoder-only models for radiology-report summarization, finding that BARTnn and GPT3davinci achieved the strongest performance, though the authors noted that none of them was ready for clinical deployment due to missing key information. To address instability in LLM outputs, Chuang et al. [50] introduced SPeC, a Soft-Prompt Calibration module designed to reduce variability caused by small prompt changes. SPeC functions as a soft-prompt encoder that minimizes information loss between the original note and the generated summary, providing a corrective calibration signal that keeps LLM outputs stable and factually grounded despite prompt variations. Other work has explored automated summarization to improve the quality of radiology histories provided to imaging interpreters. For instance, Eghbali et al. [51] developed a transformer-based system that generated brief summaries of patients' clinical notes to accompany imaging orders; radiologists reported that the summaries were useful, although they occasionally lacked subtle contextual nuances. Afshar et al. [52] further advanced this line of research by proposing a “playbook” for integrating ambient AI, like automated documentation summarization, into clinical workflows. Their pilot study demonstrated a system in which an LLM monitored provider–patient conversations and generated draft clinical notes. This pragmatic framework highlighted several operational challenges, including sensitivity to changes in note templates, and underscored the need for continuous evaluation when deploying AI systems in live clinical environments. Taken together, these studies represent early efforts to apply generative LLMs to clinical text summarization and collectively indicate that although the technology shows considerable promise, substantial refinement is required before routine clinical adoption is feasible.

3.2. RQ2: Evaluation, Validation, and Implementation Evidence

3.2.1. Overall Distribution

Standard performance metrics (e.g., F1-score, precision, recall, accuracy, and AUROC) predominated in information-extraction and classification/prediction studies. All included records reported at least one task-specific evaluative outcome; summarization and generation studies additionally used text-generation metrics and, in some cases, human or expert output assessment. Figure 3 shows that evidence beyond retrospective task performance was rare. Five records (13.9%) reported an explicit direct human comparator, whereas two (5.6%) reported a fairness or subgroup assessment, two (5.6%) reported workflow evaluation, and one (2.8%) reported clinical deployment.

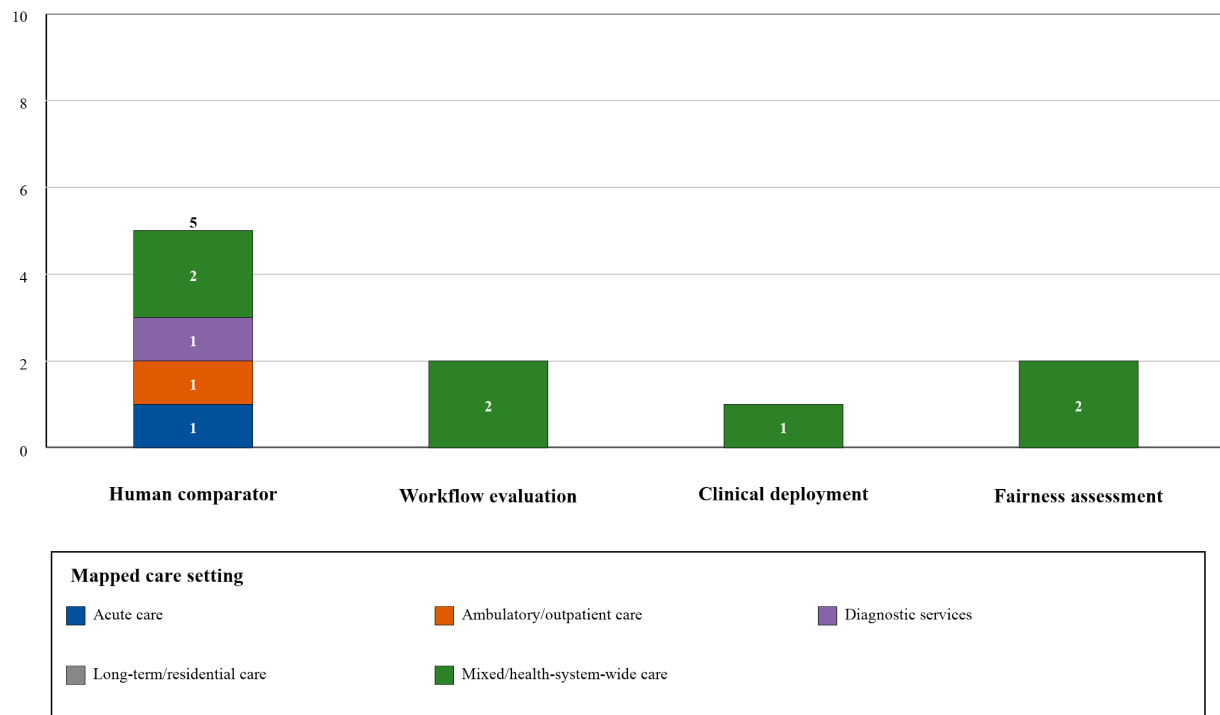


Figure 3. Study counts of reported evidence beyond retrospective task performance.

3.2.2. Patterns in Evaluation Metrics and Human Comparison

Standard performance metrics (e.g., F1-score, precision, recall, accuracy, and AUROC) were the predominant methods for evaluating information extraction and classification tasks, with every study employing at least one of these measures. Interpretability and calibration analyses were less frequently reported than discrimination or classification metrics. Only Li et al. [42] explicitly reported SHAP-based feature-importance analysis to examine clinical predictors in its model. De Hond et al. [41] additionally reported calibration measures and decision-curve analysis alongside discrimination estimates.

In the domain of summarization and note generation, evaluation strategies have expanded to assess the semantic quality of machine-generated text against reference standards. Multiple studies utilized Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores to measure textual overlap [46,49,51]. Additionally, BERTScore was employed to evaluate generation quality by calculating the cosine similarity between BERT embeddings of the generated and reference texts [49]. Although cosine similarity is commonly used in RAG workflows (e.g., [43]) and model architecture development [44], only one study employed it explicitly as a model evaluation metric [51].

Human involvement took several distinct forms, including the creation of reference standards, qualitative review of model outputs, and direct comparisons of model-versus-human performance. Under our prespecified definition, a direct human comparator required evaluation of model and human performance on matched cases; studies in which clinicians only created or verified reference standards were not classified as direct human comparisons. Only five studies met this criterion [18,29,42,51,52]. However, the comparator populations and reference standards varied substantially across studies. Cutforth et al. [18] reported a limited comparison with clinician reviewers, whereas Arzideh et al. [29] compared model performance with two medical trainees on a selected internal sample. Li et al. [42] evaluated model performance against two clinicians with more than 10 years of experience. Eghbali et al. [51] compared AI-generated clinical histories with existing imaging-order histories rather than clinician performance, and Afshar et al. [52] compared GPT-4o with professional medical coders for a secondary ICD-10 coding task. Because

comparator expertise, reference standards, and evaluation designs differed considerably, these findings should not be interpreted as evidence of generalized clinician-equivalent or clinician-superior performance.

Finally, extending validation beyond retrospective model comparison, prospective field testing in active clinical environments remains exceptionally rare. Only a single study, Afshar et al. [52], demonstrated large-scale practical feasibility by integrating an ambient AI system (Abridge AI) into the clinical workflows of 12 clinics across eight medical specialties. Beyond limited prospective evaluation, important methodological challenges, including data leakage during model development and data shift resulting from differences in patient populations, documentation practices, and healthcare settings, remain significant threats to the reliability and generalizability of clinical AI systems. Collectively, these findings indicate that although the field has made notable advances in algorithmic performance, substantial work is still required to establish clinical relevance, transparency, robustness, and safe real-world deployment.

3.2.3. Patterns in Cross-Linguistics and Contextualization

Language plays a pivotal role in model training and validation. Although several studies used predominantly English-language data sources, including MIMIC-derived datasets [40,45,48], some studies adapted transformer-based systems to non-English clinical documentation. Botero-Aguirre and García-Rivera [17] applied a Spanish-language BERT model to allergy extraction, whereas Arzideh et al. [29] evaluated German-language models for clinical-document de-identification. Kim et al. [22] developed EILEEN for alcohol-pattern extraction in a Korean clinical context by combining multilingual transformer representations with structured extraction components.

However, language-specific model development should not be equated with cross-lingual validation. The reviewed studies generally evaluated their models within the language, institution, and documentation context from which the training or test data were drawn. They did not provide common multilingual benchmarks or direct evidence that performance would transfer across languages, institutions, clinical cultures, or documentation practices. Similarly, the use of domain-specific pretrained models, such as BioBERT, ClinicalBERT, BlueBERT, and PubMedBERT, reflected efforts to align representations with biomedical terminology but did not by itself demonstrate generalizability beyond the evaluated corpus [34,44,47].

Fairness and subgroup evaluation were likewise rare. De Hond et al. [41] conducted subgroup and calibration analyses in depression-risk prediction during cancer treatment and identified underestimation of risk for female and Black patients. The only other record reporting subgroup performance was Lyu et al. [48], which presented limited sex- and age-stratified results. Together, these findings indicate that contextual, cross-lingual, and equity-related evaluation remained substantially less developed than retrospective task-level performance reporting.

4. Discussion

The current scoping review illuminated a rapidly evolving landscape wherein transformer-based language models are being actively integrated to derive clinically meaningful insights from the previously opaque domain of unstructured EHR notes. The central finding is not that one model family has established clinical superiority. Rather, our results show that different transformer-based approaches have been applied across different clinical settings and for different purposes, while the evidence needed to support generalizability, human comparability, fairness, workflow integration, and clinical deployment remains significantly limited. In the following paragraphs, we therefore interpret the

results in relation to the two research questions: first, where and for what purposes these systems have been applied; and second, what the reported evaluation evidence can, and cannot, support.

4.1. RQ1: Application Scope and Patterns of Transformer-Based Approaches

The RQ1 evidence map showed that the included literature was concentrated in information extraction/de-identification (15 studies) and classification/prediction (16 studies), whereas summarization/generation accounted for 5 studies. This pattern is consistent with the role of clinical narratives as a source of information that must first be identified, structured, or de-identified before it can support downstream clinical or operational tasks [18,22]. The smaller number of generation studies should not be interpreted as limited interest in generative approaches; study counts alone cannot establish the maturity, attention, or clinical readiness of a model approach.

Importantly, the results do not support a conclusion that one transformer approach is superior to another. Instead, model approaches appeared to align with task characteristics: encoder-only and decoder-only transformer/LLM systems were both frequently used for extraction; encoder-decoder transformers were used for summarization/generation; and hybrid and pipeline-based approaches were most common in classification/prediction. Hybrid systems often combined language model representations with structured EHR variables, retrieval components, rules, or knowledge-graph information [40–46]. Such combinations may incorporate complementary clinical information, but they do not by themselves establish improved generalizability, interpretability, or clinical benefit. Future research should therefore evaluate whether and under what conditions these system-level designs provide clinically meaningful interpretability and utility (i.e., eXplainable AI) [42,45].

Finally, the evidence map identifies an important gap in behavioral-health settings. No included studies evaluated a transformer-based application within a behavioral-health service, workflow, or service-level outcome. Alcohol-use extraction [22], psychiatric-symptom profiling [37], and depression-risk modeling in cancer care [41] addressed behaviorally or mentally relevant content, but they were not direct evidence for behavioral-health implementation in clinical settings. The findings therefore support the relevance of clinical-note methods to future behavioral-health research but not claims that their effectiveness or safety has already been established in behavioral-health services.

4.2. RQ2: Challenges to Trustworthiness, Generalizability, and Equity

RQ2 showed that retrospective task performance remained the predominant form of evaluation, whereas evidence beyond retrospective task performance was limited. Standard metrics (e.g., precision, recall, F1-score, accuracy, or AUROC) are necessary for evaluating a defined task; however, they alone do not establish whether a system will generalize across institutions, perform comparably with clinicians, or improve clinical workflows [13,14]. Relatedly, our results showed that only five studies reported an explicit direct human comparator, two reported fairness or subgroup assessment, two reported workflow evaluation, and one reported clinical deployment. In addition, data sparsity and quality remain central concerns. Many reviewed studies suffered from small sample sizes or significant class imbalance, constraining model training and often limiting model success to “low-hanging fruit” tasks that involve well-documented clinical phenomena [35]. Overall, these limitations affect the interpretation of reported performance values and reinforce the need to distinguish technical task performance from clinical effectiveness.

Generalizability is another recurring concern. Models trained on single-institution data could fail to transfer effectively to new settings due to variations in local language,

documentation practices, and patient populations [31]. This necessitates rigorous multi-site external validation before any system of this kind can be deemed robust across diverse healthcare environments. Another important direction concerns the balance between increasingly large foundation models and smaller task-specific models. Recent evidence suggests that compact, domain-specific models may offer advantages in computational efficiency, interpretability, deployment feasibility, and implementation cost, whereas general-purpose foundation models with limited task-specific fine-tuning may provide broader transferability across diverse clinical tasks [53,54]. Rather than assuming a single optimal architecture, future research should systematically compare these complementary approaches using standardized benchmarks, external and prospective validation, workflow evaluation, and clinically meaningful implementation outcomes to determine their appropriate roles in clinical decision support. Beyond predictive performance, successful implementation will also depend on economic and operational feasibility. Large foundation models often require substantial computational resources, specialized infrastructure, and ongoing maintenance, which may limit their adoption in community hospitals, outpatient clinics, and behavioral-health organizations with constrained technical capacity. In contrast, lightweight task-specific models may offer a more practical alternative for well-defined clinical tasks because they require fewer computational resources while remaining easier to deploy and maintain. Future implementation research should therefore evaluate not only predictive performance but also computational efficiency, implementation costs, scalability, and sustainability to identify deployment strategies that are both clinically effective and accessible across diverse healthcare settings.

Trustworthiness is also particularly salient for generative applications. Generative LLMs can produce coherent but incorrect outputs, as illustrated by Huang et al. [23], who reported fabricated or incorrect responses in difficult pathology-report cases. This risk underscores the need for structured human–AI partnership models in which clinicians verify AI-generated outputs against source clinical documentation before they inform patient care. The level of oversight should be proportional to clinical risk: diagnostic suggestions, risk predictions, and treatment recommendations should require explicit clinician review, while lower-risk documentation support may be monitored through routine quality assurance. Additional safeguards, including validation against source records, prospective workflow evaluation, reporting of uncertainties when feasible, and continuous post-deployment monitoring, will be essential to minimize hallucinations and ensure safe clinical implementation.

Equity and interpretability were similarly under-represented in the reviewed literature. Only two studies reported fairness or subgroup analyses. For example, De Hond et al. [41] found that a depression-risk model underestimated risk for female and Black patients, illustrating that overall discrimination metrics alone are insufficient for evaluating clinical AI systems. Differences in documentation practices, historical labeling, and data representation across demographic groups may lead clinical-note models to perpetuate or even amplify existing healthcare disparities. Consequently, future studies should routinely report subgroup-specific performance, calibration, and error analyses alongside clinically meaningful forms of interpretability [41,42,45].

Beyond fairness assessment, ethically grounded explainable human–AI partnership models should be designed to support rather than replace clinical judgment. Such systems should provide transparent, clinically interpretable explanations linked to relevant narrative evidence, incorporate routine subgroup-specific fairness evaluation, and require human verification for high-risk clinical decisions. Implementation should also include continuous monitoring for model drift, bias, and unintended consequences, supported by transparent governance frameworks addressing accountability, privacy, and patient safety.

Accordingly, evaluation should extend beyond predictive performance to include fairness, explainability, workflow integration, clinician trust, and patient-centered outcomes before routine clinical deployment.

Operational evidence remained limited. Afshar et al. [52] provided the only example of implementation in a live clinical environment, demonstrating the value of human-in-the-loop monitoring by identifying and resolving a documentation-compliance issue associated with an EHR template change. However, the primary system generated draft clinical notes from encounter audio, and the direct human comparison was limited to a secondary GPT-4o ICD-10 coding task. Accordingly, this study should be interpreted as evidence of workflow implementation rather than clinician-equivalent performance in clinical documentation.

More broadly, this implementation experience illustrates that successful clinical deployment depends not only on algorithmic performance but also on stable integration with evolving EHR workflows, continuous quality assurance, clinician training, monitoring for model drift, and governance processes that enable timely adaptation to changes in clinical documentation practices. Similarly, although several studies demonstrated successful adaptation of transformer-based language models to individual non-English languages, these findings should not be interpreted as evidence of multilingual generalizability. Robust deployment across diverse healthcare settings will require standardized multilingual benchmarks, cross-lingual external validation, and culturally representative datasets to ensure equitable performance across languages and clinical contexts.

Collectively, these findings indicate that the current evidence base has progressed considerably further in demonstrating the technical feasibility of transformer-based language models than in establishing their clinical implementation readiness. Although these models show considerable potential to support clinical decision making by extracting and synthesizing clinically meaningful information from unstructured clinical notes, the available evidence remains largely confined to retrospective task evaluation, with limited evidence supporting transportability, equitable performance, workflow integration, or safe deployment in routine clinical practice.

Furthermore, the reviewed studies employed a wide range of evaluation metrics, including F1-score, precision, recall, accuracy, AUROC, ROUGE, BERTScore, and, in several information extraction studies, Cohen's kappa, reflecting the diverse objectives of clinical NLP tasks. These complementary metrics capture different aspects of model performance, underscoring that no single measure is sufficient to establish clinical utility. In particular, agreement measures such as Cohen's kappa provide valuable information for annotation-intensive tasks involving potentially imbalanced datasets because they account for agreement beyond chance. Consistent with recent recommendations for evaluating clinical AI, future studies should adopt multidimensional evaluation frameworks that integrate discrimination, calibration, agreement, interpretability, fairness, and clinical utility, rather than relying on a single performance metric [55].

These observations are also consistent with recent reviews emphasizing that the successful adoption of large language models in healthcare depends on more than improvements in algorithmic performance. Robust implementation will require rigorous external and prospective validation, transparent governance, workflow integration, and responsible human-AI partnership models that support trustworthy clinical decision making [56–58]. Accordingly, future research should prioritize external generalizability, direct comparison with clinician performance, fairness assessment, workflow evaluation, and implementation studies that evaluate meaningful clinical, organizational, and patient-centered outcomes.

Importantly, none of the included studies evaluated transformer-based language models within behavioral-health services, workflows, or service-level outcomes. Therefore, the application of these technologies to mental health (MH) and substance use disorder

(SUD) services should be regarded as a promising future research direction rather than an established clinical application. Behavioral-health applications present distinctive challenges that extend beyond many of the clinical tasks represented in the reviewed literature. Mental health and substance use disorder documentation frequently consists of longitudinal, subjective narratives describing symptoms, psychosocial stressors, interpersonal relationships, coping strategies, and recovery trajectories rather than discrete biomedical findings. Consequently, future transformer-based systems for behavioral health should be evaluated not only for conventional information extraction and prediction but also for their ability to support clinically meaningful tasks such as longitudinal risk assessment, treatment response monitoring, recovery planning, and other forms of narrative-based clinical decision support. These applications will require greater emphasis on contextual reasoning, explainability, fairness, and prospective workflow evaluation than is currently represented in the available evidence. Given the sensitive and longitudinal nature of behavioral-health documentation, future implementation should incorporate explainable human–AI partnership models, rigorous bias mitigation, hallucination prevention, ethical governance, and attention to cross-cultural and multilingual contexts to ensure that these technologies are trustworthy, equitable, and clinically meaningful.

4.3. Limitations and Future Research

While this review provides a comprehensive evidence map of a rapidly evolving field, several limitations should be acknowledged. First, consistent with the objectives of a PRISMA-ScR scoping review, we mapped the breadth and characteristics of the evidence but did not conduct a formal risk-of-bias assessment or meta-analysis. Second, our search was limited to English-language, peer-reviewed articles indexed in PubMed, Web of Science, and PsycINFO over a 31-month period. Consequently, relevant engineering-oriented studies indexed in IEEE Xplore or the ACM Digital Library, as well as preprints available through repositories such as arXiv, may not have been captured. In addition, the rapidly evolving terminology surrounding transformer-based AI, including newer foundation models, instruction-tuning approaches, and implementation strategies such as RAG, may have resulted in some relevant studies being overlooked despite our concept-based search strategy. Future scoping reviews should therefore broaden database coverage and incorporate emerging model-specific keywords and AI terminology while maintaining methodological rigor and clinical relevance.

Although title/abstract and full-text screening were independently conducted by two reviewers using predefined eligibility criteria within Covidence, with disagreements resolved through consensus, formal inter-rater agreement statistics (e.g., Cohen’s kappa) were not prospectively calculated. While independent dual screening reduces the likelihood of selection bias, reporting inter-rater agreement would have further strengthened the transparency and reproducibility of the review. Future systematic and scoping reviews should prospectively calculate and report agreement statistics to provide an objective measure of screening consistency.

Third, this review intentionally focused on transformer-based language models because they represent the predominant paradigm for contemporary clinical language understanding and generation. As a result, potentially relevant studies employing alternative AI architectures (e.g., recurrent neural networks, convolutional neural networks, graph neural networks without transformer components, or traditional machine learning approaches) were outside the scope of this review. Future comparative reviews evaluating transformer-based and non-transformer architectures could provide a broader understanding of their relative strengths, limitations, and implementation characteristics for clinical decision support.

Finally, substantial heterogeneity in clinical tasks, model architectures, datasets, validation designs, and evaluation metrics precluded quantitative comparisons across studies. Moreover, organizational and implementation factors, including interoperability across EHR systems, implementation and maintenance costs, organizational readiness, and evolving legal and regulatory requirements, were beyond the scope of this review because they were infrequently reported in the primary studies. These factors represent important priorities for future implementation science research and are essential for translating promising AI technologies into routine clinical practice. Future studies should prioritize transparent reference standards, external and prospective validation, direct human comparison, subgroup and calibration analyses, workflow evaluation, and implementation studies. Direct evidence in behavioral health also remains limited, as none of the included studies evaluated transformer-based language models within behavioral-health workflows or service-level outcomes, underscoring the need for prospective evaluations in real-world behavioral-health settings.

5. Conclusions

Transformer-based language models have demonstrated considerable potential for extracting and leveraging clinically actionable information from unstructured clinical notes. This scoping review shows that transformer-based and hybrid approaches have been successfully applied to retrospective information extraction, classification, prediction, and text-generation tasks across diverse healthcare settings. However, the current evidence primarily supports technical feasibility rather than clinical implementation, as prospective validation, workflow evaluation, direct human comparison, fairness assessment, and real-world deployment remain limited. Importantly, none of the included studies evaluated transformer-based language models within behavioral-health services, workflows, or service-level outcomes. Therefore, the application of these technologies to behavioral health should be regarded as a promising future research direction rather than an established clinical application.

Bridging this evidence gap will require prospective, multi-site validation; standardized external benchmarking; subgroup-specific fairness evaluation; workflow integration and implementation studies; and the development of explainable, ethically grounded human–AI partnership models that support transparent, trustworthy, and equitable clinical decision making. Future evidence syntheses should also extend the temporal scope to include foundational transformer research, broaden database coverage to engineering and preprint sources, incorporate formal methodological quality appraisal where appropriate, and compare transformer-based and emerging alternative AI architectures to provide a more comprehensive understanding of the evolving clinical AI landscape.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/info17070712/s1>, Supplementary File, Completed PRISMA-ScR checklist.

Author Contributions: Conceptualization, S.H. and H.N.; methodology, S.H. and H.N.; formal analysis, S.H. and H.N.; investigation, S.H. and H.N.; resources, S.H.; data curation, S.H.; writing—original draft preparation, H.N.; writing—review and editing, S.H. and H.N.; visualization, S.H.; supervision, S.H.; project administration, S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Indiana University Indianapolis Informatics Consortium pilot Research Grant (#2290001). The funding source had no role in the design of the study, data collection, analysis, interpretation of findings, or preparation of the manuscript.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The study is a scoping review that synthesized findings from previously published, peer-reviewed journal articles and did not generate, collect, or analyze any new primary data. As such, no new datasets were created during this research. Accordingly, data sharing is not applicable because all data used in the study are derived from publicly available published sources that are fully cited in the manuscript.

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

References

1. U.S. Department of Health and Human Services. *Fiscal Year 2024 HHS Agency Financial Report*; U.S. Department of Health and Human Services: Washington, DC, USA, 2024.
2. Brown, J.; Breslau, J.; Wishon, A.; Miller, R.; Kase, C.; Dunbar, M.; Stewart, K.; Briscoombe, B.; Rose, T.; Dehus, E.; et al. *Implementation and Impacts of the Certified Community Behavioral Health Clinic Demonstration: Findings from the National Evaluation*; U.S. Department of Health and Human Services: Washington, DC, USA, 2021.
3. Hong, S.; Walton, B.; Kim, H.-W.; Rhee, T.G. Predicting the Behavioral Health Needs of Asian Americans in Public Mental Health Treatment: A Classification Tree Approach. *Adm. Policy Ment. Health Ment. Health Serv. Res.* **2023**, *50*, 630–643. [[CrossRef](#)] [[PubMed](#)]
4. Hallett, E.; Simeon, E.; Amba, V.; Howington, D.; McConnell, K.J.; Zhu, J.M. Factors influencing turnover and attrition in the public behavioral health system workforce: Qualitative study. *Psychiatr. Serv.* **2024**, *75*, 55–63. [[CrossRef](#)] [[PubMed](#)]
5. Njuguna, M.; Ngune, I.; Middlewick, Y. Mental health support workers' perspectives on barriers to and facilitators of the effective delivery of their roles: A systematic review and meta-aggregation. *Community Ment. Health J.* **2025**, *61*, 1548–1573. [[CrossRef](#)] [[PubMed](#)]
6. Ferreira, R.M. New evidence-based practice: Artificial intelligence as a barrier breaker. *World J. Methodol.* **2023**, *13*, 384–389. [[CrossRef](#)] [[PubMed](#)]
7. Sezgin, E.; McKay, I. Behavioral health and generative AI: A perspective on future of therapies and patient care. *npj Ment. Health Res.* **2024**, *3*, 25. [[CrossRef](#)] [[PubMed](#)]
8. Anand, S.K.; Kumar, S. Uncertainty analysis in ontology-based knowledge representation. *New Gener. Comput.* **2022**, *40*, 339–376. [[CrossRef](#)]
9. Arbabi, A.; Adams, D.R.; Fidler, S.; Brudno, M. Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR Med. Inform.* **2019**, *7*, e12596. [[CrossRef](#)] [[PubMed](#)]
10. Sharma, A.; Kumar, S. Machine learning and ontology-based novel semantic document indexing for information retrieval. *Comput. Ind. Eng.* **2023**, *176*, 108940. [[CrossRef](#)]
11. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [[CrossRef](#)]
12. Dang, H.A.; Tran, V.; Nguyen, L.M. Survey and analysis of hallucinations in large language models: Attribution to prompting strategies or model behavior. *Front. Artif. Intell.* **2025**, *8*, 1622292. [[CrossRef](#)] [[PubMed](#)]
13. Bedi, S.; Liu, Y.; Orr-Ewing, L.; Dash, D.; Koyejo, S.; Callahan, A.; Fries, J.A.; Wornow, M.; Swaminathan, A.; Lehmann, L.S.; et al. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA* **2025**, *333*, 319–328. [[CrossRef](#)] [[PubMed](#)]
14. Wang, L.; Wan, Z.; Ni, C.; Song, Q.; Li, Y.; Clayton, E.; Malin, B.; Yin, Z. Applications and concerns of ChatGPT and other conversational large language models in health care: Systematic review. *J. Med. Internet Res.* **2024**, *26*, e22769. [[CrossRef](#)] [[PubMed](#)]
15. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [[CrossRef](#)] [[PubMed](#)]
16. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Kastner, M.; Levac, D.; Ng, C.; Sharpe, J.P.; Wilson, K.; et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [[CrossRef](#)] [[PubMed](#)]
17. Botero-Aguirre, J.P.; García-Rivera, M.A. Natural language processing for enhanced clinical decision support in allergy verification for medication prescriptions. *Mayo Clin. Proc. Digit. Health* **2025**, *3*, 100244. [[CrossRef](#)] [[PubMed](#)]
18. Cutforth, M.; Watson, H.; Brown, C.; Wang, C.; Thomson, S.; Fell, D.; Dilys, V.; Scrimgeour, M.; Schrempf, P.; Lesh, J.; et al. Acute stroke CDS: Automatic retrieval of thrombolysis contraindications from unstructured clinical letters. *Front. Digit. Health* **2023**, *5*, 1186516. [[CrossRef](#)] [[PubMed](#)]
19. Xie, K.; Gallagher, R.S.; Shinohara, R.T.; Xie, S.X.; Hill, C.E.; Conrad, E.C.; Davis, K.A.; Roth, D.; Litt, B.; Ellis, C.A. Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. *Epilepsia* **2023**, *64*, 1900–1909. [[CrossRef](#)] [[PubMed](#)]

20. Laabidi, A.; Aissaoui, M.; Madani, M.A. Extraction of disease symptoms from free text using natural language processing techniques. In *Proceedings of the Ninth International Congress on Information and Communication Technology*; Yang, X.-S., Sherratt, R.S., Dey, N., Joshi, A., Eds.; Springer Nature: Singapore, 2024; Volume 2, pp. 549–561.
21. Eslami, B.; Dligach, D.; Strickland, B.; Azarvash, N.; Tootooni, M.S. A performance-based voting framework for assertion detection in clinical notes. In *MEDINFO 2025—Healthcare Smart × Medicine Deep*; Househ, M.S., Borycki, E., Kushniruk, A., Eds.; IOS Press: Amsterdam, The Netherlands, 2025; pp. 1125–1129.
22. Kim, H.K.; Park, Y.; Kim, Y.J.; Yi, S.; Park, Y.; So, S.; Lee, H.J.; Bae, Y.S. EILEEN: A multi-modal framework for extracting alcohol consumption patterns from bilingual clinical notes. *IEEE Access* **2025**, *13*, 25741–25751. [[CrossRef](#)]
23. Huang, J.; Yang, D.M.; Rong, R.; Nezafati, K.; Treager, C.; Chi, Z.; Wang, S.; Cheng, X.; Guo, Y.; Klesse, L.J.; et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digit. Med.* **2024**, *7*, 106. [[CrossRef](#)] [[PubMed](#)]
24. Zhang, H.; Vithanage, D.; Song, T.; Deng, C.; Yu, P. Leveraging retrieval augmented generation-driven large language models to extract dementia agitation symptoms and triggers from free-text nursing notes. In *MEDINFO 2025—Healthcare Smart × Medicine Deep*; Househ, M.S., Tariq, Z.U.A., Al-Zubaidi, M., Shah, U., Huesing, E., Eds.; IOS Press: Amsterdam, The Netherlands, 2025; pp. 799–803.
25. Sivarajkumar, S.; Tam, T.Y.C.; Mohammad, H.A.; Viggiano, S.; Oniani, D.; Visweswaran, S.; Wang, Y. Extraction of sleep information from clinical notes of Alzheimer’s disease patients using natural language processing. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 2217–2227. [[CrossRef](#)] [[PubMed](#)]
26. Hier, D.B.; Carrithers, M.A.; Platt, S.K.; Nguyen, A.; Giannopoulos, I.; Obafemi-Ajayi, T. Preprocessing of physician notes by LLMs improves clinical concept extraction without information loss. *Information* **2025**, *16*, 446. [[CrossRef](#)]
27. Chi, J.; Roupail, Y.; Hillis, E.; Ma, N.; Nguyen, A.; Wang, J.; Hofford, M.; Gupta, A.; Lyons, P.G.; Wilcox, A.; et al. EchoLLM: Extracting echocardiogram entities with light-weight, open-source large language models. *JAMIA Open* **2025**, *8*, ooaf092. [[CrossRef](#)] [[PubMed](#)]
28. Cho, H.; Yoo, S.; Kim, B.; Jang, S.; Sunwoo, L.; Kim, S.; Lee, D.; Kim, S.; Nam, S.; Chung, J.-H. Extracting lung cancer staging descriptors from pathology reports: A generative language model approach. *J. Biomed. Inform.* **2024**, *157*, 104720. [[CrossRef](#)] [[PubMed](#)]
29. Arzideh, K.; Baldini, G.; Winnekens, P.; Friedrich, C.M.; Nensa, F.; Idrissi-Yaghir, A.; Hosch, R. A transformer-based pipeline for German clinical document de-identification. *Appl. Clin. Inform.* **2025**, *16*, 31–43. [[CrossRef](#)] [[PubMed](#)]
30. Alba, P.R.; Gan, Q.; Hu, M.; Zhu, S.H.; Sherman, S.E.; DuVall, S.L.; Conway, M. Development of a natural language processing system to identify clinical documentation of electronic cigarette use. In *MEDINFO 2023—The Future Is Accessible*; Bichel-Findlay, J., Otero, P., Scott, P., Huesing, E., Eds.; IOS Press: Amsterdam, The Netherlands, 2024; pp. 659–663.
31. Floyd, S.B.; Almeldien, A.G.; Smith, D.H.; Judkins, B.; Krohn, C.E.; Reynolds, Z.C.; Jeray, K.; Obeid, J.S. Using artificial intelligence to develop a measure of orthopaedic treatment success from clinical notes. *Front. Digit. Health* **2025**, *7*, 1523953. [[CrossRef](#)] [[PubMed](#)]
32. Bowles, A.E.; Gan, Q.; Hanchrow, E.; Duvall, S.; Alba, P.R.; Shi, J. Comparative analysis of NLP models for automatic LOINC document ontology named entity recognition in clinical note titles. In *MEDINFO 2025—Healthcare Smart × Medicine Deep*; Househ, M.S., Tariq, Z.U.A., Al-Zubaidi, M., Shah, U., Huesing, E., Eds.; IOS Press: Amsterdam, The Netherlands, 2025; pp. 769–773.
33. Comito, C.; Forestiero, A.; Macrì, D.; Metlichin, E.; Giusti, G.D.; Ramacciati, N. Comparative analysis of AI algorithms on real medical data for chronic pain detection. *Int. J. Med. Inform.* **2025**, *203*, 106002. [[CrossRef](#)] [[PubMed](#)]
34. Laurentiev, J.; Kim, D.H.; Mahesri, M.; Wang, K.-Y.; Bessette, L.G.; York, C.; Zakoul, H.; Lee, S.B.; Zhou, L.; Lin, K.J. Identifying functional status impairment in people living with dementia through natural language processing of clinical documents: Cross-sectional study. *J. Med. Internet Res.* **2024**, *26*, e47739. [[CrossRef](#)] [[PubMed](#)]
35. Ameli, N.; Firoozi, T.; Gibson, M.; Lai, H. Classification of periodontitis stage and grade using natural language processing techniques. *PLoS Digit. Health* **2024**, *3*, e0000692. [[CrossRef](#)] [[PubMed](#)]
36. Jaiswal, A.; Katz, A.; Nesca, M.; Milios, E. Identifying risk factors associated with lower back pain in electronic medical record free text: Deep learning approach using clinical note annotations. *JMIR Med. Inform.* **2023**, *11*, e45105. [[CrossRef](#)] [[PubMed](#)]
37. Holmes, B.; Raymer, M.; Banerjee, T. Extraction of patient subpopulations with psychiatric symptoms using a transformer architecture. In *Proceedings of the 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–4.
38. Kennedy, E.E.; Davoudi, A.; Hwang, S.; Freda, P.J.; Urbanowicz, R.; Bowles, K.H.; Mowery, D.L. Identifying barriers to post-acute care referral and characterizing negative patient preferences among hospitalized older adults using natural language processing. *AMIA Annu. Symp. Proc.* **2023**, *2022*, 606–615. [[PubMed](#)]
39. Miller, K.; Bedrick, S.; Lu, Q.; Wen, A.; Hersh, W.; Roberts, K.; Liu, H. Dynamic few-shot prompting for clinical note section classification using lightweight, open-source large language models. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 1164–1173. [[CrossRef](#)] [[PubMed](#)]

40. Nguyen Phan, T.M.; Dao, C.-T.; Wu, C.; Wang, J.-Z.; Liu, S.; Ding, J.-E.; Restrepo, D.; Liu, F.; Hung, F.-M.; Peng, W.-C. MEDFuse: Multimodal EHR data fusion with masked lab-test modeling and large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*; ACM: New York, NY, USA, 2024; pp. 3974–3978.
41. de Hond, A.; van Buchem, M.; Fanconi, C.; Roy, M.; Blayney, D.; Kant, I.; Steyerberg, E.; Hernandez-Boussard, T. Predicting depression risk in patients with cancer using multimodal data: Algorithm development study. *JMIR Med. Inform.* **2024**, *12*, e51925. [[CrossRef](#)] [[PubMed](#)]
42. Li, J.; Wang, X.; Cai, L.; Sun, J.; Yang, Z.; Liu, W.; Wang, Z.; Lv, H. An interpretable deep learning framework for predicting liver metastases in postoperative colorectal cancer patients using natural language processing and clinical data integration. *Cancer Med.* **2023**, *12*, 19337–19351. [[CrossRef](#)] [[PubMed](#)]
43. Kolukula, N.R.; Puli, S.; Babi, C.; Kalapala, R.P.; Ongole, G.; Chinta, V.M.K. Processing of clinical notes for efficient diagnosis with feedback attention-based BiLSTM. *Med. Biol. Eng. Comput.* **2024**, *62*, 3193–3208. [[CrossRef](#)] [[PubMed](#)]
44. Abdullahi, T.; Mercurio, L.; Singh, R.; Eickhoff, C. Retrieval-based diagnostic decision support: Mixed methods study. *JMIR Med. Inform.* **2024**, *12*, e50209. [[CrossRef](#)] [[PubMed](#)]
45. Yang, Z.; Lin, Y.; Xu, Y.; Hu, J.; Dong, S. Interpretable disease prediction via path reasoning over medical knowledge graphs and admission history. *Knowl.-Based Syst.* **2023**, *281*, 111082. [[CrossRef](#)]
46. Gao, Y.; Li, R.; Croxford, E.; Caskey, J.; Patterson, B.W.; Churpek, M.; Miller, T.; Dligach, D.; Afshar, M. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI* **2025**, *4*, e58670. [[CrossRef](#)] [[PubMed](#)]
47. Daphne, S.; Rajam, V.M.A.; Hemanth, P. An ensemble patient graph framework for predictive modelling from electronic health records and medical notes. *Diagnostics* **2025**, *15*, 756. [[CrossRef](#)] [[PubMed](#)]
48. Lyu, L.; Zhu, H.; Chen, H.; Zhou, J.; Chan, R.H.M.; Lu, L. PDSNet: Patient-disease dual spatial similarity neural networks for predicting heart failure risk using short electronic health records. *IEEE J. Biomed. Health Inform.* **2025**, *30*, 676–687.
49. Chien, A.; Tang, H.; Jagessar, B.; Chang, K.-W.; Peng, N.; Nael, K.; Salamon, N. AI-assisted summarization of radiological reports: Evaluating GPT3davinci, BARTnn, LongT5booksum, LEDbooksum, LEDlegal, and LEDclinical. *Am. J. Neuroradiol.* **2024**, *45*, 244–248. [[CrossRef](#)] [[PubMed](#)]
50. Chuang, Y.N.; Tang, R.; Jiang, X.; Hu, X. SPeC: A soft prompt-based calibration on performance variability of large language models in clinical note summarization. *J. Biomed. Inform.* **2024**, *151*, 104606. [[CrossRef](#)] [[PubMed](#)]
51. Eghbali, N.; Klochko, C.; Mahdi, Z.; Alhiari, L.; Lee, J.; Knisely, B.; Craig, J.; Ghassemi, M.M. Enhancing radiology clinical histories through transformer-based automated clinical note summarization. *J. Imaging Inform. Med.* **2025**, *39*, 1031–1039. [[CrossRef](#)] [[PubMed](#)]
52. Afshar, M.; Resnik, F.; Baumann, M.R.; Hintzke, J.; Lemmon, K.; Sullivan, A.G.; Shah, T.; Stordalen, A.; Oberst, M.; Dambach, J.; et al. A novel playbook for pragmatic trial operations to monitor and evaluate ambient artificial intelligence in clinical practice. *NEJM AI* **2025**, *2*, e2406603. [[CrossRef](#)] [[PubMed](#)]
53. Selvan, R. Sustainability of large-scale artificial intelligence models in health care. *Lancet Digit. Health* **2026**, *8*, 101016. [[CrossRef](#)] [[PubMed](#)]
54. Chen, R.J.; Ding, T.; Lu, M.Y.; Williamson, D.F.K.; Jaume, G.; Song, A.H.; Chen, B.; Zhang, A.; Shao, D.; Shaban, M.; et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **2024**, *30*, 850–862. [[CrossRef](#)] [[PubMed](#)]
55. Van Calster, B.; Collins, G.S.; Vickers, A.J.; Wynants, L.; Kerr, K.F.; Barreñada, L.; Varoquaux, G.; Singh, K.; Moons, K.G.; Hernandez-Boussard, T.; et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance. *Lancet Digit. Health* **2025**, *7*, 100916. [[CrossRef](#)] [[PubMed](#)]
56. Paradowska-Stolarz, A.M. AI in modern dentistry: Hype, hope, or real transformation? *Dent. Med. Probl.* **2025**, *62*, 1013–1016. [[CrossRef](#)] [[PubMed](#)]
57. Umer, F.; Batool, I.; Naved, N. Innovation and application of large language models (LLMs) in dentistry: A scoping review. *BDJ Open* **2024**, *10*, 90. [[CrossRef](#)] [[PubMed](#)]
58. Liang, E.N.; Pei, S.; Staibano, P.; van der Woerd, B. Clinical applications of large language models in medicine. *J. Int. Med. Res.* **2025**, *53*, 03000605251347556. [[CrossRef](#)] [[PubMed](#)]

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